Johdatus todennäköisyyslaskentaan ja tilastolliseen päättelyyn

sekä

Tilastollisten menetelmien perusteet

11.2.2024

Pekka Pere
Aalto-yliopisto

Huomioita painovirheistä, virheistä ja muista puutteista sekä ehdotuksia uusiksi esimerkkiaineistoiksi pyydän osoitteeseen pekka.j.pere@aalto.fi.

Catzo / lue / etzi / ia tutki

Ote Mikael Agricolan Rucouskiriasta 1544

Sisällys

Luku 1

Johdanto

> Kaikki tutkimus on siis suhteuttavaa vertailua, joskus helppoa, joskus vaikeaa. – – Suhde – – merkitsee yhtäläisyyttä jossain asiassa ja eroa jossain muussa, eikä sitä siksi voida ajatella käyttämättä avuksi lukuja. Luvut sisältävät näin ollen kaiken sen, mikä voidaan suhteuttaa keskenään. Luvut eivät nimittäin liity pelkästään määrään, josta suhde syntyy, vaan kaikkeen siihen, mikä voi joko olemuksellisesti tai ominaisuuksiltaan olla yhtä tai erota toisistaan. Ehkä Pythagoras juuri tämän vuoksi päätteli, että kaikki sekä rakentuu että tulee ymmärretyksi lukujen voimasta.[1]
>
> Nicolaus Cusanus (1401–1464)

> Ottaessamme kirjan käteemme kysykäämme, onko siinä abstraktia järkeilyä määristä tai numeroista? Ei. Onko siinä empiiristä järkeilyä tosiasioista ja olevaisesta? Ei. Heitä se tuleen, sillä se ei voi sisältää muuta kuin viisastelua ja harhaluuloja.[2]
>
> David Hume (1711–1776)

1.1 Tehdään tästä numero

Mittaamisessa on voimaa. Mikä voidaan kertoa numeroin, se konkretisoituu. Yhteiskunnallinen mielenkiinto ja keskustelu herää, kun ongelma osoitetaan luvuilla. Toimenpiteiden onnistumista voidaan arvioida uskottavasti, jos sitä voidaan kuvata numeroin. Kvantifioitu motivoi toimiin, koska tulos on mitattavissa. Mikä mitataan, se tulee tehdyksi (Handin 2016, 38, lainaama sanonta). Mitä ja miten mitataan vaikuttaa toimiimme, yhteiskuntaamme ja tavoitteisiimme. Mahdollisuus teettää vaikkapa kyselytutkimus ja kertoa sen tuloksista on oikeus ja yksi vapaan yhteiskunnan tunnusmerkki. Yhteiskunnan numeerista

kuvaamista pidetään niin tärkeänä, että tilastolaissa (23.4.2004/280) on säädetty valtion tehtäväksi "yhteiskuntaoloja ja niiden kehitystä kuvaavien tilastojen" laatiminen "yleistä käyttöä varten".

Myös tieteessä mittaaminen on valtaa. Leonardo da Vincin (1452–1519) pyrkimys oli selvittää "täysi tieto ihmisen mitasta ja siitä, mikä on ihmisen paikka maailmankaikkeudessa – – mikä on *universale misura del huomo*, ihmisen universaali mitta" (Isaacson 2019, 243). Nykyisenlaisen tilastotieteen alkuunpolkaisija Francis Galton (1879) arvioi psykometriikkaa (tilastotieteen soveltamista psykologiaan) pohtiessaan tieteen ylipäänsä kiinnittyvän mittaamiseen ja numeroihin:

> – – mikään tietämyksenala ei ansaitse tieteen arvonimeä ennen kuin sen tutkimaa ilmiötä voidaan mitata ja kuvata numeroilla.

Fyysikko William Thomsonin (sittemmin lordi Kelvinin) näkemys 1883 on kuuluisa (Thomson 1889, 73–74):

> Jos pystyy mittaamaan ja ilmaisemaan numeroilla, mistä puhuu, tietää asiasta jotakin. Jos asiaa ei pysty mittamaan eikä esittämään sitä numeroilla, tietämys on niukkaa ja epätyydyttävää. Tietämys saattaa olla tällöin aluillaan mutta on tuskin tieteellistä, mikä asia onkaan.

Teksti on kaiverrettu Chicagon yliopiston sosiaalitieteiden rakennuksen kiveen (McCloskey 2000, 80). Luvut mahdollistavat tieteellisten teorioiden testaamisen.

Mittaamista yritetään, vaikka se olisi vaikeaa tai pohdituttaisi moraalisesti. Suomen museoliitto tiedotti 26.1.2023, että kunkin museokäynnin koetun hyvinvoinnin arvo on keskimäärin 864 euroa.[3] Galton (1901b) päätteli William Farrin (1853) laskelmista, että "huonompilaatuisen" vauvan arvo saattaa olla viisi puntaa mutta "huippuluokan" vauvan tuhansia puntia. Tänä päivänäkin ihmisalun rahallista arvoa voidaan yrittää mitata (Clarke 2021). Vaikeudesta huolimatta kvantifiointi voi olla välttämätöntä. Yhteiskunta joutuu puntaroimaan ihmiselämän ja luonnon arvoa päättäessään, kuinka paljon resursseja käytetään terveydenhoitoon tai vastatoimiin ilmastonmuutokselle (esim. Broome 1985, Almond 2006, Saarni ja Nyblin 2022, jakson 6.2 harjoitustehtävä).

Kaikki tämä ei merkitse, että numeroitavissa oleva olisi tärkeää tai että se, mitä ei voida ilmaista lukuarvolla, ei olisi merkityksellistä. Oscar Wilden pisteliäs tiivistys ihmisestä, "joka tuntee kaiken hinnan muttei minkään arvoa", on kuuluisa.

Robert Kennedy kritisoi puheessaan Kansasin yliopistolla 18.3.1968 laajemmin ja konkreettisemmin mittaamisen roolia yhteiskunnassa (John F. Kennedy Presidential Library and Museum 2017):

> Vuotuinen bruttokansantuotteemme on yli 800 miljardia dollaria – – . Se sisältää ilman saastumisen, tupakan mainonnan ja ambulanssit, joilla kiidätetään

> liikenneonnettomuuksien uhrit. Se sisältää turvalukot ja vankilat niiden rikkojille. Se sisältää punapuiden tuhoamisen ja luonnonihmeittemme menettämisen – – . Se sisältää napalmin ja ohjusten ydinkärjet sekä aseistetut autot, joilla poliisi taistelee mellakoita vastaan kaupungeissamme. Se sisältää murhaajien kiväärit ja puukot sekä televisio-ohjelmat, jotka glorifioivat väkivaltaa, jotta lapsillemme saataisiin myydyksi leluja. Silti bruttokansantuote ei kata lastemme terveyttä, heidän koulutuksensa laatua tai iloa heidän leikeistään. Se ei kata runojemme kauneutta tai avioliittojemme lujuutta, julkisen keskustelumme älykkyyttä tai viranomaistemme lahjomattomuutta. Se ei mittaa nokkeluuttamme tai rohkeuttamme, tietämystämme, oppineisuuttamme, myötätuntoamme tai kiintymystä maahamme. Se mittaa kaikkea paitsi sitä, mikä on elämässä arvokasta.

Kennedyn ajatukset ovat edelleen ajankohtaisia ja soveltuvat nykypäivän suomalaiseen yhteiskuntaan. Maailman rikkaimmaksi arvioitu nainen Liliane Bettencourt tiivisti:[4]

> Elämässä merkityksellisintä on se, mitä ei voi mitata.

Kvalitatiivisella tutkimuksella voi olla mahdollista selvittää asioita syvällisemmin kuin numeroiden avulla. Kaikkea ei voi takoa numeroiksi. Suuri kuvio on silti selvä: Numeroiden mahti on valtava. Kvantifioimalla asia konkretisoituu, siihen kiinnitetään enemmän huomiota, sen analysointi helpottuu ja tarvittaessa asialle ollaan halukkaampia tekemään jotain. Näin on sekä yhteiskunnassa että tieteessä.

1.2 Opiskelija ja tilastotiede

Tilastotiede on maailman jännittävin oppiaine. Sitä voi hyödyntää likipitäen kaikilla elämän- ja mielenkiinnonalueilla. Vahvan teorian ja modernin tietotekniikan avulla se röntgensäteiden tapaan paljastaa silmälle näkymättömän todellisuuden. (Hand 2008, 2009.) Tilastotiede on silta matematiikan ja empirian välillä.

Tilastotieteen juuret ovat empiirisissä ja yhteiskunnallisissa kysymyksissä. Jos haluaa ratkaista globaaleja ongelmia, tilastotiede on ratkaisu. Nälänhädät, työttömyys, yksinäisyys, maahanmuuton ongelmat, fyysiset tai psyykkiset sairaudet, biodiversiteetti, ilmastonmuutos. Kaikkeen saa otteen tilastotieteellä. Jos rikastuminen kiinnostaa maailman parantamista enemmän, jälleen tilastotiede on oikea ala. Työpaikkojen lukumäärän ja ansiotason ennusteet lupaavat mannaa soveltajille, varsinkin jos maustaa CV:tään koodaustaidolla ja suuntaa liike-elämään.

Paitsi että tilastotiede on hauskaa, se on hyödyllinen pää- tai sivuaine ja monissa opinnäytteissä välttämätön työkalu. Opiskelijan elämän säännönmu-

kaisuuksia on katumus opinnäytteen tekovaiheessa, miksen opiskellut enemmän tilastotiedettä.

Opiskeluasenteeksi ei sovi kiire. Luetun sivumäärän päivää kohden ei ole suotavaa olla kovin suuri. On oleellista, että teoriaa opiskellaan kynä kädessä tekemällä laskuja ja merkintöjä sivujen syrjään ja apupapereille. Tilastotieteen ymmärtämiseen tarvitaan aikaa; tilastotieteen soveltamiseen tilasto-ohjelmisto (jollei aineisto ole pieni ja tehtävä hyvin yksinkertainen). Tilastotiedettä ei kannata yrittää opetella vain lukemalla.

Tutkimusetiikka kannustaa tilastotieteen huolelliseen opiskeluun. Menetelmiä ei tule opiskella kikka-asenteella ilman kunnon ymmärrystä. Menetelmiä voi käyttää luotettavasti vain, jos ymmärtää ne.

Tilastotieteen peruskurssit ovat pisimmälle kantavia kursseja. Oman alansa tehtäviin sijoittuva maisteri hyödyntää uransa alkuvaiheessa erityistietämystään — ja mahdollisesti kvantitatiivista tietoa. Työuralla moni erkaantuu asiantuntijaroolista ja siirtyy johtajuutta vaativiin tehtäviin. Edelleen ellei enenevässä määrin hän tekee päätöksiä kvantitatiiviseen tietoon perustuen. Miten kukin eteneekin, maailma kehittyy kiivaasti suuntaan, jossa kyky ymmärtää ja käsitellä numeerista aineistoa eli dataa on yhä tärkeämpää.[5] (Hand 2008, 2009.)

Tilastotiede ei ole helppoa. Onko se liian vaikeaa vaikkapa sosiaali- tai käyttäytymistieteiden opiskelijalle, joka ei ole tarvinnut matematiikkaa muissa opinnoissaan? Karl Pearson piti toivottomana Galtonin yritystä 1899 saada antropologit ymmärtämään korrelaatiota (Pearson 1930, 57). Ajat ovat muuttuneet. Harva ajatellee tänä päivänä, että antropologi ei voisi ymmärtää nykyisin yleissivistykseen kuuluvaa korrelaatiota. Entinen mahdottomuus on nykyinen itsestäänselvyys.

Kaikki alla on opittavissa lukion lyhyen matematiikan tietopohjalta, jos on motivoitunut ja valmis käyttämään aikaa opiskeluun. Ilo uudesta kielestä ja tavasta hahmottaa maailmaa sekä kyky ymmärtää ja tutkia sitä ovat palkintoja aherruksesta.

Esimerkki. Nancy Reidin (kuva 1.1) tutkimus on matemaattista tilastotiedettä. Hän on toiminut *Canadian Journal of Statisticsin* päätoimittajana sekä Kanadan tilastoseuran (Statistical Society of Canada) että Matemaattisen tilastotieteen instituutin (Institute of Mathematical Statistics) presidenttinä, hänet on kutsuttu Yhdysvaltain Kansallisen tiedeakatemian (National Academy of Sciences) jäseneksi, hänelle on myönnetty Kanadan korkea-arvoisin kunniamerkki, Iso-Britannian Kuninkaallinen tilastoseura (Royal Statistical Society) on myöntänyt hänelle kultaisen Guy -mitalin hänen elämäntyöstään ja Amerikan tilastoseura ASA (American Statistical Association) *David R. Cox Foundations*

Kuva 1.1: Nancy Reid (18 vuotta).

of Statistics -palkinnon ensimmäisenä tämän kunnian saaneena. Reid ei alkuun ymmärtänyt tilastotiedettä ja innostui siitä vasta alkuvaikeuksien jälkeen:[6]

> Minulla ei ollut lainkaan lahjoja ohjelmointiin, joten vaihdoin pääaineeni tilastotieteeseen. – – Kaikkien täytyi ottaa toinen tilastotieteen kurssi. Se oli tietenkin meistä vaikea, emmekä ymmärtäneet sitä kunnolla, mutta sain hyvän arvosanan. Niinpä kävin seuraavaksi Tilastollisen päättelyn kurssin, mitä emme myöskään ymmärtäneet, ja saimme kaikki melko huonon arvosanan. Kurssin loppuosa oli regressiota. Se oli todella hauskaa.

Shayle Searle teki uraauurtavaa matemaattista tilastotiedettä ja hänen oppikirjojaan käytetään ympäri maailmaa. Searle (2005) kertoi vastoinkäymisistään: Hänen professorinsa kieltäytyi suosittelemasta hänelle stipendiä sanoen, että hän ei ole tarpeeksi hyvä. Hän reputti pääsykokeen. Toinen professori neuvoi, että hänen ei kannata yrittää esittää opinnäytettään väitöskirjana. Maisterin tutkinnon kokeen hän suoritti rimaa hipoen pettäen ohjaajansa odotukset. Searle neuvoi opiskelijaa: "Älä anna vastoinkäymisen tyrmätä. Älä luovuta!"

Sekä pää pyörällä olevasta opiskelijasta Reidistä että vastoinkäymisestä toiseen kompuroineesta Searlesta kehittyi molemmista ylistettyjä maailmankuuluja taitajia. □

1.3 Mitä tilastotiede on?

Tilastotieteessä olennaista on epävarmuuden arviointi, joten sisällytetään se *tilastotieteen* (*statistics*) määritelmään:

> *Tilastotiede on tiede aineiston keräämisestä, kuvauksesta ja analyysista. Tilastotieteellä arvioidaan aineistoon ja siitä tehtäviin päätelmiin liittyvää satunnaisuutta ja epävarmuutta.*

Määritelmässä yllä ei mainita tilastoja. Niillä tarkoitetaan yleensä järjestelmällisesti koottuja numeerisia yhteiskuntaa koskevia aineistoja (vrt. Luther 1993, 11). Nimestään huolimatta tilastotieteellä tutkitaan aineistoja ylipäänsä eikä vain tilastoja.

Keskeinen osa tilastotieteellistä analyysiä on *tilastollinen päättely* eli aineistoon perustuva päättely tutkittavasta satunnaisilmiöstä sekä päätelmien epävarmuuden mittaaminen. Tilastollinen päättely on tyypillisesti induktiivista: Yksittäisestä aineistosta pyritään tekemään yleispäteviä päätelmiä maailmasta. Tilastollinen päättely tehdään useimmiten *tilastollisen mallin* avulla. Tilastollinen malli asettaa satunnaisilmiölle rajoituksia kuten oletuksia havaintojen syntymekanismista tai *todennäköisyysjakauman* (luku 6). Tilastolliseen päättelyyn liittyy tyypillisesti satunnaisuutta mittaavien tunnuslukujen ja luottamusvälien laskua sekä hypoteesien testausta (luvut 10–11). Tilastollinen malli on matemaattinen yksinkertaistus maailmasta. Malli selkeyttää ajattelua ja helpottaa päättelyä.

1.4 Muuta

Tällaisten viivojen välissä on teoreettisesti vaikeampaa tai täydentävää tietoa tai osoitan lukijalle parempia menetelmiä tilanteisiin, joissa päätekstissä opettamani menetelmät eivät ole riittäviä. Lisäysten toivon herättävän kiinnostusta teoriaan ja auttavan aitojen tutkimusongelmien kanssa painivia.

Luku 2

Aineistot

> Jos meillä on aineisto, katsotaan sitä. Jos meillä on vain mielipiteitä, mennään minun mielipiteelläni.[7]
>
> Jim Barksdale (1943–)

> Onko muuta aineistoa? Usein on, ja toinen aineisto voi olla hyvin hyödyllinen.[8]
>
> Bradley Efron (1938–)

Luvussa käsitellään aineistoja tärkeyden, luotettavuuden ja mitta-asteikkojen näkökulmista. Näkökulma siirtyy empiirisestä teoreettiseen.

Ennen nykytiedettä tieto ajateltiin olevan johdettavissa aiemmasta ymmärryksestä loogisella päättelyllä. Nykyään ajatellaan, että tiedon tulee pohjautua empiiriseen aineistoon. Pyrkimys on oikea, kunhan aineisto on luotettavaa. Luvussa varoitetaan tilastojen ja aineistojen ongelmista. Aineiston tarkoituksenmukaisuutta ja soveltuvuutta otantateorian, kyselyjen ja tilastollisen päättelyn kannalta pohditaan luvussa 8.

2.1 Keskeisin käsite

Aineisto on tilastotieteen keskeisin käsite. Ilman aineistoa — teoreettista tai empiiristä — ei ole tilastotiedettäkään. Jos asiaa ei ole tutkittu, mahdollinen eikä lainkaan harvinainen syy on, että aineistoa on vaikea saada. Priima aineisto on arvokas.

Aineisto koostuu havainnoista. Mitä enemmän havaintoja, sitä informatiivisempi aineisto yleensä on, jos se on kerätty tarkoituksenmukaisesti.

2.2 Mitta-asteikot

Sopiva tilastotieteellinen menetelmä riippuu mitta-asteikosta, jolla havaintoja on mitattu. Sopivuus riippuu nimenomaan käytetystä mitta-asteikosta; ei välttämättä mitatusta muuttujasta. Muuttujaa saatettaisiin voida mitata toisellakin mitta-asteikolla. Neljä toistaan tarkempaa mitta-asteikkoa tavataan erotella (Stevenson 1946).

Karkein mitta-asteikko on *luokka-asteikko* (*nominal scale*). Sillä havainnot jaotellaan luokkiin, jotka eroavat toisistaan laadullisesti. Luokkia ei voida asettaa järjestykseen.

Esimerkki. Työmarkkinoilla olevat ovat työttömiä tai työllisiä. Yliopistossa on monta tiedekuntaa. Wikipedia luettelee puolisen tuhatta musiikkityyliä: Ambient, ooppera, progressiivinen rock, rap jne.[9] Luokilla ei ole esimerkeissä ilmeistä järjestystä. □

Monia asioita voidaan mitata *järjestysasteikolla* (*ordinal scale*). Tällaiset mittaukset voidaan asettaa järjestykseen, mutta luokkien välisiä eroja ei voida mitata numeerisesti.

Esimerkki. Jaetaan työmarkkinoilla olevat ihmiset työttömiin, osa-aikatyötä tai täyspäivätyötä tekeviin. Äänestäjät saatetaan jakaa poliittisen kantansa mukaan luokkiin, jotka kuvaavat poliittisen kannan liberaaliutta tai konservatiivisuutta. Luokat voidaan järjestää työmäärän tai aseman liberaali–konservatiivi -akselilla mukaan. □

Useimmiten tilastotieteellisiä analyyseja tehdään *välimatka-asteikolla* (*interval scale*) mitatuilla muuttujilla. Mittausten erojen — välimatkojen — suuruudet ovat tällöin todettavissa numeerisesti.

Esimerkki. Mitataan tunnin tarkkuudella ihmisten tekemät työtunnit (esim. kuukaudessa). Mitataan lämpötila ulkona 0.5 celsiusasteen tarkkuudella. Molemmissa tilanteissa voidaan todeta numeroin ilmaistavissa oleva ero työtuntien tai lämpötilojen välillä. □

Suhdeasteikko (*ratio scale*) on informatiivisin asteikko. Suhdeasteikollinen mittaus on tehty välimatka-asteikolla. Lisäksi asteikolla on nollapiste, joka kuvaa mitattavan ominaisuuden täydellistä puuttumista. Jos yhden suhdeasteikolla tehdyn mittauksen tulos on x ja toisen $2x$, niin voidaan sanoa, että ensimmäisen mittauksen tilanteessa mitattavaa suuretta on kaksinkertaisesti jälkimmäisen mittauksen tilanteeseen nähden. Välimatka- muttei suhdeasteikollisesti mitatun suureen tilanteessa vastaava kuvaus ei ole mielekäs.

Juuri mikään tilastotieteellinen menetelmä ei edellytä suhdeasteikollisesti mitattuja havaintoja. Sen merkitys on tilastotieteen kannalta lähinnä tulkinnallinen edellä todetulla tavalla.

Mittausasteikoille pätee, että jos muuttuja on mitattavissa

- suhdeasteikollisesti, on se mitattavissa myös välimatka-, järjestys- ja luokka-steikollisesti.
- välimatka-asteikollisesti, on se mitattavissa myös järjestys- ja luokka-asteikollisesti.
- järjestysasteikollisesti, on se mitattavissa myös luokka-asteikollisesti.

Esimerkki. Työtunteja voidaan mitata välimatka-asteikollisesti, ja työtunneilla on 0-arvo, joka kuvaa tilannetta, jossa työtä ei tehdä lainkaan. Työtunnit ovat suhdeasteikollinen muuttuja. Jos joku teki kuukaudessa 40 tuntia työtä ja toinen 20 tuntia, ensiksi mainittu teki töitä kaksinkertaisen määrän toiseksi mainittuun verrattuna.

Celsiusasteilla on 0-arvo, joka ei kuitenkaan kuvaa lämmön täydellistä puuttumista. Jos lämpötila on 20 celsiusastetta, lämpötila ei ole kaksinkertainen verrattuna päivään, jolloin lämpötila on 10 celsiusastetta. Sitä ilmentää se, että lämpötilaa voidaan mitata fahrenheitasteilla, ja että vastaavat fahrenheitasteet ovat 68 ja 50. Celsius- ja fahrenheitasteiden 0-arvoilla lämpötila ei ole absoluuttisessa mininissään. Siksi kumpikaan asteikko ei ole suhdeasteikollinen. Kelvinasteikko olisi: 0 kelvinastetta on lämpötila, jota kylmempää ei voi olla. Edellä todetut lämpötilat ovat kelvinasteissa 293.15 ja 283.15. Edellinen lämpötila on noin 1.04-kertainen jälkimmäiseen verrattuna. □

Aina ei ole selvää, millä mitta-asteikolla suure on mitattu. Tällöin on turvallisinta valita karkeammalla mitta-asteikolla toimiva tilastotieteellinen menetelmä.

Esimerkki. Pro gradu -tutkielmia on arvosteltu ja arvostellaan erilaisilla asteikoilla. Yksi asteikko on tällainen: *approbatur*, *lubenter approbatur*, *non sine laude approbatur*, *cum laude approbatur*, *magna cum laude approbatur*, *eximia cum laude approbatur*, *laudatur*. Arvosanat ovat yksikäsitteisessä järjestyksessä. Onko asteikko myös välimatka-asteikollinen, eli ovatko arvosanat koodattavissa numeroiksi ja olisivatko numeroiden erotukset tulkittavissa arvosanojen välisiksi etäisyyksiksi? Eritoten ovatko välimatkat huonoimmasta (*approbatur*) arvosanasta seuraavaan tai toiseksi parhaasta parhaimpaan (*laudatur*) yhtä suuria

kuin muiden arvosanojen väliset etäisyydet? Joissain yliopistoissa tai oppiaineissa myönnetään huonoin tai paras arvosana hyvin harvoin. Se viittaa mahdollisuuteen, että tällaisen arvosanan saaneet tutkielmat eroaisivat muista tutkielmista enemmän kuin yhden arvosanan ero viereiseen arvosanaan ehdottaa.

Nykyään käytetään tyypillisesti arvosana-asteikkoa 1, 2, 3, 4, 5. Tällöinkään vierekkäisten arvosanojen väliset tosiasialliset etäisyydet eivät ole välttämättä kaikki samoja, vaikka arvosanat on koodattu numeroiksi. □

Mitata saatetaan, vaikka se olisi objektiivisesti mahdotonta. Suhdeasteikkokaan ei takaa mittaamisen pätevyyttä. Kyselytutkimuksilla ja ajatuskokeilla on pyritty selvittämään, kuinka paljon arvokkaampana ihmiset pitävät vaikkapa raskaana olevaa naista kuin naista. Se, että asteikkoja voi soveltaa, ei merkitse, että suure todella olisi objektiivisesti suhde-, välimatka- tai järjestysasteikollinen.

Toisaalta vaikka mittaukset olisivat epätarkkoja, tilastotiedettä voidaan usein hyödyntää hedelmällisesti. Itse asiassa tilastotiede on tarpeen nimenomaan tilanteissa, joihin liittyy jonkinlaista epätarkkuutta. Asiayhteydestä riippuu, kuinka suuri ongelma itse mittaamisen epätarkkuus on. Ylipäänsä pätee, mitä tarkemmat mittaukset, sitä luotettavammat analyysit.

Luku 3

R: Nano-opas

> Asiat, jotka on opittava tekemään, opitaan vain tekemällä niitä.[10]
>
> Aristoteles (384–322 eKr.)

Älä vain lue tätä lukua! Tottuneet R:n käyttäjät hyppäävät tästä seuraavaan lukuun. Muiden kannattaa tulkita ohje näin: Luvun oppiaines ei avaudu vain lukemalla. Luku tulee käydä läpi tietokoneen kanssa R-käskyjä toteuttaen.

R on ohjelmointikieli, joka sopii erityisen hyvin tilastotieteellisten analyysien ja kuvioiden tekemiseen. Ross Ihaka ja Robert Gentleman loivat R:n kaupallisen S-kielen vastineeksi Aucklandin yliopistossa Uudessa-Seelannissa. Nimi R kunnioittaa S:ää ja viittaa R:n luojien etukirjaimiin. S-kielen tärkeimpiä kehittäjiä oli John Chambers. Edellä mainittujen lisäksi R:ää ovat kehittäneet ja kehittävät lukuisat tutkijat ympäri maailmaa. Kehitystyötä johdetaan R-säätiöstä (*R Foundation*), joka on rekisteröity Wieniin Itävaltaan. Versio 1.0 julkaistiin helmikuussa 2000. R:stä tulee vuosittain useita päivityksiä. R on niin laajasti käytetty, että sitä voi kutsua tilastotieteen *lingua francaksi*.

Tärkeitä syitä R:n menestykseen ovat, että R:llä voi tehdä mitä erilaisimpia analyyseja ja että se on luotettava. Yksi syy R:n suosion takana on epäilemättä myös sen ilmaisuus. Se on koulutuksellisesti tärkeä näkökulma. Yliopisto-opintojen tarkoitus on kouluttaa ihmisiä opintojen jälkeistä elämää varten. R-koulutus ei mene hukkaan, sillä R:ää voivat kaikki käyttää yliopisto-opintojen jälkeenkin.

Lisätietoja R:stä on The Comprehensive R Archive Network (CRAN) -sivulla `https://www.r-project.org/`. CRAN on valtaisa tietovarantoverkosto.

3.1 R:n asennus

Tässä opastetaan asennus Windows-käyttöjärjestelmään:

- Mene sivulle `https://ftp.acc.umu.se/mirror/CRAN/`.[11]
- Klikkaa linkkiketju *Download R for Windows* → *install R for the first time* → *Download R "versionumero" for Windows* ja kuvaketta *Tallenna tiedosto*. R:n asennusohjelma imuroituu nyt tietokoneellesi (vie vähän aikaa).
- Avaa imuroitu tiedosto (löytyy selaimesi oikean yläkulman valikoista tai tiedostonhallinnasta kohdasta Ladatut tiedostot (*Downloads*). Hyväksy klikkaamalla asennus ja seuraavat esiintulevat oletusarvoiset ehdotukset ja lopussa ruksaa kuvakkeen työpöydälle ja pikakäynnistysnäppäimen luomiset. R-ohjelma asentuu.
- Tietokoneesi aloitusnäkymässä on nyt R-logo.

3.2 R:n käyttö

Merkintöjä:

- komentokehote >
- sijoitusoperaatio `<-`
- R:n palautteen *n*. alkio `[n]`
- R:n jatkamiskehote (esimerkki alempana) `+`
- kommentti (esimerkki alempana) `#`
- R:n palaute (luentomateriaalin merkintätapa) `##`
- R-palautteesta poistettua tekstiä (luentomateriaalin merkintätapa) `- - .`

R käynnistyy kaksoisklikkaamalla R-logoa. R:ää komennetaan kirjoittamalla tekstiä komentokehotemerkillä > alkavalle komentoriville ja painamalla syöttönäppäintä ↲.

Yksinkertaisimmillaan R on (symbolinen) laskin:

```
> 1+2
## [1] 3
> a <- 1
> b <- 2
> a
## [1] 1
> b
## [1] 2
> a+b
## [1] 3
```

Ensin laskettiin $1+2=3$. R:n palaute oli `[1] 3`. Siinä `[1]` osoittaa palautteen 1. alkion — tässä 3:n. Kaksi risuaitaa `##` ovat luentomateriaalin käytäntö osoittaa R:n palaute. R ei tuota tällaista merkintää.

Seuraavaksi ohjattiin lukuarvot 1 ja 2 muuttujien `a` ja `b` arvoiksi sijoitusoperaatiolla `<-`. Komennoilla `a` ja `b` katsottiin, millaisia `a` ja `b` ovat. Käskyllä `a+b` laskettiin summa ja saatiin vastaukseksi 3.

Palaute voi koostua useammasta alkiosta, jolloin R osoittaa hakasuluissa, kuinka mones palautteen alkio on rivillä ensimmäisenä. Komento `seq(1,30,1)` (*sequence*) tuottaa luvut 1:stä 30:een yhden välein (`- -` osoittaa poistettua R-palautetta):

```
> seq(1,30,1)
## [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 - -
## [26] 26 27 28 29 30
```

Palaute koostuu luvuista $1, \ldots, 30$. Toisen rivin ensimmäinen alkio on palautteen 26. alkio. Siksi R on tulostanut sen edelle `[26]`. Yleensä alkioiden järjestysnumero ei ole yhtä helposti hahmotettavissa. Silloin R:n käytäntö on avuksi.

Huom! Komentokehotetta $>$ ei osoiteta komentojen edellä jatkossa.

Havaintoja voi tuoda monella tavalla R:ään. Yhden muuttujan havainnot saa R:ään kätevästi ketjutuskomennolla `c(·)` (*concatenate*):

```
x <- c(49,35,32,39,45)
x
## [1] 49 35 32 39 45
y <- c(45,37,30,
## +
32,40)
y
## [1] 45 37 30 32 40
```

Havainnot luettiin ensin muuttujaan `x`. Sen jälkeen katsottiin, että `x` todella koostuu niistä. Toiset havainnot luettiin seuraavaksi kahdessa erässä muuttu-

jaan `y`. Yllä `+` on jatkamiskehote, joka ilmoittaa, että R odottaa komennon loppuosaa. Seuraavalla rivillä komento täydennettiin loppuun. Huom! Jatkamiskehote `+` ei suorita yhteenlaskua!

Suuret aineistot kannattaa tuoda R:ään esimerkiksi `read.table`-komennolla. Sitä ei havainnollisteta tässä.

R erottelee isot ja pienet kirjaimet:

```
X
## Error: object 'X' not found
x
## [1] 49 35 32 39 45
```

R ei tunnista `X`:ää, koska sitä ei määritelty edellä. Vain pieni `x` määriteltiin.

Huom1! Kokeile komentoja edellä! Tietokoneohjelmia oppii vain käyttämällä niitä. Huom2! Komentoja kutsutaan R:ssä funktioiksi. Ne eivät ole funktioita matemaattisessa mielessä. Selvyyden vuoksi tässä puhutaan komennoista.

R:ssä on lukuisia komentoja, joilla analysoida aineistoja. Muuttujan `x` otoskeskiarvo, otoshajonta ja muita tunnuslukuja saadaan `mean(x)`-, `sd(x)`- ja `summary(x)`-komennoilla:

```
mean(x)
## [1] 40
sd(x)
## [1] 7
summary(x)
## Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32      35      39      40      45      49
```

Otoskeskiarvo ja -hajonta ovat 40 ja 7. Aineiston pienin ja suurin havainto ovat 32 ja 49. Mediaani on 39, ja 1. ja 3. kvartiili ovat 35 ja 45.

Muuttujien `x` ja `y` otoskorrelaatio lasketaan `cor(x,y)`-komennolla:

```
cor(x,y)
## [1] 0.8725105
```

Tarkempia tietoja laskuista edellä saa komennolla `help`: `help(mean)`, `help(sd)` tai `help(summary)`. Edellä lasketut tunnusluvut selitetään jaksoissa 6.1 ja 8.2.

R operoi sujuvasti lukuisilla jakaumilla kuten normaali-, χ^2-, t- ja F-jakaumilla (luku 7). Komennot `pnorm(x)` ja `qnorm(x)` laskevat standardinormaalijakauman kertymäfunktion arvon pisteessä `x` ja vastaavan kvantiilin pisteessä `x` ($p \leftrightarrow$ *probability*; $q \leftrightarrow$ *quantile*) (jakso 6.1). Vastaavat komennot normaalijakaumalle odotusarvolla `m` ja keskihajonnalla `s` ovat `pnorm(x,m,s)` ja `qnorm(x,m,s)` (jakso 6.2). Esimerkkejä:

```
pnorm(-1.964)
## [1] 0.02476505
qnorm(0.02476505)
## [1] -1.964
pnorm(-1.964,-2,2)
## [1] 0.5071806
qnorm(0.5071806,-2,2)
## [1] -1.964
```

Huom! Normaalijakauma määritellään oppikirjoissa järjestään odotusarvon (μ) ja varianssin (σ^2) — ei keskihajonnan (σ) kuten edellä — avulla: $\mathsf{N}(\mu, \sigma^2)$.

Vastaavat komennot χ^2-, `t`- ja `F`-jakaumille ovat `pchisq(x,df)`, `pt(x,df)` ja `pf(x,df1, df2)` sekä `qchisq(x,df)`, `qt(x,df)` ja `qf(x,df1, df2)`. Niissä `df`, `df1` ja `df2` viittaavat jakauman vapausasteisiin (*degrees of freedom*).

Samalla rivillä voidaan antaa useampia komentoja ;-merkillä erotettuna:

```
pt(-1.964,20); qt(0.03179091,20)
## [1] 0.03179091
## [1] -1.964
```

3.3 Hyviä käytäntöjä

Usein komennot kannattaa ajaa tiedostosta käsin ja lopuksi tallettaa tiedosto: Valitse R:ssä *File* (näytön vasen yläkulma) ja *New script*. R:n tekstieditori avautuu. Kirjoita tarvitsemasi komennot peräkkäisille omille riveille, ja aja ne painamalla näppäinyhdistelmää Ctrl-R kunkin rivin kohdalla. Kaikki komennot saa toteutettua peräjälkeen maalaamalla (Windows-komento) komentorivit ja painamalla Ctrl-R.

Talleta tiedosto valitsemalla *File* ja *Save as*. Tiedoston voi hakea uudelleen käytettäväksi valinnoilla *File* ja *Open script*. Näin voi helposti toistaa aiemmat analyysinsa tai muokata niitä samantapaisiin uusiin tehtäviin. Jos antaa tiedoston jollekulle, hän voi toistaa ja tarkistaa analyysit.

R sivuuttaa risuaidalla `#` merkityt kommenttirivit. Ne helpottavat hahmottamista, mitä on tehty:

```
x <- c(49,35,32,39,45)
# y on laskettu Albert Neron artikkelissa ‘‘Elämän salaisuus’’:
y <- c(45,37,30,32,40)
```

Kommentoi koodiasi!

3.4 Korjaaminen ja peruminen

Virheellisen komennon voi korjata painamalla nuoli ylös -näppäintä ↑:

```
x <- c(49 35 32 39 45)
## Error: unexpected numeric constant in "x <- c(49 35"
x <- c(49,35,32,39,45)
x
## [1] 49 35 32 39 45
```

Nuolinäppäin tuo virheellisen komennon `x <- c(49 35 32 39 45)` komentoriville, josta sen voi korjata helposti lisäämällä pilkut lukujen väliin. (Tehty 3. rivillä yllä.)

R:n tekstieditorissa korjaaminen on erityisen helppoa, koska komento jää esille sen ajamisen jälkeen.

Jos R pyytää täydentämään komentoa, muttet halua tai osaa, peru toiminto poistumisnäppäimellä Esc:

```
y <- c(45,37,33,
## +
```

Jos nyt huomaat antaneesi kolmannen lukuarvon väärin (33 eikä 30), keskeytä komento painamalla Esc.

3.5 Ohjelmoinnista ja paketeista

R:n yksinkertainen käyttö ei vaadi ohjelmointitaitoja. Hyvin paljon saa tehtyä simppeleillä komennoilla tai lyhyillä komentoketjuilla.

Monia vaativampia tehtäviä varten on lisäosia eli paketteja (*package*). Niitä on jo yli 20 000 (`https://cran.r-project.org/web/packages/`). Esimerkiksi psykologeille erityisen sopivia tilastollisia menetelmiä on koottu psych-pakettiin. Se haetaan Internetistä omalle koneelle ja otetaan käyttöön komennoilla `install.packages("psych")` ja `library(psych)`. Ensimmäistä pakettia asennettaessa R kysyy, miltä palvelimelta paketti haetaan. Sopiva valinta on esimerkiksi Ruotsissa (*Sweden*) sijaitseva palvelin, jolle CRANin tietovaranto on kopioitu.

Monimutkaisiin tehtäviin löytyy usein valmiita komentojonoja R-oppaista tai Internetistä. Koodit voi kopioida ja säätää omiin tarpeisiin. Koodin tekijään tulee viitata, kun näin tekee. Esimerkiksi erikoisen kuvion piirtäminen voi olla tällainen tehtävä. Tällöinkin vaihtoehto saattaa olla ladata sopiva paketti (esim. ggplot2).

3.6 Lopettaminen ja viittaaminen

Lopeta R-istunto `quit()`- tai lyhyemmin `q()`-komennolla:

```
q()
```

Vastaa R:n esittämiin seuraaviin kysymyksiin kieltävästi. Istuntosi on sen jälkeen päättynyt.

Jos käytät opinnäytteessäsi tai muussa teoksessasi R:ää, kerro se. Ohjeen viittaamiseen saat `citation()`-komennolla:

```
citation()
## To cite R in publications use:
##
##   R Core Team (2022). R: A language and environment for statistical
##   computing. R Foundation for Statistical Computing, Vienna, Austria.
##   URL https://www.R-project.org/.
 - -
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

(Lainausmerkkejä on muokattu yllä.)

3.7 Oppaita ja neuvoja

R-opintoja voi syventää vaikkapa seuraavien opusten avulla: Braun ja Murdoch (2021), Crawley (2013), Kabacoff (2022) ja Väkeväinen (2018). Oppaita ei ole tarkoitettu luettaviksi kannesta kanteen! Oppaasta kannattaa tyypillisesti lukea vain alkuluvut, ja sen jälkeen konsultoida sitä tarpeen tullen. Oppaita kannattaa opiskella ei vain lukien vaan samalla itse koodeja kokeillen. Internetistä löytyy usein nopeasti apu mitä erilaisimpiin kysymyksiin R:n käytöstä. Internetistä löytyvät myös ohjeet helppolukuisen R-koodin kirjoittamiseen (`https://google.github.io/styleguide/Rguide.html` ja `https://style.tidyverse.org/`; haettu 5.10.2021). Oppaita ei tarvita luentomateriaalin ymmärtämiseen.

Luku 4

Todennäköisyyslaskentaa

> – – elämän tärkeimmät kysymykset – – ovat suurimmilta osin vain todennäköisyysongelmia.[12]
>
> Pierre-Simon Laplace (1749–1827)

> Elämän tärkein asia on ammatin valinta; siitä päättää sattuma.[13]
>
> Blaise Pascal (1623–1662)

> Ihmisen elämänkulussa, kuten urassa ja muissa valinnoissa, on lopulta erittäin paljon sattumaa.[14]
>
> Kari Raivio (1940–)

> Vaikka sitä omahyväisesti kuvittelee, että tämä illansuuta kohti kaartuva elämä on rakentunut omille valinnoille, yhä vastaansanomattomammaksi käy totuus: ylivoimainen osa elämästä on ollut silkan sattuman sanelemaa.[15]
>
> Pekka Sauri (1954–)

> Sattumaahan tämä kaikki on, mutta en ole harmitellut.[16]
>
> Yrjö Kukkapuro (1933–)

Laplacesta (1840/1902, 196) todennäköisyyslaskenta on pohjimmiltaan laskennaksi pelkistettyä maalaisjärkeä. Venkateshin mielestä (2013, xxvii–xviii, 3) todennäköisyyslaskenta on matematiikan aloista geometrian ohella intuitiivisin ja käytännönläheisin, ja todennäköisyyslaskennan intuitiivisuus on myös hänen opiskelijoidensa kokemus. Hand (2014, 69) kuvaa todennäköisyyslaskennan päinvastoin tunnetusti matematiikan aloista intuitionvastaisimpana. Cobbista (2015) todennäköisyyden käsite on käsistälipsuva ja intuition ja teorian välinen kuilu on matematiikan aloista suurin todennäköisyyslaskennassa. Tijms (2012,

214) viittaa näkemykseen, että millään muulla matematiikan alalla ei asiantuntija erehdy yhtä helposti. Spiegelhalterista (2011) vuosikymmenien aiheen parissa puurtamisen jälkeenkin todennäköisyyslaskenta on epäintuitiivista, vaikeaa ja jotkut sen käsitteet konstikkaita. Jos häneltä kysyy todennäköisyydestä, hän ei luota vainuunsa vaan vastaa vasta paneuduttuaan ongelmaan. Savagen (1976, 466) mukaan kukaan tässä maailmassa ei tiedä, mitä on — suurimman koskaan eläneen tilastotieteilijän Ronald Fisherin luomus — fidusiaalinen todennäköisyys. Kirjoittajasta todennäköisyyslaskenta on käytännönläheistä mutta intuitio ja todennäköisyys eivät aina kohtaa, ja se on osa aiheen viehätystä.

Todennäköisyyslaskenta on tilastollisen päättelyn kivijalka. Lisäksi todennäköisyyslaskenta on hauskaa. Ja merkityksellistä.

Esimerkki.[17] Yhdysvallat suunnitteli maailman ensimmäisen ydinpommin Manhattan-projektissa II maailmansodan aikana. Ennen ensimmäistä koeräjäytystä, jonka täytyi olla suuri, Richard Hammingiä pyydettiin tarkistamaan lasku. Hamming ajatteli tyrkkäävänsä tehtävän alaiselleen. Hamming kysyi, mitä pitäisi tarkistaa. Tutkija vastasi: Todennäköisyys, että koeräjäytys sytyttää maapallon ilmakehän. Hamming päättikin tarkistaa laskun itse. Seuraavana päivänä Hamming kertoi tutkijalle, että laskunsa ovat oikein mutta häntä askarruttavat happeen ja typpeen liittyvät kaavat, sillä niihin liittyviä kokeita ei ole mahdollista tehdä tarvittavilla energiamäärillä. Tutkija totesi, että hän pyysi Hammingia tarkistamaan laskut eikä fysiikan kaavoja ja poistui. Hamming jäi hermostuneena kävelemään edestakaisin portaita kysyen itseltään, mitä on mennyt tekemään. Hän on mukana riskeeraamassa kaikkea elämää maapallolla ymmärtämättä tarkistamansa laskun perusteita. Ystävä näki Hammingin ja kysyi, mikä häntä vaivaa. Hamming kertoi. Ystävä lohdutti: Unohda koko juttu. Kukaan ei tule koskaan syyttämään sinua. □

4.1 Otosavaruus, tapahtuma ja satunnaismuuttuja

Koe (*experiment*) on menettely, josta seuraa tulos, jonka tulosvaihtoehdot ovat määritellyt. Mielenkiinnon kohteena ovat kokeet, joiden lopputulokseen voi liittyä ja tyypillisesti liittyy satunnaisuutta. Kokeen tulos ei silloin ole välttämättä sama, jos koe uusitaan, vaikka olosuhteet olisivat muuttumattomat. Koe on tässä lavea käsite eikä viittaa esimerkiksi kokeeseen, jonka tarkoitus on vahvistaa tai saada uutta tietoa.

Kokeen *otosavaruus* (*sample space*) S on sen kaikkien mahdollisten tulos-

vaihtoehtojen eli alkeistapahtumien joukko. Monesti tulosvaihtoehtoja (a_i) on äärellinen määrä ja ne voidaan luetella: $S = \{a_1, a_2, \ldots, a_n\}$.

Tapahtuma (*event*) A on otosavaruuden osajoukko. Alkeistapahtuma-nimitys korostaa, että alkeistapahtuma on yksinkertaisin tapahtuma.

Esimerkki. Lantin heitto. Otosavaruus $S = \{kruuna, klaava\}$. Alkeistapahtumia on kaksi. □

Esimerkki. Nopan heitto. Otosavaruus $S = \{1, 2, 3, 4, 5, 6\}$. Alkeistapahtumia on kuusi. Tapahtuma A voisi olla, että nopan silmäluku on parillinen: $A = \{2, 4, 6\}$. □

Esimerkki. Kahden nopan heitto. Otosavaruus $S = \{(1,1), (1,2), \ldots, (1,5), (1,6), (2,1), (2,2), \ldots, (6,5), (6,6)\}$. Sen alkiot ovat siis:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Alkeistapahtumia on kolmekymmentäkuusi. Tapahtuma A voisi olla esimerkiksi, että molempien noppien silmäluku on sama ja parillinen: $A = \{(2,2), (4,4), (6,6)\}$. □

Satunnaismuuttuja (*random variable*) tavataan määritellä kuvaukseksi otosavaruudesta reaalilukujen joukkoon. Intuitiivisempi määritelmä (Everitt 2003) on, että satunnaismuuttuja on muuttuja, joka saa arvoja jonkin *todennäköisyysjakauman* (jakso 6.1) määräämällä tavalla. Satunnaismuuttujaa merkitään usein isolla ja sen *toteumaa* (*realization*) pienellä kirjaimella (vaikkapa X ja x).

Esimerkki. Lantin heitto (jatkoa). Kuvataan "kruuna" ykköseksi ja "klaava" nollaksi. Lantin heittoon liittyy nyt satunnaismuuttuja X, joka voi saada arvon 1 tai 0. Kun on saatu kruuna, $x = 1$. □

4.2 Todennäköisyyden määritelmiä

Todennäköisyys (*probability*) voidaan määritellä monin tavoin. Seuraavassa selitetään kolme keskeisintä määritelmää. de Finettin (1974, x) kuuluisa näkemys on, että todennäköisyyttä ei ole olemassa. Hän tarkoitti, ettei ole yhtä objektiivista todennäköisyyttä vaan on monia subjektiivisia näkemyksiä siitä. Savagen

(1972, 2) mukaan sitten Baabelin tornin on harvoin ollut täydellisempää erimielisyyttä ja kommunikaatiokatkosta kuin kysymyksistä, mitä todennäköisyys on ja kuinka se liittyy tilastotieteeseen. Nykytilanne ei ole yhtä jyrkkä. Silti edelleenkin voidaan arvioida, että todennäköisyyden käsitteestä vallitsee "massiivinen epäselvyys" (Weisberg 2014, xii). Määrittelytavasta alla riippumatta todennäköisyys noudattaa samoja laskusääntöjä. Todennäköisyyden muita tulkintoja, filosofiaa ja historiaa käsittelevät Galavotti (2013), Gillies (2000), Gorroochurn (2012, luku 14), Hacking (2006), Niiniluoto (1975) ja von Plato (1994).

4.2.1 Klassinen todennäköisyys

Olkoot otosavaruuden $S = \{a_1, a_2, \ldots, a_n\}$ kaikki tulosvaihtoehdot *symmetrisiä* eli yhtätodennäköisiä:

$$\mathsf{P}(a_i) = \frac{1}{n}.$$

Yllä on merkitty todennäköisyyttä $\mathsf{P}(\cdot)$:llä.

Määritellään tapahtuma A symmetristen tulosvaihtoehtojen avulla. Tapahtuman A *klassinen todennäköisyys* on tällöin

$$\mathsf{P}(A) = \frac{A\text{:lle suotuisten tulosvaihtoehtojen lukumäärä}}{S\text{:n tulosvaihtoehtojen lukumäärä}}. \tag{4.1}$$

Esimerkki. Pallojen poimiminen. Olkoon pussissa 10 vihreää, 20 oranssia ja 30 keltaista palloa. Poimitaan pussista sattumanvaraisesti pallo. Todennäköisyys, että saatu pallo on vihreä, on 1/6:

$$\mathsf{P}(\text{vihreä}) = \frac{10}{10 + 20 + 30} = \frac{1}{6}. \quad \square$$

Esimerkki. Kahden nopan heitto (jatkoa). Kaikki 36 silmälukuparia ovat yhtä todennäköisiä (tämä perustellaan myöhemmin). Jos $A = \{(2,2), (4,4), (6,6)\}$, niin tapahtuman A todennäköisyys on 1/12:

$$\mathsf{P}(A) = \frac{3}{36} = \frac{1}{12}. \quad \square$$

Monet tilanteet eivät sovi klassisen todennäköisyyden määritelmään. Näin käy, jos vaikkapa tulosvaihtoehdot eivät ole symmetrisiä tai tulosvaihtoehtoja on ääretön määrä, jolloin ne eivät ole lueteltavissa.

4.2.2 Frekventistinen todennäköisyys

Frekventistinen todennäköisyys tapahtumalle A määritellään todennäköisyyden

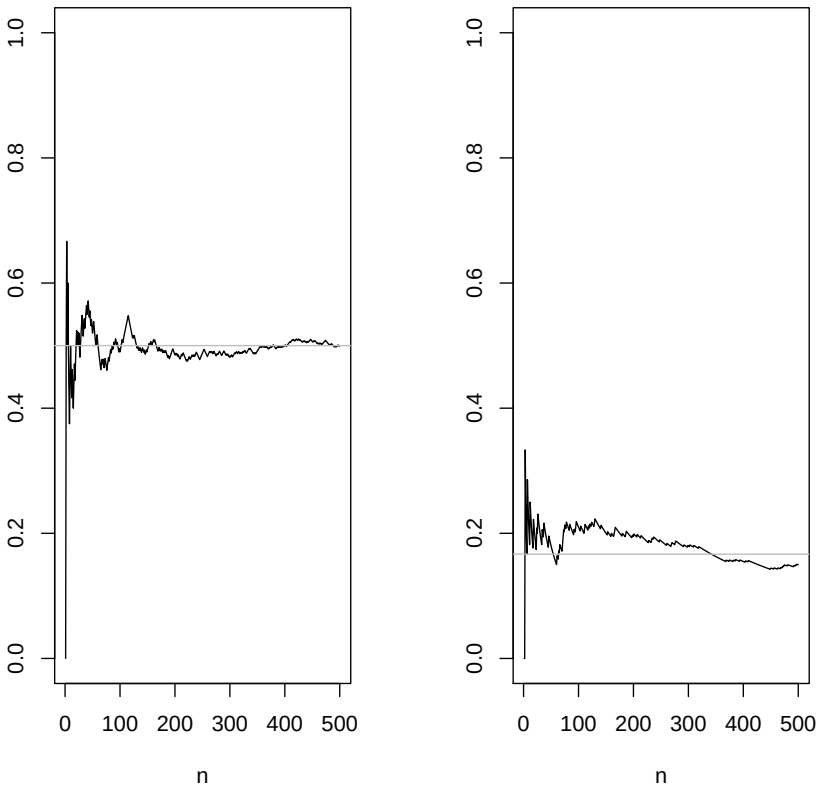
$$\mathsf{P}_n(A) = \frac{f_n}{n} \tag{4.2}$$

raja-arvona. Yllä f_n on tapahtuman A frekvenssi ja $\mathsf{P}_n(A)$ tapahtuman A havaittu todennäköisyys n:ssä riippumattomasti tehdyssä satunnaiskokeessa. Tapahtuman A frekventistinen todennäköisyys seuraa antamalla toistokokeiden lukumäärän kasvaa kohti ääretöntä. Tällaisen raja-arvon laskun oletuksena on, että satunnaiskoetta voidaan toistaa samanlaisissa olosuhteissa rajatta ja kokeet ovat riippumattomia.

Esimerkki. Pallojen poimiminen (jatkoa). Todennäköisyys vihreän pallon nostamiselle on frekventistisen näkemyksen mukaan vihreiden pallojen nostojen määrä f_n jaettuna kaikkien poimintojen määrällä n, kun se kasvaa kohti ääretöntä (olettaen, että nostettu pallo palautetaan pussiin ja pallot sekoitetaan kunkin poiminnan jälkeen). Sadan poiminnan jälkeen $f_{100} = 20$ ja $\mathsf{P}_{100}(A) = 20/100 = 1/5$. Kahdensadan poiminnan kohdalla $f_{200} = 35$ ja $\mathsf{P}_{200}(A) = 35/200 = 0.175$. Kun poimintoja on 1 000, $f_{1000} = 160$ ja $\mathsf{P}_{1000}(A) = 160/1000 = 0.16 \approx 1/6$. Vihreiden pallojen osuus menee suurilla havaintomäärillä kohti 1/6:tta. (Teoreettinen perustelu esitetään myöhemmin.) □

Esimerkki. Lantin heitto ja pallojen poimiminen (jatkoa). R-ohjelmistolla simuloitiin lantin heittämistä ja pallon poimimista. Kuvassa 4.1 vasemmalla havainnollistetaan kruunujen suhteellisen frekvenssin (f_n/n yhtälössä (4.2)) kehittymistä, kun heittojen lukumäärä $n = 1, \ldots, 500$. Oikealla kuvassa ovat vastaavat vihreän pallon poimimisen suhteelliset frekvenssit. Osuudet näyttävät suppenevan kohti 1/2:hta ja 1/6:tta. Harjoitustehtävässä lasketaan eri n:n arvoilla todennäköisyyksiä, joilla osuudet ovat lähellä 1/2:hta tai 1/6:tta. □

Todennäköisyyttä ei voida määritellä frekventistisesti, jos kyseessä on ainutkertainen tapahtuma. Tällaisiksi tapahtumiksi voisi mieltää esimerkiksi alkuräjähdyksen tai että tietty henkilö päättää ostaa maitoa kaupasta tiettynä päivämääränä tiettynä kellonaikana. Kaava yllä ei anna myöskään vastausta sellaisiin kysymyksiin kuin, mikä on todennäköisyys, että Jumala on olemassa.



Kuva 4.1: Kruunujen ja vihreiden pallojen osuuden kehitykset lantin heitossa ja pallon poiminnassa.

4.2.3 Subjektiivinen todennäköisyys

Ihmisillä on eri mielipiteitä ja niin myös todennäköisyyksistä. Näitä yksilöllisiä todennäköisyyksiä — tai uskomuksen asteita — kutsutaan *subjektiivisiksi todennäköisyyksiksi*. Ne ovat subjektiivisia, koska ihmiset saattavat arvioida todennäköisyyksiä erilailla samojenkin tietojen perusteella.

Esimerkki. Talouden asiantuntijat esittivät Yle Uutisissa 15.5.2012[18] 0.50:n ja 0.80:n välillä olevia arvioita todennäköisyydestä, jolla Kreikka eroaa eurosta (uutisessa todennäköisyydet on esitetty prosentteina 0:n ja 100:n välillä):

> Johtavat talousasiantuntijat pitävät valtiovarainministeriön arviota Kreikasta mahdollisesti aiheutuvista tappioista liian optimistisena. Ministeriö arvioi viime viikolla, että tappioita tulisi Suomelle korkeintaan 400 miljoonaa euroa. – – Kaikki asiantuntijat pitivät todennäköisenä, että Kreikka eroaa eurosta. Arviot eron todennäköisyydestä vaihtelevat 50–80 prosentin välillä.
>
> Pasi Holm, PTT:n toimitusjohtaja
> Saattaa olla, että kahdenvälisen lainan tappiot ovat 500 miljoonaa euroa ja väliaikaisen kriisirahaston kautta 140 miljoonaa euroa. Lisäksi EKP:n ja eurojärjestelmän kautta joitakin satoja miljoonia euroa. – Kreikan eurosta eroamisen todennäköisyys: 70 prosenttia.
>
> Seija Ilmakunnas, Palkansaajien tutkimuslaitoksen johtaja
> Pidän mahdollisena, että koko kahdenvälinen laina (miljardi euroa) menetetään. ERVV-osuuden riski on pienempi, mutta sitä ei voi arvioida tuntematta vakuusjärjestelyn yksityiskohtia. – Kreikan eurosta eroamisen todennäköisyys: 50 prosenttia.
>
> Vesa Kanniainen, Kansantaloustieteen professori
> Kokonaistappio 2–3 miljardia euroa, sisältäen lainat ja takuut sekä eurojärjestelmän kautta syntyvät pääomatappiot. – Kreikan eurosta eroamisen todennäköisyys: 75 prosenttia.
>
> Vesa Vihriälä, Etlan toimitusjohtaja
> Maksimitappio voisi olla 1.84 miljardia euroa. Tappio jäänee tätä pienemmäksi, mutta ei välttämättä VM:n arvioimaan 400 miljoonaan. Lisäksi Suomen Pankille voi tulla tappioita siitä, jos Kreikan keskuspankki ei pysty vastaamaan veloistaan EKP:lle. – – Kreikan eurosta eroamisen todennäköisyys: 80 prosenttia.

Todennäköisyyden frekventistinen tulkinta ei ole luonteva Kreikan eurosta eroamisen mahdollisuutta arvioitaessa. Yksikään maa ei ole aiemmin eronnut eurosta, eikä eroon johtavista seikoista ole kaikkea tarvittavaa tietoa. Asiantuntijat kertovat omista lähtökohdistaan eri suuruisen arvion eron todennäköisyydestä. Ne ovat heidän subjektiivisia arvioitaan Kreikan eurosta eroamisen todennäköisyydestä. □

Subjektiivinen todennäköisyys voi olla mahdollista selvittää, vaikka ihminen ei sitä kertoisi tai osaisi kertoa yhtä avoimesti kuin esimerkissä edellä. Yksilön subjektiivinen todennäköisyys määritellään joskus vetokertoimen (*vastasuhde*, *odds*)

$$\frac{S}{V} = \frac{\mathsf{P}(A)}{1 - \mathsf{P}(A)}, \tag{4.3}$$

avulla. Yllä $S > 0$ ja $V > 0$ (euroa) ovat sijoitus (mahdollisesti tappio) ja voitto vedonlyönnissä tapahtumasta A ja $\mathsf{P}(A) \in (0, 1)$ on yksilön subjektiivinen todennäköisyys tapahtumalle A. ("$\in$" luetaan "kuuluu joukkoon" tai tässä yhteydessä "kuuluu välille".) Yhtäsuuruus perustellaan alla.

Idea on, että yksilö ilmaisee subjektiivisen todennäköisyytensä reilussa (*fair*) vedonlyönnissä. Oletetaan, että pelaaja sijoittaa vedonlyöntiin S euroa, voittaa V euroa, jos A tapahtuu mutta häviää sijoituksensa S euroa, jos A:ta ei tapahdu. Sille todennäköisyys on pelaajan mielestä $1-\mathsf{P}(A)$. Tällainen vedonlyönti on reilu, jos pelaajan mielestä hän ei odotetun tuoton (jakso 6.3) mielessä hyödy vedonlyönnistä eli odotettu tuotto on 0 euroa:

$$\mathsf{P}(A) \times V + [1 - \mathsf{P}(A)] \times (-S) = 0.$$

Jaetaan yhtälö puolittain V:llä ja todetaan, että yhtäsuuruuden (4.3) täytyy olla voimassa, jotta yhtälö pätisi:

$$\mathsf{P}(A) \times 1 - [1 - \mathsf{P}(A)] \times \frac{S}{V} = 0.$$

Pelaajan subjektiivinen todennäköisyys tapahtumalle A saadaan ratkaisemalla yhtäsuuruus (4.3) $\mathrm{P}(A)$:n suhteen:

$$\mathsf{P}(A) = \frac{S}{S + V}.$$

Subjektiivinen todennäköisyys voidaan siten selvittää, kun tiedetään vetokerroin pelaajan mielestä reilussa vedonlyönnissä.

Esimerkki. Tentin läpäisy. Opiskelija on varsin vakuuttunut, että hän läpäisee tentin. Hän on valmis lyömään siitä vetoa vetokertoimella $S/V = 4/1$. Hänen subjektiivinen todennäköisyytensä tentin läpäisemiselle on tällöin $S/(S + V) = 4/(4 + 1) = 0.8$. □

Vedonlyöntiasetelma on keinotekoinen: Moni ei suostuisi uhkapeliin, jonka odotettu tuotto on nolla. Joku ei ryhtyisi uhkapeliin koskaan. Asetelma ei toimi monen mielestä tärkeimmän kysymyksen kohdalla, jatkuuko elämä kuoleman jälkeen. Uskomuksesta riippumatta kannattaisi aina lyödä tuonpuoleisen elämän puolesta vetoa, koska veto ratkeaisi ainoastaan elämän jatkuessa ja ainoastaan silloin olisi mahdollista kerätä vedon tuotto. Asiasta voisi huoletta lyödä vetoa mielivaltaisen suurella vetokertoimella, mikä edellä olevan laskun mukaan merkitsisi uskoa tuonpuoleiseen elämään oleellisesti todennäköisyydellä 1. (Mellor

1973.) Johto edellä on silti tavanomainen subjektiivisen todennäköisyyden yhteydessä. Ajatus lienee, että "pakottamalla" yksilö vedonlyöntiin, hänen subjektiivinen todennäköisyytensä määrittyy. Vetokerroin S/V on helpohko hahmottaa, ja johto yhdistää sen kätevästi subjektiivisen todennäköisyyden suuruuteen. Idea juontaa Thomas Bayesin postuumiin (Richard Pricen esitelmöimään) tutkimukseen 1763 asti.

Muitakin menetelmiä selvittää subjektiivinen todennäköisyys on. Pähkinänkuoriesitys on Haighin (2012) kirjassa.

Todennäköisyyden arviointi voi olla vaikeaa. Joskus yksilö osaa asettaa tapahtumat järjestykseen todennäköisyyden mukaan (tapahtuma on toista todennäköisempi) muttei pysty nimeämään niille (subjektiivisiakaan) todennäköisyyksiä tai vedonlyöntivalmiuttaan. Joskus voi olla vaikea asettaa tapahtumia edes todennäköisyysjärjestykseen. Tällaisissa tilanteissa subjektiivista todennäköisyyttä ei voi määritellä.

Esimerkki. Subjektiivinen todennäköisyys voi olla hyvin huono arvio.[19] Koronapandemia alkoi Wuhanin kaupungista Kiinassa joulukuussa 2019. Suomalainen johtava lääkäri arvioi 24.1.2020, että todennäköisyys olisi "varmasti yksi miljoonasta", että Ivalon terveyskeskukseen hakeutuneilla Wuhanista tulleilla kiinalaisturisteilla olisi koronavirus. Arviota uutisoitiin laajasti. Meni viisi päivää, ja Wuhanista tulleella kiinalaisturistilla todettiin koronavirus Lapin keskussairaalassa. □

Esimerkki. Todennäköisyyttä ei osata arvioida. Suomen Pankin ekonomisti katsoo, että Yhdysvaltain ja Kiinan kauppasodan pahimman skenaarion todennäköisyyttä on mahdoton arvioida. Helsingin Sanomat 3.10.2019:[20]

> Suomen Pankki julkaisi – – uuden riskiarvonsa siitä, kuinka pahasti kauppasodan kärjistyminen voisi vahingoittaa euroalueen taloutta. Pahimmassa tapauksessa euroalueen talouskasvu voisi hidastua yli 2.5 prosenttiyksikköä eli edessä olisi vaikea taantuma. – – "On mahdotonta arvioida, mikä olisi tämän riskiarvion toteutumisen todennäköisyys. Rahoitusmarkkinoiden häiriöt ovat aina yllättäviä ja niiden laajuutta on erittäin vaikea ennakoida", sanoo Suomen Pankin Rahapolitiikka ja kansainvälinen talous -toimiston päällikkö Hanna Freystätter.

Työeläkevakuutusyhtiö Varman toimitusjohtaja Risto Murto toteaa vastaavasti Venäjän ja Ukrainan sodan laajentumisen mahdollisuudesta. Arvopaperi, tammikuu 2022 (s. 31):

> Sota voi yllättää kahdella tapaa: laajentumisella tai rauhoittumisella. Kykymme laskea näille [vaihtoehdoille] järkeviä todennäköisyyksiä on kuitenkin todella pieni.

□

4.3 Joukko-oppia

Tapahtumat koostuvat otosavaruuden S osajoukoista. Olkoot A ja B kaksi tapahtumaa (osajoukkoa) siinä.

Tapahtumien A ja B yhdiste (unioni) koostuu niistä tulosvaihtoehdoista (alkioista), jotka kuuluvat joko A:han tai B:hen tai molempiin. A:n ja B:n yhdistettä merkitään $A \cup B$.

Tapahtumien A ja B leikkaus koostuu niistä tulosvaihtoehdoista, jotka kuuluvat sekä A:han että B:hen. A:n ja B:n leikkausta merkitään $A \cap B$.

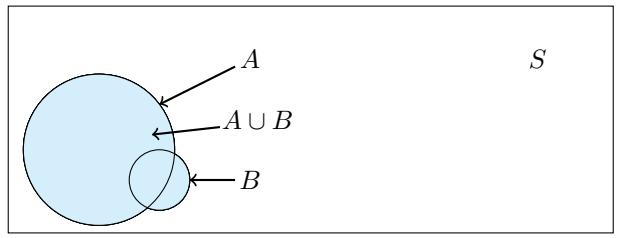
Jos A tapahtuu aina, kun B tapahtuu, B:n määrittävät tulosvaihtoehdot ovat A:n määrittävien tulosvaihtoehtojen osajoukko. Tällöin merkitään $B \subseteq A$ tai $B \subset A$, jos A ja B eivät ole samoja tapahtumia. Jälkimmäisessä tapauksessa B:n sanotaan olevan A:n aito osajoukko.

Jos $A \cap B = \emptyset$ (tyhjä joukko), niin A ja B ovat *erillisiä* eli toisensa poissulkevia (*disjoint, mutually exclusive*) tapahtumia.

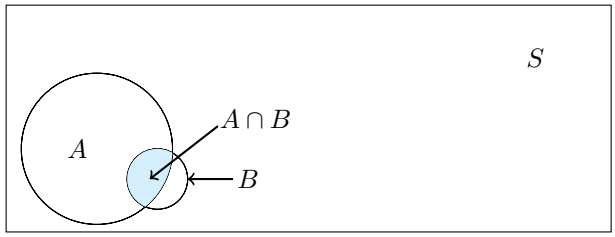
Tapahtuman A *vastatapahtuma* (*complement event*) muodostuu niistä tulosvaihtoehdoista, jotka eivät kuulu A:han. A:n vastatapahtumaa merkitään A^C:lla.

Tapahtumia voi olla useampia — vaikkapa kolme: A, B ja C.

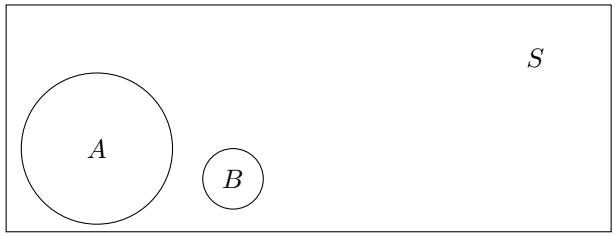
Joukko-opillisia operaatioita havainnollistetaan usein *Venn-diagrammeilla* (kuvat 4.2–4.9):



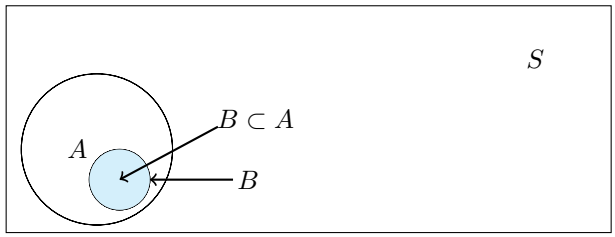
Kuva 4.2: A:n ja B:n yhdiste.



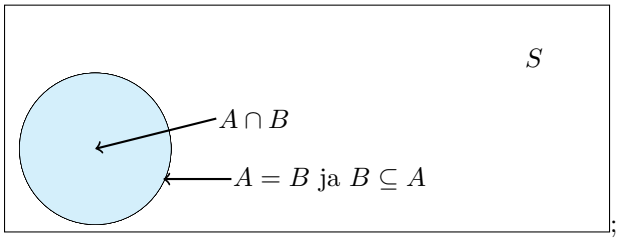
Kuva 4.3: A:n ja B:n leikkaus.



Kuva 4.4: A ja B ovat erillisiä.

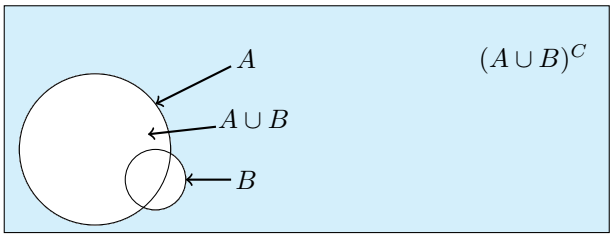


Kuva 4.5: Kaikki B:n tapahtumat ovat myös A:n tapahtumia.

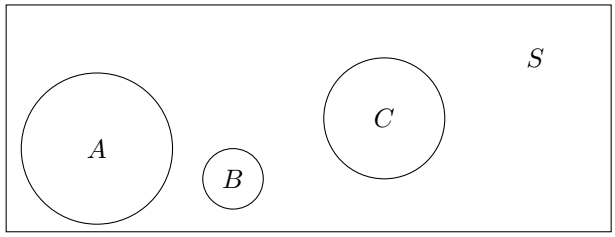


Kuva 4.6: A ja B ovat sama tapahtuma.

Kuva 4.7: A:n vastatapahtuma



Kuva 4.8: A:n ja B:n yhdisteen vastatapahtuma.



Kuva 4.9: A, B ja C ovat erillisiä.

4.4 Todennäköisyyslaskennan laskusääntöjä

Kuulukoon tapahtuma A otosavaruuteen S. Tapahtuman A todennäköisyys on vähintään 0:

$$\mathsf{P}(A) \geq 0.$$

Todennäköisyys, että jokin otosavaruuden tapahtumista tapahtuu, on 1:

$$\mathsf{P}(S) = 1.$$

Olkoot A ja B otosavaruuteen kuuluvia erillisiä tapahtumia ($A \cap B = \emptyset$). Niiden yhdisteen todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B).$$

Edellisistä oletuksista voidaan johtaa todennäköisyyslaskennan laskusäännöt alla.[21]

Minkä tahansa tapahtuman A todennäköisyys on korkeintaan 1:

$$\mathsf{P}(A) \leq 1.$$

A:n vastatapahtuman todennäköisyys on

$$\mathsf{P}(A^C) = 1 - \mathsf{P}(A).$$

Otosavaruuteen kuulumattoman eli loogisesti mahdottoman tapahtuman todennäköisyys on 0:

$$\mathsf{P}(\emptyset) = 0.$$

Jos $B \subseteq A$, niin

$$\mathsf{P}(B) \leq \mathsf{P}(A).$$

Olkoot A ja B erillisiä tapahtumia. *Erillisten tapahtumien yhteenlaskusäännön* mukaan tapahtumien A ja B yhdisteen todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B). \tag{4.4}$$

Sääntö yleistyy suoraviivaisesti. Olkoot tapahtumat $A_1, \ldots, A_n$, erillisiä ($A_i \cap A_j = \emptyset$ kaikille $i \neq j$). Tällöin yhdistetyn tapahtuman $\cup_{i=1}^{n} A_i$ todennäköisyys on

$$\mathsf{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathsf{P}(A_i). \tag{4.5}$$

Yllä $\cup_{i=1}^{n} A_i$ on kaikkien tapahtumien $A_1, \ldots, A_n$ yhdiste $A_1 \cup \cdots \cup A_n$. Summamerkintä $\sum_{i=1}^{n}$ tarkoittaa, että sen oikealla puolella olevassa termissä indeksi i saa arvot 1:stä n:ään ja että termit lasketaan yhteen: $\sum_{i=1}^{n} \mathsf{P}(A_i) = \mathsf{P}(A_1) + \cdots + \mathsf{P}(A_n)$. Ellipsis (… tai $\cdots$) kuvaa esiinkirjoittamattomia termejä. Ne on pääteltävä yhteydestä.

Yleistä yhteenlaskusääntöä käytetään, jos tapahtumat A ja B eivät ole erillisiä ($A \cap B \neq \emptyset$). Sen mukaan tapahtumien yhdisteen todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B). \tag{4.6}$$

Yleinen yhteenlaskusääntö n:n tapahtuman $A_1, \ldots, A_n$ tilanteessa[22] on

$$\mathsf{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i=1}^{j} \sum_{j=1+1}^{n} \mathsf{P}(A_i \cap A_j) + \sum_{i=1}^{j} \sum_{j=1+1}^{k} \sum_{k=j+1}^{n} \mathsf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathsf{P}(A_1 \cap \cdots \cap A_n).$$

Kun tapahtumia on kaksi (A ja B), kaava typistyy säännöksi (4.6). Elfving (1956, 30–31) todistaa tuloksen induktioperiaatteella.

Säännöt ovat intuitiiviset. Ympyröiden kokoja kuvissa 4.2–4.9 voi ajatella tapahtumien todennäköisyyksinä, jos laatikon koko on yksi. Mahdottoman (todennäköisyyden mielessä) tapahtuman todennäköisyys on nolla ja varman yksi. Muunlaisen tapahtuman todennäköisyys on nollan ja yhden välillä. Erillisten tapahtumien A ja B yhdiste vastaa tapahtumaa, että jompikumpi niistä tapahtuu. Tällaisen yhdistetyn tapahtuman todennäköisyys on erillisten tapahtumien todennäköisyyksien summa (kuva 4.4). Tapahtuma A ja sen vastatapahtuma A^C ovat erillisiä ja kattavat koko otosavaruuden, joten niiden todennäköisyyksien summa on yksi (kuva 4.7). Kuva 4.2 selventää, että tapahtumista, jotka eivät ole erillisiä, muodostetun yhdistetyn tapahtuman todennäköisyyttä ei voi laskea suoraviivaisesti summaamalla tapahtumien todennäköisyyksiä. Summasta pitää vähentää todennäköisyys tulosvaihtoehdolle “molemmat tapahtuvat”. Sen todennäköisyys tulisi muuten ynnättyä kahdesti (kuvat 4.2 ja 4.3). Kuvat 4.5 ja 4.6 havainnollistavat, että $\mathsf{P}(B) \leq \mathsf{P}(A)$, jos A tapahtuu aina, kun B tapahtuu. Jos erillisiä tapahtumia on kolme A, B ja C (kuva 4.9), niiden yhdisteen $A \cup B \cup C$ todennäköisyys saadaan summaamalla kunkin tapahtuman todennäköisyydet kaavan (4.5) mukaisesti.

Esimerkki. Nopan heitto (jatkoa). Silmäluku 3.5 ei kuulu otosavaruuteen, ja sen todennäköisyys on 0. Varmasti eli todennäköisyydellä 1 saadaan jokin otosavaruuden $\{1, 2, 3, 4, 5, 6\}$ muodostavista silmäluvuista eli alkeistapahtumista $\{1\}, \ldots, \{6\}$. Kunkin otosavaruuteen kuuluvan silmäluvun todennäköisyys on $0 < 1/6 < 1$. Kukin silmäluku on erillinen tapahtuma, sillä kahta silmälukua ei voi tulla samanaikaisesti.

Olkoon tapahtuma $A = \{2, 4, 6\}$ ja tapahtuma $B = \{1\}$. Tapahtumat ovat erillisiä. Yhdistetyn tapahtuman $A \cup B = \{1, 2, 4, 6\}$ todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) = \frac{3}{6} + \frac{1}{6} = \frac{2}{3}.$$

Vastatapahtuman $(A \cup B)^C = \{3, 5\}$ todennäköisyys voidaan laskea summana erillisten alkeistapahtumien todennäköisyyksistä

$$\mathsf{P}(A \cup B)^C = \frac{2}{6} = \frac{1}{3}$$

tai säännön

$$\mathsf{P}((A \cup B)^C) = 1 - \mathsf{P}(A \cup B) = 1 - \frac{2}{3} = \frac{1}{3}$$

avulla (vrt. kuva 4.8).

Olkoon $B = \{2\}$. Nyt A ja B eivät ole erillisiä: $A \cap B = \{2\}$. $\mathsf{P}(B) = \mathsf{P}(A \cap B) = 1/6$. Yhdistetyn tapahtuman $A \cup B = \{2, 4, 6\}$ todennäköisyys on

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B) = \frac{3}{6} + \frac{1}{6} - \frac{1}{6} = \frac{1}{2}.$$

Se voitaisiin laskea myös vertaamalla (yhtätodennäköisten) suotuisten tulosvaihtoehtojen lukumäärää kaikkien tulosvaihtoehtojen lukumäärään:

$$\mathsf{P}(A \cup B) = \frac{3}{6} = \frac{1}{2}.$$

Olkoon $B = \{1, 2\}$. A ja B eivät ole erillisiä: $A \cap B = \{2\}$. $\mathsf{P}(B) = 1/3$ ja $\mathsf{P}(A \cap B) = 1/6$. Yhdistetyn tapahtuman $A \cup B = \{1, 2, 4, 6\}$ todennäköisyys on

$$\mathsf{P}(A \cup B) = \frac{3}{6} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}. \quad \square$$

Esimerkki. Nopan heitto (jatkoa). Todennäköisyyslaskut nopan heitosta kuvaavat yhtälailla mitä tahansa koetta, jossa otosavaruus koostuu kuudesta yhtä todennäköisestä erillisestä tulosvaihtoehdosta. Sellaisia voisivat olla

- yhden tai useamman työntekijän valinta kuuden tasaveroisen hakijan joukosta.
- äänestettävän kansanedustajaehdokkaan valinta kuuden tasavahvan ehdokkaan joukosta. Ehdokkaisiin voitaisiin taas liittää ominaisuuksia silmälukujen tilalle kuvaamaan puoluetta, kantaa tiettyyn poliittiseen kysymykseen jne.
- sivuaineiden valinta yliopistossa kuuden yhtäkiinnostavan vaihtoehdon joukosta. (Niiden sisältö voi olla opiskelijan näkökulmasta tuntematon ja satunnainen.)
- ylipäänsä yhden tai useamman asian, esineen tai ihmisen valinta kuuden tasavahvan vaihtoehdon joukosta.

Nopan heitto -laskut yleistyvät tilanteisiin, joissa tasavahvoja erillisiä vaihtoehtoja on kuudesta poikkeava määrä. Laskujen numeeriset tulokset eivät ole samat mutta periaatteet ovat. □

4.5 Ehdollinen todennäköisyys ja riippumattomuus

Olkoot A ja B tapahtumia otosavaruudessa ja $\mathsf{P}(B) > 0$ (mahdottomalle tapahtumalle ei voi ehdollistaa). A:n *ehdollinen todennäköisyys* ehdolla B on

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}. \tag{4.7}$$

Merkinnän " $\mid$ " vasemmalla puolella on tapahtuma, jonka todennäköisyyttä määritellään ja oikealla puolella tapahtuma, jolle ehdollistetaan. Kaava on tärkeä. Siitä johdetaan myöhemmin tulosääntö (4.9) ja riippumattomuusehdot (4.10) ja (4.11).

Venn-diagrammit avittavat taas ymmärtämistä. Kuvassa 4.3 osajoukko $A \cap B$ on pieni osuus S:stä eli $\mathsf{P}(A \cap B)$ on pieni. Suhteutettuna osajoukon B kokoon $A \cap B$ on silti suurehko ja siten $\mathsf{P}(A \mid B)$:kin on. A tapahtuu peräti aina kuvassa 4.5, jos B tapahtuu. Tällöin $\mathsf{P}(A \mid B) = 1$, vaikka $\mathsf{P}(A)$ olisi pieni. A:n ehdollinen todennäköisyys on 1 myös kuvan 4.6 tilanteessa, jossa A ja B ovat samat tapahtumat. Kun ehdollistetaan tapahtumalle B, otosavaruus S korvautuu B:llä.

B:n todennäköisyys ehdolla A on vastaavasti

$$\mathsf{P}(B \mid A) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(A)}. \tag{4.8}$$

Esimerkki. Nopan heitto (jatkoa). Olkoon $A = \{2, 4, 6\}$ ja $B = \{1\}$. Tällöin $A \cap B = \emptyset$, jonka todennäköisyys on nolla. Ehdollistaminen romauttaa todennäköisyyden:

$$\mathsf{P}(A) = \frac{1}{2}$$

mutta

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{0}{1/6} = 0$$

(kuva 4.4). Jos $B = \{2, 4\}$, niin $A \cap B = \{2, 4\}$. Nyt ehdollistaminen räjäyttää todennäköisyyden:

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{1/3}{1/3} = 1$$

(kuva 4.5). □

Esimerkki. Myös subjektiiviset todennäköisyydet voivat muuttua paljon ehdollistavan informaation muuttuessa. Yle Uutiset 1.3.2017:[23]

> Yle Uutiset haastatteli keväällä 2012 neljää talouden asiantuntijaa Kreikan kriisistä. – – Nyt haastattelimme samat asiantuntijat uudelleen. – – Maailma on muuttunut, niin myös tutkijan käsitys. — Kreikan rooli pakolaiskriisin hoitamisessa on ollut iso. Painostus euroeroon ei tunnu enää niin soveliaalta kuin vielä pari vuotta sitten, Ilmakunnas arvioi. Muita syitä Ilmakunnas löytää Kreikan omasta kannasta — helleenit haluavat tätä nykyä pysyä rahaliitossa – – . – – Ilmakunnaksen mukaan euroalue haluaa ylläpitää mielikuvaa rahaliiton peruuttamattomuudesta. – – Kanniainen päätyy pienen pohdinnan jälkeen fifty-fifty -arvioon. – – Kreikan euroeron todennäköisyys on noin 10 prosenttia viiden vuoden aikajaksolla, tutkimusjohtaja Pasi Holm sanoo.

Asiantuntijat esittivät 15.5.2012 suurempia arvioita todennäköisyydestä, jolla Kreikka eroaa eurosta (jakso 4.2.3). Ilmakunnaksen arvio on muuttunut 0.50:tä pienemmäksi, Kanniaisen 0.75:stä 0.50:een ja Holmin 0.70:stä 0.10:een. □

Esimerkki.[24] Olkoon todennäköisyys

- 0.2, että A tapahtuu mutta B ei tapahdu ($A \cap B^C$).
- 0.1, että B tapahtuu mutta A ei tapahdu ($B \cap A^C$).

- 0.6, että kumpikaan ei tapahdu ($(A \cup B)^C$). Mikä on todennäköisyys, että A tapahtuu, jos B tapahtuu eli ehdollinen todennäköisyys $\mathsf{P}(A \mid B)$?

  Merkitään

$$\mathsf{P}(\text{kumpikaan ei tapahdu}) = \mathsf{P}((A \cup B)^C) = 0.6.$$

Toisaalta

$$\mathsf{P}(A \cup B) = 1 - 0.6 = 0.4 = \mathsf{P}(A \cap B^C) + \mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C)$$

(kuva 4.10). Näin ollen

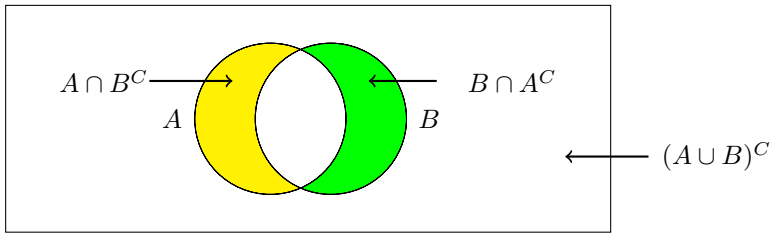
$$\mathsf{P}(A \cap B) = 0.4 - 0.2 - 0.1 = 0.1$$

ja

$$\mathsf{P}(B) = \mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C) = 0.1 + 0.1 = 0.2$$

(kuva 4.10). Kysytty ehdollinen todennäköisyys on

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C)} = \frac{0.1}{0.2} = 0.5. \ \square$$



Kuva 4.10: $\mathsf{P}(A \cup B) = \mathsf{P}(A \cap B^C) + \mathsf{P}(A \cap B) + \mathsf{P}(B \cap A^C)$.

Joskus sanotaan, että B myötävaikuttaa A:han, jos

$$\mathsf{P}(A \mid B) > \mathsf{P}(A).$$

Vastaavasti saatetaan sanoa, että B estää A:ta, jos

$$\mathsf{P}(A \mid B) < \mathsf{P}(A).$$

Verbejä “myötävaikuttaa” ja “estää” ei tule tulkita kirjaimellisesti. Tapahtumien välillä ei tarvitse olla syy-seuraussuhdetta, vaikka tapahtuman A todennäköisyys muuttuu B:n tapahtumisen myötä. Ylipäänsäkään todennäköisyyksien laskuun ei tule liittää ajatusta kausaliteetista ellei ole muuta tietoa siitä.

Esimerkki. Harrastukset. Tapahtuma A on “harrastaa innokkaasti videopelejä” ja C on “kuntoilee aktiivisesti”.[25] Innokkaat videopelaajat ovat ehkä keskimääräistä harvemmin aktiivisia kuntoilijoita ja päinvastoin. Tällöin $\mathsf{P}(A \mid C) < \mathsf{P}(A)$ ja $\mathsf{P}(C \mid A) < \mathsf{P}(C)$. Kuntoilu voi vaikuttaa videopelaamisen todennäköisyyteen muttei välttämättä vaikuta esimerkiksi haluun videopelata saati estä sitä. Biologiset taipumukset tai vanhempien ohjaus jommankumman harrastuksen pariin voivat olla syitä harrastukseen, eivätkä harrastukset sinällään vaikuta toisiinsa. □

Kaavoista (4.7) ja (4.8) seuraa *tulosääntö*

$$\mathsf{P}(A \cap B) = \mathsf{P}(A \mid B) \times \mathsf{P}(B) = \mathsf{P}(B \mid A) \times \mathsf{P}(A). \tag{4.9}$$

Sille on paljon käyttöä.

Esimerkki. Pallojen poimiminen (jatkoa). Pussissa on 10 vihreää, 20 oranssia ja 30 keltaista palloa. Poimitaan umpimähkäisesti pussista pallo, jota ei palauteta pussiin, ja ongitaan sen jälkeen pussista umpimähkäisesti toinen pallo. Mikä on todennäköisyys, että 1. pallo on vihreä ja 2. pallo keltainen? Entä todennäköisyys, että molemmat pallot ovat vihreitä?

Todennäköisyys saada ensin vihreä pallo on $\mathsf{P}(\text{1. pallo vihreä}) = 1/6$. Tämän tapahtuman jälkeen pussissa on 9 vihreää, 20 oranssia ja 30 keltaista palloa. Todennäköisyys saada seuraavaksi keltainen pallo on

$$\mathsf{P}(\text{2. pallo keltainen} \mid \text{1. pallo vihreä}) = \frac{30}{9 + 20 + 30} = \frac{30}{59}.$$

Todennäköisyys saada ensin vihreä ja sitten keltainen pallo on

$$\begin{aligned}&\mathsf{P}(\text{1. pallo vihreä ja 2. pallo keltainen})\\ &\quad = \mathsf{P}(\text{2. pallo keltainen} \mid \text{1. pallo vihreä}) \times \mathsf{P}(\text{1. pallo vihreä})\\ &\quad = \frac{30}{59} \times \frac{1}{6} = \frac{5}{59} \approx 0.085.\end{aligned}$$

Todennäköisyys saada kaksi vihreää palloa on

$$\mathsf{P}(\text{1. ja 2. pallo vihreä})$$

$$= \mathsf{P}(\text{2. pallo vihreä} \mid \text{1. pallo vihreä}) \times \mathsf{P}(\text{1. pallo vihreä})$$
$$= \frac{9}{59} \times \frac{1}{6} = \frac{3}{118} \approx 0.025. \ \square$$

Tapahtumat A ja B ovat *riippumattomia* (*independent*), jos

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \times \mathsf{P}(B). \tag{4.10}$$

Tällöin

$$\mathsf{P}(A \mid B) = \mathsf{P}(A) \tag{4.11}$$

kaavassa (4.9), jos $\mathsf{P}(B) > 0$. Riippumattomien tapahtumien todennäköisyyteen ei vaikuta, onko toista tapahtunut vai ei. Jos A on mahdoton tapahtuma ($\mathsf{P}(A) = 0$), riippumattomuusehto (4.10) toteutuu. Mahdollinen tapahtuma on siten riippumaton mahdottomasta tapahtumasta. Vastaavasti mahdollinen tapahtuma on riippumaton myös varmasta tapahtumasta ($\mathsf{P}(A) = 1$).

Kaava (4.10) määrittelee riippumattomuuden, vaikka pätisi $\mathsf{P}(B) = 0$. Ehdon (4.10) mukaan riippumattomuus on symmetrinen ominaisuus: Jos A on riippumaton B:stä, myös B on riippumaton A:sta. Sama symmetrisyysominaisuus seuraa tulosäännöstä (4.9) ja ehdosta (4.11), jos sekä $\mathsf{P}(B) > 0$ että $\mathsf{P}(A) > 0$ (jolloin molemmille tapahtumille voidaan ehdollistaa):

$$\mathsf{P}(A \mid B) \times \mathsf{P}(B) = \mathsf{P}(A) \times \mathsf{P}(B) = \mathsf{P}(B) \times \mathsf{P}(A) = \mathsf{P}(B \mid A) \times \mathsf{P}(A).$$

Reunimmaiset yhtäsuuruudet seuraavat riippumattomuusoletuksesta. Kummassakaan tapahtumassa ei ole informaatiota toisesta.

Esimerkki. Kortin peluu (kuva 4.11). Korttipakka koostuu 52 kortista (pakassa ei ole jokereita). Kortit jakaantuvat neljään maahan: pataan, ristiin, herttaan ja ruutuun. Kunkin maan kortit on numeroitu 1–13. Numeroa 1 kutsutaan ässäksi ja numeroita 11–13 sotilaaksi, kuningattareksi ja kuninkaaksi. Vedetään korttipakasta sattumanvaraisesti kortti. Kunkin yksittäisen kortin todennäköisyys tulla valituksi on 1/52. Ovatko tapahtumat hertta ja kuningas riippumattomia?

Tapahtumat ovat riippumattomia, mikäli

$$\mathsf{P}(\text{hertta}) \times \mathsf{P}(\text{kuningas}) = \frac{1}{52},$$

[1]Kuva: Picryl (`https://picryl.com/media/card-players-alexander-laureus-nationalmuseum-19982-0bed3c`; haettu 7.5.2023). Alkuperäisteos kuuluu Ruotsin kansallismuseon kokoelmiin.

Kuva 4.11: Alexander Lauréus (1783–1823): Kortinpelaajat.[1]

joka on todennäköisyys, että saadaan sekä hertta että kuningas (herttakuningas).

Todennäköisyys saada hertta tai kuningas ovat

$$\mathsf{P}(\text{hertta}) = \frac{13}{52} = \frac{1}{4}$$

ja

$$\mathsf{P}(\text{kuningas}) = \frac{4}{52} = \frac{1}{13}.$$

Niiden tulo on

$$\mathsf{P}(\text{hertta}) \times \mathsf{P}(\text{kuningas}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52}.$$

Saatiin herttakuninkaan todennäköisyys. Tapahtumat hertta ja kuningas ovat riippumattomia.

Tarkistetaan, että ehdolliset ja ehdollistamattomat todennäköisyydet ovat samat:

$$\mathsf{P}(\text{hertta} \mid \text{kuningas}) = \frac{\mathsf{P}(\text{hertta ja kuningas})}{\mathsf{P}(\text{kuningas})} = \frac{1/52}{4/52} = \frac{1}{4} = \mathsf{P}(\text{hertta}) \text{ ja}$$

$$\mathsf{P}(\text{kuningas} \mid \text{hertta}) = \frac{\mathsf{P}(\text{kuningas ja hertta})}{\mathsf{P}(\text{hertta})} = \frac{1/52}{13/52} = \frac{1}{13} = \mathsf{P}(\text{kuningas}). \quad \square$$

Riippumattomuus on eri asia kuin erillisyys. Jos A ja B ovat erillisiä, ne eivät voi tapahtua yhtä aikaa (kuva 4.4). Tällöin toinen on ikään kuin este toisen tapahtumiselle, informaatiota on, eivätkä tapahtumat voi olla riippumattomia, jos niillä on positiivinen todennäköisyys. Tällöin riippumattomuusehto ei toteudu:

$$\mathsf{P}(A \cap B) = 0 \neq \mathsf{P}(A) \times \mathsf{P}(B).$$

Yhtäsuuruus seuraa erillisyysoletuksesta. Erilliset tapahtumat ovat riippumattomia vain, jos toisen tapahtuman (tai molempien) todennäköisyys on 0.

Esimerkki.[26] Jos tapahtumat A ja B ovat riippumattomia ja sekä $\mathsf{P}(A) > 0$ että $\mathsf{P}(B) > 0$, ovatko A ja B^C riippumattomia? Kyllä:

$$\mathsf{P}(B^C \mid A) = 1 - \mathsf{P}(B \mid A) = 1 - \mathsf{P}(B) = \mathsf{P}(B^C).$$

Toinen yhtäsuuruus seuraa A:n ja B:n riippumattomuudesta. Koska ehdollinen on ehdollistamaton todennäköisyys, A ja B^C ovat riippumattomia (riippumattomuusehto (4.11)). Vaihtamalla A:n ja B:n rooleja seuraa, että myös A^C ja B ovat riippumattomia, jos A ja B ovat. $\square$

Seuraava tulos mahdollistaa vastatapahtumien todennäköisyyksien kertomisen todennäköisyyslaskuissa, kun tiedetään, että tapahtumat ovat riippumattomia.

Esimerkki.[27] Olkoot A ja B riippumattomia tapahtumia. Ovatko A^C ja B^C riippumattomia tapahtumia?

Yleisestä yhteenlaskusäännöstä (4.6) saadaan

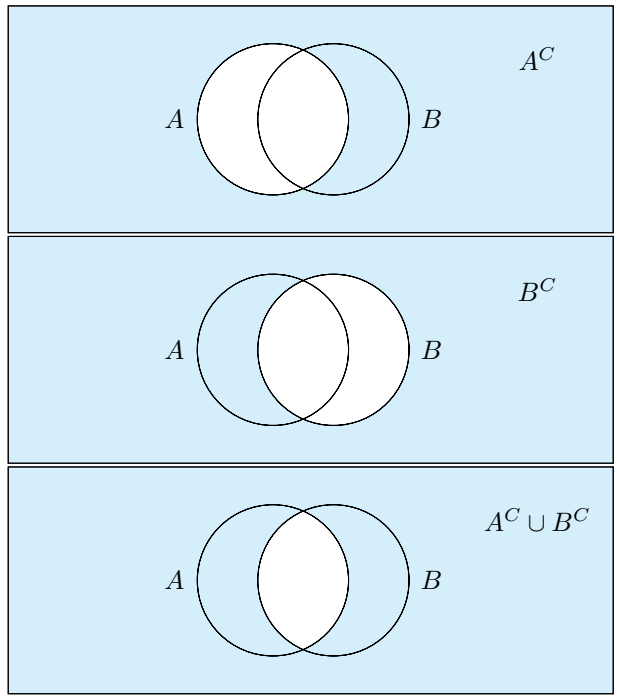
$$\mathsf{P}(A^C \cup B^C) = \mathsf{P}(A^C) + \mathsf{P}(B^C) - \mathsf{P}(A^C \cap B^C).$$

Toisaalta

$$\mathsf{P}(A^C \cup B^C) = \mathsf{P}((A \cap B)^C) = 1 - \mathsf{P}(A \cap B)$$

(DeMorganin sääntö). Venn-diagrammit kuvassa 4.12 havainnollistavat. Yhtälöistä seuraa, että

$$\begin{aligned} 1 - \mathsf{P}(A \cap B) &= \mathsf{P}(A^C) + \mathsf{P}(B^C) - \mathsf{P}(A^C \cap B^C) \\ &= 1 - \mathsf{P}(A) + 1 - \mathsf{P}(B) - \mathsf{P}(A^C \cap B^C). \end{aligned}$$



Kuva 4.12: DeMorganin sääntö.

Riippumattomuusoletuksen mukaan $\mathsf{P}(A \cap B) = \mathsf{P}(A) \times \mathsf{P}(B)$. Näin ollen

$$\begin{aligned}\mathsf{P}(A^C \cap B^C) &= 1 - \mathsf{P}(A) + 1 - \mathsf{P}(B) - [1 - \mathsf{P}(A) \times \mathsf{P}(B)] \\ &= [1 - \mathsf{P}(A)] \times [1 - \mathsf{P}(B)] \\ &= \mathsf{P}(A^C) \times \mathsf{P}(B^C).\end{aligned}$$

Koska $\mathsf{P}(A^C \cap B^C) = \mathsf{P}(A^C) \times \mathsf{P}(B^C)$, niin A^C ja B^C ovat riippumattomia tapahtumia. □

Esimerkki. Kahden nopan heitto (jatkoa). Aiemmin todettiin, että kaikki 36 silmälukuparia $(1,1), (1,2), \ldots, (6,5), (6,6)$ ovat yhtä todennäköisiä. Perustellaan se. Noppien heitot ovat riippumattomia. Kunkin silmäluvun todennäköisyys on $1/6$. Kunkin silmälukuparin (i,j) $(i,j = 1,\ldots,6)$ todennäköisyys on riippumat-

tomuuden perusteella 1/36:

$$\mathsf{P}(i,j) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

Vaikkapa silmälukuparien (1,2), (2,1) ja (2,2) todennäköisyys on sama 1/36. □

Esimerkki. Sisarukset. Oletetaan, että tytöt (T) ja pojat (P) syntyvät toisistaan riippumatta samalla todennäköisyydellä 1/2.[28] Perheessä on kaksi lasta, joista vanhempi on tyttö. Mikä on todennäköisyys, että nuorempi lapsista on tyttö?

Otosavaruus on

$$S = \{\mathrm{TT}, \mathrm{TP}, \mathrm{PT}, \mathrm{PP}\}.$$

Sisarukset ovat pareissa ikäjärjestyksessä (vanhempi ensin). Kunkin parin todennäköisyys on $(1/2) \times (1/2) = 1/4$.

Nuorempi sisarus on tyttö pareissa TT ja PT. Vanhempi sisarus on tyttö pareissa TT ja TP. Sovelletaan ehdollisen todennäköisyyden kaavaa (4.7):

$$\begin{aligned} \mathsf{P}(\{\mathrm{TT} \cup \mathrm{PT}\} \mid \{\mathrm{TT} \cup \mathrm{TP}\}) &= \frac{\mathsf{P}(\{\mathrm{TT} \cup \mathrm{PT}\} \cap \{\mathrm{TT} \cup \mathrm{TP}\})}{\mathsf{P}(\mathrm{TT} \cup \mathrm{TP})} \\ &= \frac{\mathsf{P}(\mathrm{TT})}{\mathsf{P}(\mathrm{TT} \cup \mathrm{TP})} \\ &= \frac{1/4}{1/4 + 1/4} = \frac{1}{2}. \end{aligned}$$

Todennäköisyys, että nuorempi sisaruksista on tyttö, on 1/2. □

Esimerkki. Sisarukset (jatkoa). Sisarusparadoksi I. Perheessä on kaksi lasta, joista ainakin toinen on tyttö. Mikä on todennäköisyys, että toinenkin on tyttö?

Ehdon "ainakin toinen on tyttö" rajaama otosavaruus on {TT, TP, PT}. Ehdollinen todennäköisyys on

$$\begin{aligned} \mathsf{P}(\mathrm{TT} \mid \mathrm{TT}, \mathrm{TP}, \mathrm{PT}) &= \frac{\mathsf{P}(\mathrm{TT} \cap \{\mathrm{TT}, \mathrm{TP}, \mathrm{PT}\})}{\mathsf{P}(\mathrm{TT}, \mathrm{TP}, \mathrm{PT})} = \frac{\mathsf{P}(\mathrm{TT})}{\mathsf{P}(\mathrm{TT}, \mathrm{TP}, \mathrm{PT})} \\ &= \frac{1/4}{1/4 + 1/4 + 1/4} = \frac{1}{3}. \end{aligned}$$

Kun satunnaisesti poimitaan kaksilapsinen perhe, jossa ainakin toinen lapsista on tyttö, todennäköisyydellä 1/3 molemmat lapset ovat tyttöjä. □

Riippumattomuus ei ole transitiivinen ominaisuus: Jos sekä A ja B että B ja C ovat riippumattomia tapahtumia, niin A ja C eivät välttämättä ole.

Esimerkki.[29] Otosavaruus koostuu kymmenestä yhtä todennäköisestä alkeistapahtumasta $\{1, \ldots, 10\}$ ($\mathsf{P}(\{i\}) = 1/10, i = 1, \ldots, 10$). Määritellään tapahtumat $A = \{2, 3, 4, 5, 6\}$, $B = \{4, 5, 6, 7, 8, 9\}$ ja $C = \{2, 4, 6, 8, 10\}$. Niiden todennäköisyydet ovat

$$\mathsf{P}(A) = \frac{5}{10} = \frac{1}{2},\ \mathsf{P}(B) = \frac{6}{10} = \frac{3}{5} \quad \text{ja} \quad \mathsf{P}(C) = \frac{5}{10} = \frac{1}{2}.$$

Ehdollisten todennäköisyyksien laskua varten kirjataan $A \cap B = \{4, 5, 6\}$, $B \cap C = \{4, 6, 8\}$ ja $A \cap C = \{2, 4, 6\}$ sekä niiden todennäköisyydet:

$$\mathsf{P}(A \cap B) = \mathsf{P}(B \cap C) = \mathsf{P}(A \cap C) = \frac{3}{10}.$$

Huomataan, että tapahtumat A ja B sekä B ja C ovat riippumattomia, sillä niiden ehdollistetut ja ehdollistamattomat todennäköisyydet ovat yhtäsuuret:

$$\begin{aligned}
\mathsf{P}(A \mid B) &= \frac{3/10}{6/10} = \frac{1}{2} = \mathsf{P}(A), \\
\mathsf{P}(B \mid A) &= \frac{3/10}{5/10} = \frac{3}{5} = \mathsf{P}(B), \\
\mathsf{P}(B \mid C) &= \frac{3/10}{5/10} = \frac{3}{5} = \mathsf{P}(B) \quad \text{ja} \\
\mathsf{P}(C \mid B) &= \frac{3/10}{6/10} = \frac{1}{2} = \mathsf{P}(C).
\end{aligned}$$

Tapahtuman A todennäköisyys muuttuu, jos se ehdollistetaan tapahtumalle C tai toisinpäin:

$$\begin{aligned}
\mathsf{P}(A \mid C) &= \frac{3/10}{5/10} = \frac{3}{5} \neq \mathsf{P}(A) \quad \text{ja} \\
\mathsf{P}(C \mid A) &= \frac{3/10}{5/10} = \frac{3}{5} \neq \mathsf{P}(C).
\end{aligned}$$

A ja B ovat riippumattomia, B ja C ovat riippumattomia, mutta A ja C eivät ole. □

Esimerkki. Harrastukset (jatkoa). Määritellään tapahtumiksi A = "harrastaa innokkaasti videopelejä", B = "pelaa shakkia" ja C = "kuntoilee aktiivisesti". Voisi kuvitella, että harrastukset A ja B olisivat riippumattomia eli että videopelaamisen todennäköisyys ei riippuisi shakinpelaamisesta ja päinvastoin:

$\mathsf{P}(A \mid B) = \mathsf{P}(A)$ ja $\mathsf{P}(B \mid A) = \mathsf{P}(B)$. Kuntourheilu ja shakki saattaisivat nekin olla riippumattomia harrastuksia: $\mathsf{P}(B \mid C) = \mathsf{P}(B)$ ja $\mathsf{P}(C \mid B) = \mathsf{P}(C)$. Silti saattaisi päteä aiemmin ounasteltu videopelaamisen ja kuntoilun käänteinen yhteys eli riippuvuus: $\mathsf{P}(A \mid C) < \mathsf{P}(A)$ ja $\mathsf{P}(C \mid A) < \mathsf{P}(C)$. □

Monen tapahtuman riippumattomuus. Kahden erillisen tapahtuman todennäköisyyksien yhteenlaskusääntö (4.4) yleistyy yksinkertaisella tavalla kaavaksi (4.5) tapahtumille $A_1, \ldots, A_n$. Niiden riippumattomuusehto ei yleisty yhtä suoraviivaisesti. Ne ovat riippumattomia, jos kaikki yhtäsuuruudet alla pätevät:

$$\mathsf{P}(A_i \cap A_j) = \mathsf{P}(A_i) \times \mathsf{P}(A_j), \quad \mathsf{P}(A_i \cap A_j \cap A_k) = \mathsf{P}(A_i) \times \mathsf{P}(A_j) \times \mathsf{P}(A_k), \ldots,$$
$$\mathsf{P}(A_1 \cap \cdots \cap A_n) = \mathsf{P}(A_1) \times \cdots \mathsf{P}(A_n).$$

Yllä alaindeksit $i, j, k, \ldots$ saavat kaikki mahdolliset arvot $1, \ldots n$ ja ovat kaikki eri suuria kullakin rivillä.[30] Parittainen riippumattomuus ($\mathsf{P}(A_i \cap A_j) = \mathsf{P}(A_i) \times \mathsf{P}(A_j)$, $i \neq j$) ei takaa tapahtumien riippumattomuutta, jos tapahtumia on kolme tai enemmän.

Esimerkki. Tapahtumat A, B ja C ovat riippumattomia, jos kaikki yhtälöt alla pätevät:

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \times \mathsf{P}(B), \quad \mathsf{P}(A \cap C) = \mathsf{P}(A) \times \mathsf{P}(C), \quad \mathsf{P}(B \cap C) = \mathsf{P}(B) \times \mathsf{P}(C) \text{ ja}$$
$$\mathsf{P}(A \cap B \cap C) = \mathsf{P}(A) \times \mathsf{P}(B) \times \mathsf{P}(C). \ \square$$

Monesti oletetaan, että tapahtumat ovat riippumattomia. Oletuksen taustalla on enemmän ehtoja kuin asiaan perehtymättä ehkä arvaisi.

Riippumattomuusehto (4.10) yleistyy usean tapahtuman tilanteeseen: Tapahtumien $A_1, \ldots, A_n$, ollessa riippumattomia

$$\mathsf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathsf{P}(A_1) \times \mathsf{P}(A_2) \times \cdots \times \mathsf{P}(A_n). \tag{4.12}$$

Todennäköisyyslaskut ovat usein yksinkertaisia tai yksinkertaistuvat tavattomasti, jos voidaan olettaa riippumattomuus ja kaava on käytettävissä.

Esimerkki. Kolmen lantin heitto. Heitetään lanttia kolme kertaa peräjälkeen riippumattomasti. Merkitään kruuna = H (*heads*) ja klaava = T (*tails*). Heittosarjat ovat alkeistapahtumia ja muodostavat otosavaruuden $S = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HTT}, \mathrm{THH}, \mathrm{THT}, \mathrm{TTH}, \mathrm{TTT}\}$. Riippumattomuuden perusteella kunkin alkeistapahtuman todennäköisyys on

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.$$

□

Esimerkki. Kolmen lantin heitto (jatkoa). Asiaan perehtymätön saattaisi ajatella, että heittosarjan HHH todennäköisyys olisi pienempi kuin vaihtelevammin kruunuja ja klaavoja sisältävän heittosarjan. Ajatukselle on nimikin: Pelurin virhepäätelmä (*gambler's fallacy*). Siihen sortunee helpommin, jos on heitetty vieläkin useampia kertoja peräjälkeen kruuna. Maallikko voi ajatella, että seuraavaksi täytyy tulla klaava.

Lasketaan todennäköisyys kolmannelle kruunalle, kun kaksi aiempaa heittoa ovat tuottaneet kruunan. Kahden ensimmäisen heiton tuloksen rajaama otosavaruus on $B = \{\mathrm{HHH}, \mathrm{HHT}\}$. Merkitään $A = \{\mathrm{HHH}\}$. $\mathsf{P}(B) = 1/8 + 1/8 = 1/4$ (riippumattomuuden ja erillisyyden perusteella), $\mathsf{P}(A) = 1/8$ ja

$$\begin{aligned}\mathsf{P}(A \mid B) &= \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(\mathrm{HHH} \cap \{\mathrm{HHH},\, \mathrm{HHT}\})}{\mathsf{P}(\mathrm{HHH},\, \mathrm{HHT})} = \frac{\mathsf{P}(\mathrm{HHH})}{\mathsf{P}(\mathrm{HHH},\, \mathrm{HHT})} \\ &= \frac{1/8}{1/4} = \frac{1}{2}.\end{aligned}$$

Kruunan todennäköisyys kahden kruunan heiton jälkeen on edelleen 1/2. Vastaavasti voitaisiin osoittaa, että vaikkapa 11. kruunan todennäköisyys on 1/2, vaikka sitä ennen olisi heitetty 10 kruunaa. □

Esimerkki. Tentin läpäisy (jatkoa). Uhkapelurin virhepäätelmä ei rajoitu tapahtumasarjoihin, joissa yksittäisen tapahtuman todennäköisyys on 1/2. Tekeekö opiskelija uhkapelurin virhepäätelmän, jos hän on reputtanut monta kertaa tentin ja ajattelee, että seuraavalla kerralla tentin täytyy mennä läpi? Ei välttämättä. Jos opiskelija on ahkeroinut tenttien välissä, on todennäköisyys läpäistä tentti kasvanut. Mikäli opiskelija yrittää kerta toisensa jälkeen tenttiä samoilla tiedoilla ja uskoo, että seuraavalla kerralla hänen täytyy jo läpäistä tentti, hän tekee uhkapelurin virhepäätelmän — riippumatta siitä, mikä on tentin läpäisemisensä todennäköisyys. □

Esimerkki. Yle Uutiset 9.5.2011:[31]

> Japanilainen ydinvoimayhtiö – – on päättänyt sulkea toistaiseksi Hamaokan ydinvoimalan. Japanin hallitus pyysi yhtiötä sulkemaan laitoksen, koska se on erityisen altis luonnononnettomuuksille. – – Hamaokan ydinvoimala sijaitsee 200 kilometriä Tokiosta länteen. Voimalaa on pidetty maailman vaarallisimpana, koska se sijaitsee mannerlaattojen saumakohdassa. – – Japanin maanjäristysasiantuntijoiden mukaan on 87 prosentin todennäköisyys sille, että alueella tapahtuu seuraavien 30 vuoden aikana maanjäristys, joka on suuruudeltaan kahdeksan Richterin asteikolla. Järistyksen odotetaan synnyttävän vastaavanlaisen tsunamin kuin Fukushiman ydinvoimalan hukuttanut jättiaalto.

Oletetaan, että maanjäristykset ovat riippumattomia ja että todennäköisyys

ainakin yhdelle maanjäristykselle 30 vuoden aikana on 0.87.[32] Mikä on todennäköisyys π maanjäristykselle vuoden aikana?

Koska maanjäristykset ovat riippumattomia, niin ovat myös niiden vastatapahtumat "ei-maanjäristykset" (esimerkki s:lla 49). Ilmaistaan todennäköisyys maanjäristykselle vuoden aikana vastatapahtuman avulla:

$$\begin{aligned}\pi &= \mathsf{P}(\text{maanjäristys vuoden aikana}) \\ &= 1 - \mathsf{P}(\text{ei maanjäristystä vuoden aikana}) \\ &= 1 - (1 - \pi).\end{aligned}$$

Riippumattomuudesta ja oletuksesta 0.87:n todennäköisyydestä seuraa, että

$$\begin{aligned}&\mathsf{P}(\text{maanjäristys ainakin kerran 30 vuoden aikana}) \\ &\quad = 1 - \mathsf{P}(\text{ei maanjäristystä 30 vuoden aikana}) = 1 - (1 - \pi)^{30} = 0.87.\end{aligned}$$

Ratkaistaan π $(0 < \pi < 1)$:

$$\begin{aligned}(1 - \pi)^{30} = 1 - 0.87 = 0.13 \quad &\Leftrightarrow \quad 1 - \pi = (0.13)^{1/30} \quad \Leftrightarrow \\ \pi = 1 - (0.13)^{1/30} &\approx 0.066.\end{aligned}$$

Viimeisen laskun voi tehdä R:ssä komennolla `1-0.13^(1/30)`. Todennäköisyys maanjäristykselle vuoden aikana on noin 0.066. □

Tulosääntö (4.9) yleistyy useamman tapahtuman tilanteeseen. Tekniikka on ajatella useampaa ehdollistavaa tapahtumaa yhtenä. Olkoon tapahtumia kolme: A, B ja C. Merkitään $D = B \cap A$. Tulosäännön mukaan

$$\begin{aligned}&\mathsf{P}(C \cap B \cap A) \\ &\quad = \mathsf{P}(C \cap D) = \mathsf{P}(C \mid D) \times \mathsf{P}(D) = \mathsf{P}(C \mid B \cap A) \times \mathsf{P}(B \cap A) \\ &\quad = \mathsf{P}(C \mid B \cap A) \times \mathsf{P}(B \mid A) \times \mathsf{P}(A).\end{aligned} \tag{4.13}$$

Jos tapahtumia on n $(A_1, \ldots, A_n)$, tulosääntö yleistyy samalla tekniikalla näin:

$$\begin{aligned}\mathsf{P}(A_n \cap A_{n-1} \cap \cdots \cap A_1) &= \mathsf{P}(A_n \mid \{A_{n-1} \cap \cdots \cap A_1\}) \\ &\quad \times \mathsf{P}(A_{n-1} \mid \{A_{n-2} \cap \cdots \cap A_1\}) \times \mathsf{P}(A_{n-2} \mid \{A_{n-3} \cap \cdots \cap A_1\}) \\ &\quad \times \ldots \times \mathsf{P}(A_2 \mid A_1) \times \mathsf{P}(A_1).\end{aligned}$$

Esimerkki.[33] Sisarukset (jatkoa). Perheessä on kaksi lasta ($S = \{\text{TT, TP, PT, PP}\}$ ja kunkin parin todennäköisyys on 1/4). Tapaat sattumalta toisen lapsista

kutsuilla ja huomaat, että hän on tyttö. Mikä on todennäköisyys, että lapsista toinenkin on tyttö? Oletetaan, että tapaat lapsen samalla todennäköisyydellä 1/2 riippumatta hänen sukupuolestaan.

Merkitään A = "vanhempi lapsi on tyttö", B = "nuorempi lapsi on tyttö" ja C = "lapsi, jonka tapaat on tyttö". Tällöin $\mathsf{P}(A) = \mathsf{P}(B) = 1/2$, koska ehdot täyttäviä sisaruspareja on kaksi otosavaruudessa. Myös $\mathsf{P}(C) = 1/2$, sillä oletettiin, että sattumanvaraisesti tavattava lapsi on yhtätodennäköisesti tyttö kuin poika. Tapahtuman molemmat lapset ovat tyttöjä todennäköisyys on $\mathsf{P}(A \cap B) = 1/4$, koska sellaisia sisaruspareja on yksi (TT) otosavaruudessa. Huomataan, että $A \cap B \cap C = A \cap B$, sillä jos molemmat lapset ovat tyttöjä, myös satunnaisesti tavattavan lapsen täytyy olla tyttö! Havaitaan, että perheen lapsista toinenkin on tyttö todennäköisyydellä 1/2, jos sattumalta tavattu lapsi on tyttö:

$$\mathsf{P}(A \cap B \mid C) = \frac{\mathsf{P}(A \cap B \cap C)}{\mathsf{P}(C)} = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(C)} = \frac{1/4}{1/2} = \frac{1}{2}. \quad \square$$

Esimerkki. Sisarukset (jatkoa). Sisarusparadoksi II. Perheessä on kaksi lasta. Mikä on todennäköisyys, että lapsista toinenkin on tyttö, jos tiedetään, että ainakin toinen lapsista on talvella syntynyt tyttö? Oletetaan, että lapset syntyvät toisistaan riippumattomasti todennäköisyydellä 1/4 kunakin vuodenaikana ja että sukupuoli ja vuodenaika ovat riippumattomia tapahtumia.

Laskettava todennäköisyys on

$$\begin{aligned}&\mathsf{P}(\text{molemmat tyttöjä} \mid \text{ainakin yksi talvityttö})\\ &= \frac{\mathsf{P}(\text{molemmat tyttöjä ja ainakin yksi talvityttö})}{\mathsf{P}(\text{ainakin yksi talvityttö})}\end{aligned}$$

Riippumattomuuden perusteella todennäköisyys syntyä tyttönä talvella on $(1/2) \times (1/4) = 1/8$. Todennäköisyys nimittäjässä saadaan näin:

$$\begin{aligned}\mathsf{P}(\text{ainakin yksi talvityttö}) &= 1 - \mathsf{P}(\text{ei yhtään talvityttöä})\\ &= 1 - \mathsf{P}(\text{1. lapsi ei ole talvityttö ja 2. lapsi ei ole talvityttö})\\ &= 1 - \left(1 - \frac{1}{8}\right)^2 = 1 - \left(\frac{7}{8}\right)^2 = \frac{15}{64}.\end{aligned}$$

Todennäköisyyden osoittajassa päättely:

$$\mathsf{P}(\text{molemmat tyttöjä ja ainakin yksi talvityttö})$$

$$
\begin{aligned}
&= \mathsf{P}(\text{molemmat tyttöjä ja ainakin yksi talvilapsi}) \\
&= \frac{1}{4} \times \mathsf{P}(\text{ainakin yksi talvilapsi}) \\
&= \frac{1}{4} \times [1 - \mathsf{P}(\text{kumpikaan ei talvilapsi})] \\
&= \frac{1}{4} \times [1 - \mathsf{P}(\text{1. ei talvilapsi ja 2. ei talvilapsi})] \\
&= \frac{1}{4} \times \left[1 - \left(\frac{3}{4}\right)^2\right] \\
&= \frac{7}{64}.
\end{aligned}
$$

Ensimmäinen yhtäsuuruus pätee, koska viitatut osajoukot (joukkojen leikkaukset) ovat samat! Seuraava yhtäsuuruus tulee sukupuoli- ja vuodenaikatapahtumien riippumattomuudesta, minkä johdosta kahden tytön todennäköisyys (1/4) voidaan kertoa talvilapsen syntymiseen liittyvällä todennäköisyydellä.

Todennäköisyys, että perheessä on kaksi tyttöä, kun perheessä on ainakin yksi talvella syntynyt tyttö, on

$$\mathsf{P}(\text{molemmat tyttöjä} \mid \text{ainakin yksi talvityttö}) = \frac{7/64}{15/64} = \frac{7}{15}\ !$$

Todennäköisyys 7/15 sijoittuu aiemmissa sisarusesimerkeissä laskettujen todennäköisyyksien 1/3 ja 1/2 väliin. □

Tapahtumat A ja B ovat *ehdollisesti riippumattomia* (*conditionally independent*), jos

$$\mathsf{P}(A \cap B \mid C) = \mathsf{P}(A \mid C) \times \mathsf{P}(B \mid C).$$

Yllä C on tapahtuma, jolle ehdollistetaan ($\mathsf{P}(C) > 0$). Tapahtumat voivat olla ehdollisesti riippumattomia, vaikka ne eivät olisi riippumattomia, tai ne voivat olla riippumattomia, vaikka ne eivät olisi ehdollisesti riippumattomia. Jos tapahtumat ovat riippumattomia ehdolla C, ne eivät välttämättä ole riippumattomia ehdolla C^C.

Esimerkki.[34] Kaksi lanttia. Ehdollisesta riippumattomuudesta ei seuraa riippumattomuutta. Lanteista toinen (“kultainen lantti” tai “lantti1”) on tavanomainen: Kruunan todennäköisyys on 1/2. Toinen lantti (“hopeinen lantti” tai “lantti2”) on erikoinen: Kruunan todennäköisyys on 3/4.

Arvotaan jommankumman värinen lantti ja heitellään sitä. Heittojen tulokset (kruuna tai klaava) ovat riippumattomia ehdolla arvotunvärinen lantti. Kruunujen todennäköisyys kylläkin riippuu arvotusta lantista.

Jos lantteja ei voi erottaa väristä, heittojen tulokset eivät ole riippumattomia. Arvottua lanttia heitetään ja saadaan kruuna. Se viittaa siihen, että on valittu "lantti2". Seuraavankin heiton tulos olisi tällöin todennäköisemmin kruuna. Heittojen tulokset eivät ole riippumattomia.

Olkoot C lantin arvonnan tulos ("lantti1" tai "lantti2") ja A ja B ensimmäisen ja toisen lantin heiton tulos (kruuna tai klaava). A ja B ovat nyt riippumattomia ehdolla C: Ensimmäinen heitto ei anna informaatiota toisen heiton tuloksesta. A ja B eivät ole riippumattomia, sillä A:ssa on informaatiota B:stä: Jos A oli kruuna, se viittaa siihen, että on valittu "lantti2", jolloin on todennäköisempää, että B:kin on kruuna. Esimerkkiä jatketaan harjoitustehtävässä. □

Esimerkki. Puhelimen soiminen. Riippumattomuudesta ei seuraa ehdollista riippumattomuutta. Vanhuksella on kaksi ystävää A ja B. Kunakin päivänä A voi soittaa riippumatta siitä, soittaako B ja päinvastoin. Kun puhelin soi (ehdollistava tapahtuma), A ja B eivät ole riippumattomia: Jos soittaja on A, soittaja ei voi olla B ja päinvastoin. □

Esimerkki. Kahdenlaiset kurssit. Riippumattomuudesta ehdolla C ei seuraa riippumattomuus ehdolla C^C. Opiskelija on erikoisessa yliopistossa. Toisilla kursseilla arvosana on sama opiskelun määrästä riippumatta; toisilla ahertaminen palkitaan hyvällä arvosanalla. Merkitään ensimmäisenlaista kurssia C:llä (jälkimmäisenlaista C^C:lla), hyvää arvosanaa A:lla (huonoa A^C:lla) ja ahertamista B:llä (ahertamattomuutta B^C:lla). Tällöin A ja B ovat riippumattomia ehdolla C mutteivät ole riippumattomia ehdolla C^C. □

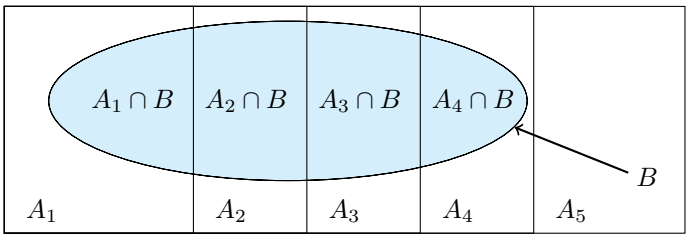
4.6 Kokonaistodennäköisyys ja Bayesin kaava

Olkoon otosavaruus ositettu erillisiin tapahtumiin A_i, joilla on kaikilla positiivinen todennäköisyys ($S = \cup_{i=1}^{n} A_i, A_i \cap A_j = \emptyset$ ja $\mathsf{P}(A_i) > 0, i = 1, \ldots, n$). Tällöin tapahtuman B todennäköisyys on

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i). \tag{4.14}$$

Kaavaa kutsutaan *kokonaistodennäköisyyden laiksi*. Se perustellaan näin: $B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$. Leikkaukset ovat erillisiä. Erillisten

tapahtumien yhteenlaskusäännön (4.5) ja tulosäännön (4.9) mukaan $\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \cap A_i) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i)$. Kuvan 4.13 Venn-diagrammissa otosavaruus koostuu 5 erillisestä tapahtumasta A_i ($n = 5$) ja tapahtumaa B merkkaa sininen ellipsi. Kuvan tilanteessa $B \cap A_5 = \emptyset$, jolloin termi $\mathsf{P}(B \mid A_5) = 0$ kokonaistodennäköisyyden kaavassa (4.14).



Kuva 4.13: Kokonaistodennäköisyyden lasku.

Esimerkki.[35] Todistusaineiston väärentäminen. Helsingin Sanomat 5.2.2016 ja 23.10.2023:

> – – tutkinnan puolueettomuus nousi keskeiseksi teemaksi – – Helsingin käräjäoikeudessa – – . Jari Aarnion asianajaja Riitta Leppiniemi ja jutun tutkinnanjohtaja, kihlakunnansyyttäjä Jukka Haavisto ottivat tiukasti yhteen – – . Leppiniemi kyseenalaisti koko Aarnio-tutkinnan ja Haaviston puolueettomuuden. Hän huomautti, että Haavisto oli huumepoliisien ensimmäisessä virkarikosjutussa toisena syyttäjänä. – – Aarnio on – – pyytänyt, ettei keskusrikospoliisin rikosylikomisario Rabbe von Hertzen osallistuisi jutun tutkintaan. Aarnio on nimennyt von Hertzenin julkiseksi vihamiehekseen. – – Von Hertzeniltä kysyttiin myös salanauhoituksesta, jonka poliisi oli tehnyt huumerikostutkijoiden risteilyllä – – . Keskustelun lomassa von Hertzen luonnehtii Aarniota rosvoksi. – – Istunnossa myös selviteltiin – – miten Aarnion tontilta Porvoosta löydettiin 65 000 euron rahakätkö – – . – – Aarnion mukaan kätkö on lavastus.
>
> Syyttäjän mukaan järjestyksenvalvojat pahoinpitelivät joukolla ihmisen – – ja lavastivat hänet pahoinpitelyn jälkeen huumerikoksesta.

Oikeudessa puolustus väittää poliisin väärentäneen todistusaineistoa tai muuten toimineen väärin. Todennäköisyys, että puolustus pystyy vakuuttamaan oikeuden, että niin on tapahtunut, on 0.70. Tällöin oikeus tuomitsee syytetyn todennäköisyydellä 0.15. Muussa tilanteessa tuomion todennäköisyys on 0.80. Mikä on todennäköisyys, että oikeus tuomitsee syytetyn?

Merkitään B:llä, A_1:llä ja A_2:lla tapahtumia oikeus tuomitsee syytetyn, oikeus katsoo poliisin toimineen väärin ja oikeus ei katso poliisin toimineen väärin.

Annettujen tietojen mukaan $\mathsf{P}(B \mid A_1) = 0.15$, $\mathsf{P}(B \mid A_2) = 0.80$, $\mathsf{P}(A_1) = 0.70$ ja $\mathsf{P}(A_2) = 1 - 0.70 = 0.30$. Tuomion todennäköisyys on 0.345:

$$\begin{aligned}\mathsf{P}(B) &= \mathsf{P}(B \mid A_1)\mathsf{P}(A_1) + \mathsf{P}(B \mid A_2)\mathsf{P}(A_2) \\ &= 0.15 \times 0.70 + 0.80 \times 0.30 \\ &= 0.345. \ \square\end{aligned}$$

Bayesin kaava[36] on

$$\mathsf{P}(A_j \mid B) = \frac{\mathsf{P}(B \mid A_j)\mathsf{P}(A_j)}{\mathsf{P}(B)} = \frac{\mathsf{P}(B \mid A_j)\mathsf{P}(A_j)}{\sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i)}, \tag{4.15}$$

jossa $1 \leq j \leq n$. Kaavaa kutsutaan myös käänteistodennäköisyyden kaavaksi, koska tapahtuman (B) ja sen ehdon (A_j) roolit on käännetty kaavan vasemmalla puolella. Kaavan perustelu on lyhyt: $\mathsf{P}(A_j \mid B) = \mathsf{P}(A_j \cap B)/\mathsf{P}(B) = \mathsf{P}(B \mid A_j)\mathsf{P}(A_j)/\mathsf{P}(B)$, jossa jälkimmäinen yhtäsuuruus seuraa tulosäännöstä (4.9). Sijoittamalla nimittäjään kokonaistodennäköisyyden kaava (4.14) saadaan Bayesin kaava.

Esimerkki. Todistusaineiston väärentäminen (jatkoa). Oikeus tuomitsi syytetyn. Mikä on todennäköisyys, että oikeus katsoi poliisin toimineen väärin?

Bayesin kaavan perusteella todennäköisyys on noin 0.304:

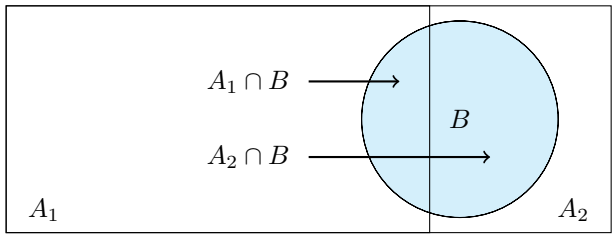
$$\begin{aligned}\mathsf{P}(A_1 \mid B) &= \frac{\mathsf{P}(A_1 \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(B \mid A_1)\mathsf{P}(A_1)}{\mathsf{P}(B \mid A_1)\mathsf{P}(A_1) + \mathsf{P}(B \mid A_2)\mathsf{P}(A_2)} \\ &= \frac{0.15 \times 0.70}{0.345} \approx 0.304.\end{aligned}$$

Kuvan 4.14 Venn-diagrammi havainnollistaa. Relevantti otosavaruus on tapahtuman B, oikeus tuomitsee syytetyn, rajaama. Tapahtumat A_1 ja A_2 ovat erillisiä. A_1:n pinta-ala on 0.7, A_2:n pinta-ala on 0.3, otosavaruuden $A_1 \cup A_2$ pinta-ala on 1 ja B:n pinta-ala on 0.345. Koska $\mathsf{P}(A_1 \mid B) \approx 0.304$, niin $\mathsf{P}(A_2 \mid B) = 1 - \mathsf{P}(A_1 \mid B) \approx 0.696$. Diagrammin pinta-alat toteuttavat (likimäärin) yhtälöt

$$\begin{aligned}\mathsf{P}(A_1 \cap B) &= \mathsf{P}(B \mid A_1)\mathsf{P}(A_1) = 0.15 \times 0.70 = 0.105, \\ \mathsf{P}(A_2 \cap B) &= \mathsf{P}(B \mid A_2)\mathsf{P}(A_2) = 0.8 \times 0.30 = 0.240, \\ \mathsf{P}(B \mid A_1) &= \frac{\mathsf{P}(A_1 \cap B)}{\mathsf{P}(A_1)} = \frac{0.105}{0.7} = 0.15,\end{aligned}$$

$$\mathsf{P}(B \mid A_2) = \frac{\mathsf{P}(A_2 \cap B)}{\mathsf{P}(A_2)} = \frac{0.240}{0.3} = 0.8,$$
$$\mathsf{P}(A_1 \mid B) = \frac{\mathsf{P}(A_1 \cap B)}{\mathsf{P}(B)} = \frac{0.105}{0.345} \approx 0.304 \quad \text{ja}$$
$$\mathsf{P}(A_2 \mid B) = \frac{\mathsf{P}(A_2 \cap B)}{\mathsf{P}(B)} = \frac{0.240}{0.345} \approx 0.696.$$

Tapahtumaa A_1 otosavaruudessa B kuvaava pinta-ala on noin 0.304-osaa tapahtumaa B kuvaavasta pinta-alasta. Tapahtumaa A_2 otosavaruudessa B kuvaava pinta-ala on noin 0.696-osaa tapahtumaa B kuvaavasta pinta-alasta. □



Kuva 4.14: Bayesin kaavan geometrinen tulkinta.

Esimerkki. HIV. HIV-infektoituneita oli 14.2.2010 mennessä Suomessa tilastoitu 2 618, joista 497 oli kuollut. Infektiot esiintyvät valtaosin ikäryhmässä 15–64-vuotta. Tämänikäisiä suomalaisia oli 3 543 084 vuoden 2008 lopussa. HIV-infektion todennäköisyys tässä ikäryhmässä on siten noin $(2\,618 - 497)/3\,543\,084$ eli 0.0006. Lääketieteellinen testi indikoi infektiota todennäköisyyksillä 0.997 ja 0.015, kun henkilö on tai ei ole infektoitunut (testi ei anna aina oikeata vastausta kummassakaan tilanteessa).[37]

Mikä oli vuonna 2010 todennäköisyys, että satunnaisesti valittu ja testattu 15–64-vuotias suomalainen on HIV-infektoitunut, kun hänelle tehty testi indikoi infektiota?

Merkitään I:llä infektoitunutta, E:llä ei-infektoitunutta ja $+$:lla positiivista testitulosta (testi indikoi infektiota). Annetuista tiedoista saadaan todennäköisyydet $\mathsf{P}\ (I) = 0.0006$, $\mathsf{P}(E) = 0.9994$, $\mathsf{P}(+ \mid I) = 0.997$ ja $\mathsf{P}(+ \mid E) = 0.015$.

Sijoitetaan ne Bayesin kaavaan:

$$\begin{aligned} \mathsf{P}(I \mid +) &= \frac{\mathsf{P}(+ \mid I)\mathsf{P}(I)}{\mathsf{P}(+ \mid I)\mathsf{P}(I) + \mathsf{P}(+ \mid E)\mathsf{P}(E)} \\ &= \frac{0.997 \times 0.0006}{0.997 \times 0.0006 + 0.015 \times 0.9994} \\ &\approx 0.038. \end{aligned}$$

Todennäköisyys, että satunnaisesti valittu 15–64-vuotias suomalainen on HIV-infektoitunut, kun HIV-testitulos on positiivinen, on noin 0.04. Virus on niin harvinainen, että positiivinen testitulos johtuu tavattomasti useammin virheellisestä testituloksesta kuin infektiosta. □

Esimerkki. HIV (jatkoa). Kansanterveyslaitos (nykyinen Terveyden ja hyvinvoinnin laitos) ja Aids-tukikeskus lähettivät kyselylomakkeen sekä HIV:tä testaavan sylkitestin seksuaali- ja sukupuolivähemmistöjen lehden tilaajarekisterin miesten osoitteisiin 2006. Homo- tai biseksuaalisista miehistä HIV-infektoituneeksi ilmoittautui tai tutkimuksessa havaittiin infektoituneiksi 4.6 %.[38] Mikä on todennäköisyys, että tilaajarekisteristä satunnaisesti valittu homo- tai biseksuaalinen mies on HIV-infektoitunut, kun hänelle tehty testi indikoi infektiota?

Infektion yleisyyttä kuvaavat todennäköisyydet ovat nyt $\mathsf{P}(I) = 0.046$ ja $\mathsf{P}(E) = 0.954$. Sijoitetaan ne yhdessä testin ominaisuuksia kuvaavien todennäköisyyksien (samat kuin edellä) kanssa Bayesin kaavaan:

$$\begin{aligned} \mathsf{P}(I \mid +) &= \frac{\mathsf{P}(+ \mid I)\mathsf{P}(I)}{\mathsf{P}(+ \mid I)\mathsf{P}(I) + \mathsf{P}(+ \mid E)\mathsf{P}(E)} \\ &= \frac{0.997 \times 0.046}{0.997 \times 0.046 + 0.015 \times 0.954} \\ &\approx 0.762. \end{aligned}$$

Todennäköisyys, että satunnaisesti valittu lehden tilaajarekisterin homo- tai biseksuaalinen mies on HIV-infektoitunut, kun HIV-testitulos on positiivinen, on noin 0.76. Ero todennäköisyyteen edellisessä esimerkissä on suuri ja johtuu HIV-infektion huomattavasti suuremmasta yleisyydestä tässä ryhmässä verrattuna 15–64-vuotiaisiin suomalaisiin. Edelleenkin kuitenkin testin virhemahdollisuudesta johtuen ($\mathsf{P}(+ \mid E) = 0.015$) HIV-infektoituneeksi testattu homo- tai biseksuaalinen tilaajamies on noin 0.24:n todennäköisyydellä infektoitumaton. □

Diagnostisen tai muun luokittelevan testin todennäköisyyttä tunnistaa sairaus tai muu ominaisuus kutsutaan testin *herkkyydeksi* (*sensitivity*). Todennäköisyys,

jolla testi tunnistaa sairauden tai ominaisuuden puuttumisen, on testin *tarkkuus* (*specificity*).

Esimerkki. HIV (jatkoa). HIV-testin herkkyys ja tarkkuus ovat 0.997 ja $1 - 0.015 = 0.985$. □

Olkoot tapahtumat A ja B. Lasketaan Bayesin kaavalla (4.15) todennäköisyydet $\mathsf{P}(A \mid B)$ ja $\mathsf{P}(A^C \mid B)$, ja jaetaan saadut yhtälöt puolittain. $\mathsf{P}(B)$:t supistuvat pois. Tulos on

$$\frac{\mathsf{P}(A \mid B)}{\mathsf{P}(A^C \mid B)} = \frac{\mathsf{P}(B \mid A)}{\mathsf{P}(B \mid A^C)} \times \frac{\mathsf{P}(A)}{\mathsf{P}(A^C)}. \tag{4.16}$$

Oikeanpuoleisin osamäärä on A:n ja sen vastatapahtuman todennäköisyyksien suhde *priorivastasuhde* (*prior odds*). Yhtälön vasemmalla puolella on A:n ja sen vastatapahtuman ehdollisten todennäköisyyksien suhde *posteriorivastasuhde* (*posterior odds*).

Vastasuhde (*odds*) kertoo, kuinka moninkertainen on tapahtuman todennäköisyys verrattuna sen vastatapahtuman todennäköisyyteen.[39] Priorivastasuhde on tämä suhde ennen ehdollistamista; posteriorivastasuhde sen jälkeen. Mitä enemmän priori- ja posteriorivastasuhteet eroavat, sitä enemmän informaatiota on ehdollistaminen B:lle tuottanut.

Vastasuhde saa arvoja nollasta ylöspäin (kuva 4.15, jossa todennäköisyyttä merkitään π:llä). Mahdottomalle tapahtumalle vastasuhde on 0. Jos todennäköisyys on pieni, se ja vastasuhde ovat suurinpiirtein yhtäsuuria. Jos tapahtuman todennäköisyys on sama kuin sen vastatapahtuman (0.5), vastasuhde on 1. Jos vastasuhde on yhtä suurempi, todennäköisyys tapahtumalle on suurempi kuin sen vastatapahtumalle. Vastasuhde suurenee nopeasti todennäköisyyden lähestyessä yhtä.

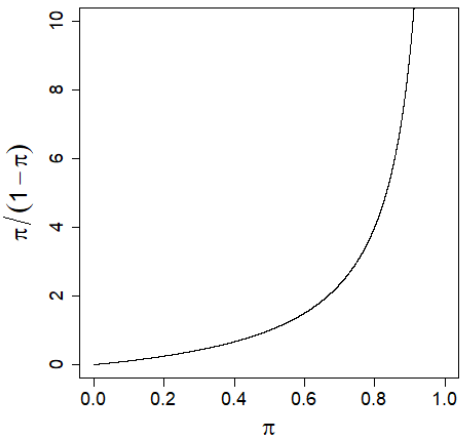
Esimerkki. Tentin läpäisy (jatkoa). Opiskelija arvioi, että hän läpäisee tentin todennäköisyydellä 0.8. Vastasuhde on varsin suuri:

$$\frac{0.8}{0.2} = 4.$$

Opiskelijan mielestä tentin hyväksymistodennäköisyys on 4 kertaa niin suuri kuin tentin hylkäämistodennäköisyys. □

Esimerkki. Lastensuojeluilmoitus. Ikäryhmästä 0–2-vuotta 5.9 %:sta oli tehty lastensuojeluilmoitus vuonna 2022.[40] Poimitaan satunnaisesti 0–2-vuotias. Asiakkuuden todennäköisyys 0.059 ei ole lukuna suuri, joten se ja vastasuhde 0.063 poikkeavat vähän:

$$\frac{0.059}{1 - 0.059} = \frac{0.059}{0.941} \approx 0.063.$$



Kuva 4.15: Vastasuhde $\pi/(1-\pi)$.

Lastensuojeluilmoituksen todennäköisyys on noin 0.063-kertaa niin suuri kuin todennäköisyys, ettei lastensuojeluilmoitus ole tehty. Toisaalta todennäköisyys, ettei lastensuojeluilmoitusta ole tehty, on noin $0.941/0.059 \approx 16$-kertaa lastensuojeluilmoituksen todennäköisyys. Jokaista lastensuojeluilmoitusta kohden on noin 16 lasta, joista ei ole tehty lastensuojeluilmoitusta.

13–15-vuotiaista 13.6 %:sta oli tehty lastensuojeluilmoitus vuonna 2022. Lastensuojeluilmoituksen todennäköisyys 0.136 ja vastasuhde 0.157 eroavat enemmän kuin edellä, sillä todennäköisyys on nyt suurempi:

$$\frac{0.013}{1-0.013} \approx 0.157.$$

□

Esimerkki. HIV (jatkoa).[41] HIV-infektion (I; I^C on E) priorivastasuhde suomalaisten ikäryhmässä 15–64-vuotta on lähes sama kuin infektion todennäköisyys:

$$\frac{\mathsf{P}(I)}{\mathsf{P}(E)} = \frac{0.0006}{0.9994} \approx 0.0006.$$

Selitys on infektion pieni todennäköisyys, jolloin vastasuhteen jakaja on lähes 1.

Positiivisen HIV-testitulokn jälkeinen todennäköisyys infektiolle on noin 0.038. Se on edelleen sen verran pieni, että posteriorivastasuhde on lähes sama:

$$\frac{\mathsf{P}(I \mid +)}{\mathsf{P}(E \mid +)} \approx \frac{0.038}{1 - 0.038} \approx 0.040.$$

Seksuaali- ja sukupuolivähemmistöjen lehden tilaajarekisterin miesten keskuudessa priori- ja posteriorivastasuhteet eroavat suuresti:

$$\frac{\mathsf{P}(I)}{\mathsf{P}(E)} = \frac{0.046}{0.954} \approx 0.048$$

ja

$$\frac{\mathsf{P}(I \mid +)}{\mathsf{P}(E \mid +)} \approx \frac{0.762}{0.238} \approx 3.205.$$

Posteriorivastasuhde (3.205) on moninkertainen verrattuna sen lähtökohtana olevaan todennäköisyyteen (0.762). On noin kolme kertaa todennäköisempää, että tilaajarekisterin mies on HIV-infektoitunut kuin että ei ole, kun testin tulos on positiivinen. □

Olkoon vastasuhde v. Sen määrittävästä yhtälöstä voidaan ratkaista todennäköisyys:

$$\frac{\mathsf{P}(A \cdot)}{\mathsf{P}(A^C \cdot)} = \frac{\mathsf{P}(A \cdot)}{1 - \mathsf{P}(A \cdot)} = v \quad \Leftrightarrow \quad \mathsf{P}(A \cdot) = \frac{v}{1+v}. \tag{4.17}$$

Yllä $\mathsf{P}(A \cdot)$ on ehdollistamaton tai ehdollinen todennäköisyys riippuen siitä, onko v priori- vai posteriorivastasuhde. Jos vastasuhde on muotoa $v = a/b$, jossa a ja b ovat positiivisia kokonaislukuja, niin todennäköisyyden voi päätellä erityisen kätevästi näin:

$$\frac{\mathsf{P}(A \cdot)}{\mathsf{P}(A^C \cdot)} = \frac{a}{b} \quad \Leftrightarrow \quad \mathsf{P}(A \cdot) = \frac{a}{a+b}. \tag{4.18}$$

Esimerkki. HIV (jatkoa). Seksuaali- ja sukupuolivähemmistöjen lehden tilaajamiehille $v = 3.205$. Kaavasta (4.17) voidaan laskea ehdollinen todennäköisyys HIV-kantajuudelle:

$$\mathsf{P}(I \mid +) \approx \frac{3.205}{1 + 3.205} \approx 0.762.$$

Saatiin aiemmin laskettu todennäköisyys. □

Esimerkki. Tentin läpäisy (jatkoa). Vastasuhde tentin läpimenolle on opiskelijan mielestä $4 = 4/1$. Läpimenon (subjektiivinen) todennäköisyys on kaavan (4.18) mukaan $4/(4+1) = 4/5 = 0.8$. □

Bayesin kaava on *johdonmukainen* (*coherent*): On samantekevää, tuleeko uusi ehdollistava tieto kerralla vai ehdollisesti riippumattomissa erissä. Saman tiedon huomioiva ehdollinen todennäköisyys on molemmilla tavoilla laskettuna sama. Kaava (4.16) päivittää kätevästi ehdollisen todennäköisyyden, kun saadaan uutta informaatiota.

Esimerkki. HIV (jatkoa). Tehdään ikäryhmään 15–64-vuotta kuuluvalle suomalaiselle kaksi HIV-testiä. Testit ovat riippumattomia ehdolla tutkittavan infektoituneisuustila (on tai ei ole). Testien herkkyys ja tarkkuus ovat 0.997 ja 0.985. Molempien testien mukaan tutkittu kantaa HIV:tä. Mikä on todennäköisyys, että hänessä on HIV?

Merkitään virusta indikoivaa 1. ja 2. testitulosta $+_1$:llä ja $+_2$:lla ja infektoituneisuutta I:llä (I^C on E). Sovelletaan kaavaa (4.16):

$$\begin{aligned}\frac{\mathsf{P}(I \mid +_1 \cap +_2)}{\mathsf{P}(E \mid +_1 \cap +_2)} &= \frac{\mathsf{P}(+_1 \cap +_2 \mid I)}{\mathsf{P}(+_1 \cap +_2 \mid E)} \times \frac{\mathsf{P}(I)}{\mathsf{P}(E)} \\ &\approx \frac{0.997^2}{0.015^2} \times 0.0006 \\ &\approx 2.646.\end{aligned}$$

Priorivastasuhde $\mathsf{P}(I)/\mathsf{P}(E) \approx 0.0006$ laskettiin edellä. Approksimatiivisen yhtäsuuruuden jälkeinen osamäärä seuraa testien riippumattomuudesta ehdolla, että tutkittava on infektoitunut (osoittaja) tai ei ole (nimittäjä). Kaavasta (4.17) saadaan kysytyksi todennäköisyydeksi 0.726:

$$\mathsf{P}(I \mid +_1 \cap +_2) \approx \frac{2.646}{1 + 2.646} \approx 0.726.$$

Jos kaksi testiä indikoi HIV-infektiota, infektoituneisuus on melko todennäköistä.

Sama tulos voidaan laskea kahdessa vaiheessa. Huomataan, että $\mathsf{P}(I \mid +_1 \cap +_2) = \mathsf{P}(+_2 \cap I \cap +_1)/\mathsf{P}(+_1 \cap +_2) = \mathsf{P}(+_2 \mid I \cap +_1)\mathsf{P}(I \mid +_1)\mathsf{P}(+_1)/\mathsf{P}(+_1 \cap +_2)$. Jälkimmäinen yhtäsuuruus seuraa yhtälöstä (4.13). Samoin $\mathsf{P}(E \mid +_1 \cap +_2) = \mathsf{P}(+_2 \mid E \cap +_1)\mathsf{P}(E \mid +_1)\mathsf{P}(+_1)/\mathsf{P}(+_1 \cap +_2)$. Näin ollen

$$\frac{\mathsf{P}(I \mid +_1 \cap +_2)}{\mathsf{P}(E \mid +_1 \cap +_2)} = \frac{\mathsf{P}(+_2 \mid I \cap +_1)}{\mathsf{P}(+_2 \mid E \cap +_1)} \times \frac{\mathsf{P}(I \mid +_1)}{\mathsf{P}(E \mid +_1)}$$

$$\approx \frac{0.997}{0.015} \times 0.040$$
$$\approx 2.646.$$

Yllä on käytetty 1. testin jälkeistä posteriorivastasuhdetta 0.040 päivitettynä priorivastasuhteena kaavassa (4.16). Posteriorivastasuhde on sama kuin edellä laskettu. Sijoittamalla $v = 2.646$ kaavaan (4.17) kysytyksi todennäköisyydeksi saadaan taas 0.726. □

4.7 Puukaavio

Puukaavio visualisoi satunnaisilmiöitä ja auttaa hahmottamaan, kuinka yhdistetty tapahtuma koostuu useammasta yksinkertaisemmasta tapahtumasta. Se helpottaa yhdistettyjen tapahtumien todennäköisyyksien ymmärtämistä ja laskemista.

Puukaavioilla voidaan laskea todennäköisyyksiä, jos tutkittavalla prosessilla on yksi alkutila, useita vaihtoehtoisia lopputiloja, joista yksi toteutuu ja välissä on tapahtumia, jotka johtavat erillisiin, toisensa poissulkeviin, tapahtumaketjuihin. Todennäköisyyksiä voidaan laskea puukaaviolla kahden säännön avulla:

- Reitin todennäköisyys on siihen johtavien tapahtumien ehdollisten todennäköisyyksien tulo (tulosääntö).

- Tapahtuman todennäköisyys on siihen johtavien reittien todennäköisyyksien summa (yhteenlaskusääntö).

Esimerkki. Tytön saaminen.[42] Pariskunta päättää tehdä lapsia, kunnes on saatu tyttö. Neljää lasta enempää ei kuitenkaan ryhdytä tekemään.

Oletetaan, että tyttöjä ja poikia syntyy todennäköisyydellä 0.5 riippumatta aiemmin syntyneiden lasten sukupuolesta ja että kerrallaan syntyy yksi lapsi. Mikä on todennäköisyys, että pariskunta saa tytön?

Puukaaviossa kuvassa 4.16 punertaviin tyttö-tapahtumiin (oksien kärkiin) päästään lähtöpisteestä (T/P; puun juuri) neljää eri reittiä (haarautunutta oksaa) pitkin. Reitin todennäköisyys saadaan lasten sukupuolien riippumattomuuden perusteella riippumattomien tapahtumien tulosäännöstä (4.12). Kukin tyttö-tapahtuma on erillinen, ja kuhunkin vie vain yksi reitti. Todennäköisyys saada 1:senä, 2:sena, 3:ntena tai 4:ntenä lapsena tyttö on riippumattomuuden johdosta $1/2$, $(1/2)^2 = 1/4$, $(1/2)^3 = 1/8$ tai $(1/2)^4 = 1/16$. Erillisten tapahtumien

yhteenlaskusäännön (4.5) perusteella reittien todennäköisyydet voidaan laskea yhteen. Todennäköisyys saada tyttö on 15/16:

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16}.$$

□

Esimerkki.[43] Useimmat tutkimustulokset vääriä. De Long ja Lang (1992) pohtivat, ovatko kaikki taloustieteelliset hypoteesit vääriä. Ioannidis (2005, 2019) kysyi, miksi useimmat julkaistut tutkimustulokset ovat vääriä. Jälkimmäinen on *PLoS*-lehden (*Public Library of Science*) luetuin artikkeli (kirjoitushetkellä noin 3.2 miljoonaa lukua), ja sen on katsottu johtaneen metodologiseen kriisin monilla empiirisillä tieteenaloilla (Spiegelhalter 2017). Szucs ja Ioannidis (2017) rummuttavat edelleen samaa viestiä. Mistä on kysymys?

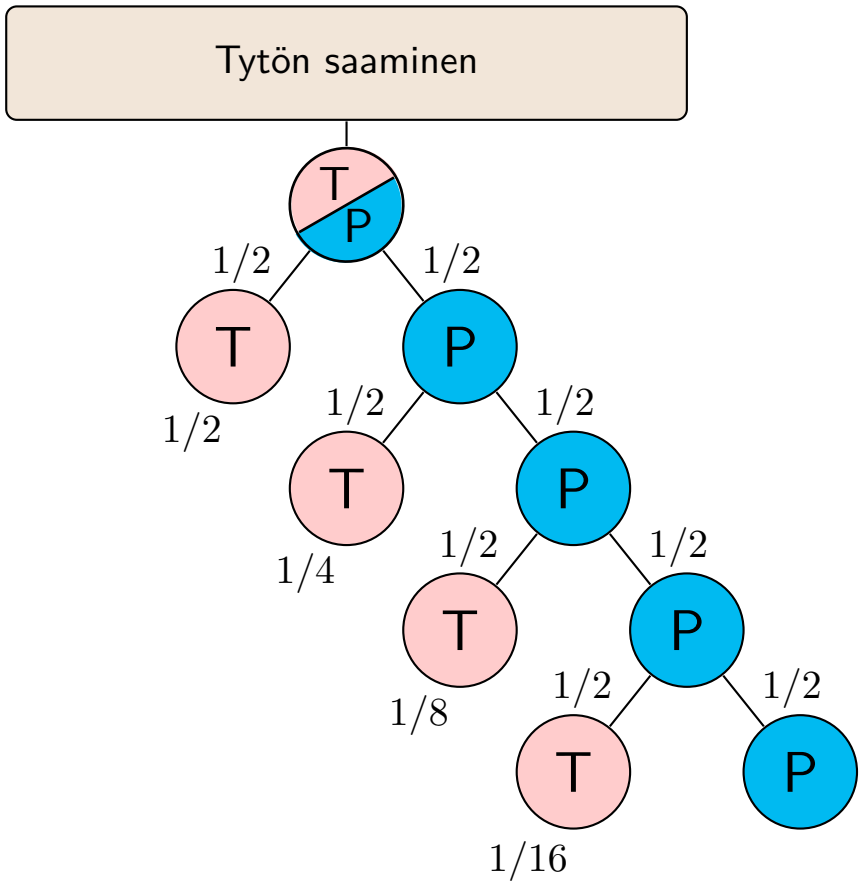
Tehdään tiedettä, ja pohditaan tuloksia Ioannidisin hengessä. Oletuksia:

- Tutkittava teoria on oikea todennäköisyydellä 0.1.
- Jos teoria on oikea, 0.5:n todennäköisyydellä tieteellinen menetelmä
  - ehdottaa teorian oikeutta (testin herkkyys).
  - ehdottaa teorian vääryyttä.
- Jos teoria on väärä, tieteellinen menetelmä todennäköisyydellä
  - 0.05 ehdottaa teorian oikeutta.
  - 0.95 ehdottaa teorian vääryyttä (testin tarkkuus).

Oletukset ovat (tieteenalasta ja sovelluskohteesta riippuen) kohtuullisen realistisia (esim. useimmat teoriat ovat vääriä).

Kuvan 4.17 puukaavio kuvaa polut, joita pitkin päädytään neljään vaihtoehtoon (niiden todennäköisyydet suluissa):

- Teoria on väärä ja menetelmän mukaan väärä ($0.9 \times 0.95 = 0.855$).
- Teoria on väärä mutta menetelmän mukaan oikea ($0.9 \times 0.05 = 0.045$).
- Teoria on oikea mutta menetelmän mukaan väärä ($0.1 \times 0.5 = 0.05$).
- Teoria on oikea ja menetelmän mukaan oikea ($0.1 \times 0.5 = 0.05$).



Kuva 4.16: Tytön saamisen todennäköisyys korkeintaan neljällä yrityksellä.

Menetelmä on varsin toimiva: Se kertoo todennäköisyydellä $0.855 + 0.05 = 0.905$ oikein, onko teoria oikea vai väärä.

Kuvio muuttuu hälyttäväksi, jos tutkijat hiljaa hautaavat tyrmätyt $(85.5+5)$ % $= 90.5$ % teorioista tai eivät saa tyrmäyksiä julkaistua. Julkisuuteen pongahtavat tällöin vain $(4.5+5)$ % $= 9.5$ % tuloksista, joiden mukaan teoria on oikea. Tällaisessa maailmassa teoria olisi oikea vain todennäköisyydellä 0.53, vaikka se olisi julkaistun tutkimuksen mukaan tosi:

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{0.05}{0.045 + 0.05} \approx 0.53$$

(kaava (4.7)). Yllä on merkitty A:lla teorian oikeutta ja B:llä teorian oikeutta menetelmän mukaan. Olisi siis oleellisesti yhtä todennäköistä, että julkaistu tutkimus pitäisi paikkansa kuin ei pitäisi.

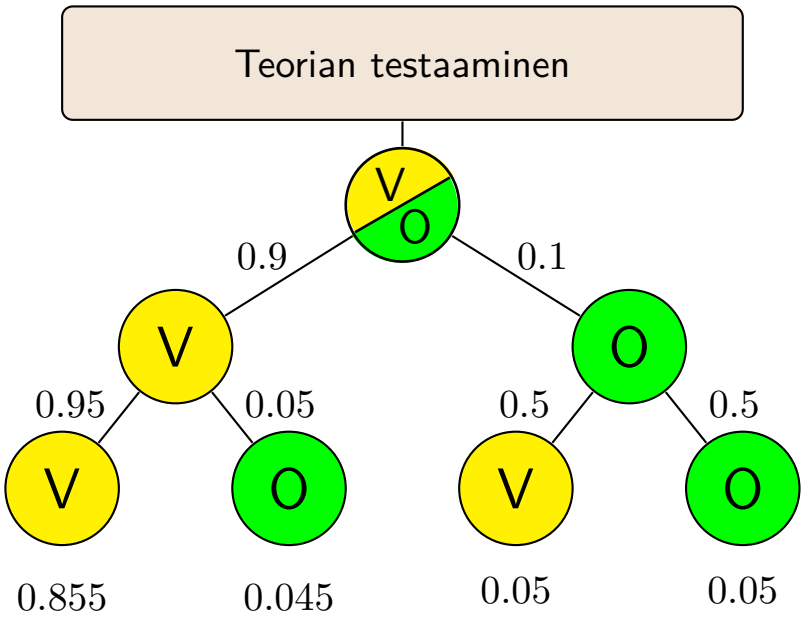
Tukka lienee noussut monilla pystyyn. Laskun mukaan Ioannidis liioitteli vain hieman. □

4.8 Kokonaistodennäköisyyden ja Bayesin kaavan ehdollistaminen

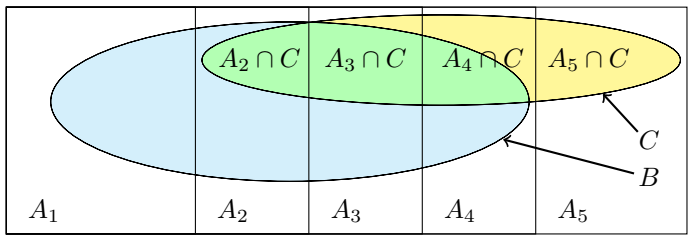
Otosavaruus S on ositettu erillisiin tapahtumiin A_i, $i = 1, \ldots, n$. Tapahtuman B todennäköisyys ehdolla tapahtuma C on

$$\mathsf{P}(B \mid C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C). \tag{4.19}$$

Todennäköisyys lasketaan nyt tapahtuman C rajaamassa otosavaruudessa (vrt. kokonaistodennäköisyyden kaava (4.14)). Oletus on, että $\mathsf{P}(C) > 0$ ja $\mathsf{P}(A_i \cap C) > 0$, $i = 1, \ldots, n$ (muuten kaikkia kaavan ehdollisia todennäköisyyksiä ei ole määritelty). Perustelu: $\mathsf{P}(B \mid C) = \mathsf{P}(B \cap C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \cap A_i \cap C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \cap C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C)\mathsf{P}(C)/\mathsf{P}(C) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C)$. Kuvan 4.18 Venn-diagrammin tilanteessa sininen ellipsi on tapahtuma B, keltainen ellipsi on tapahtuma C. Todennäköisyys $\mathsf{P}(B \mid C)$ on ellipsien vihreän leikkauksen alan suhde keltaisen ellipsin alaan. Kuvassa $\mathsf{P}(A_1 \cap C) = 0$, joten kaavassa (4.19) tulee asettaa $i = 2, \ldots, 5$. Lisäksi kuvan tilanteessa pätee $\mathsf{P}(B \mid A_2 \cap C) = 1$ ja $\mathsf{P}(B \mid A_5 \cap C) = 0$.



Kuva 4.17: Teorian oikeuden arvioinnin tulemat ja niiden todennäköisyydet.



Kuva 4.18: Ehdollisen kokonaistodennäköisyyden lasku.

Esimerkki. Terroristi iskee. Vastuulliseksi ilmoittautuvat kansainväliset terroristiryhmät $A_1, \ldots, A_n$, joista yksi on vastuullinen. Ryhmien aiemmasta aktiviteetista arvioituna ne ovat tekijöitä todennäköisyyksillä $\mathsf{P}(A_i)$. Iskun tekota-

pa ja siitä jäänyt muu todistusaineisto C viittaa ryhmään A_i todennäköisyydellä $\mathsf{P}(A_i \mid C)$ ja sen sisällä terroristin kansallisuuteen B todennäköisyydellä $\mathsf{P}(B \mid A_i \cap C)$. Se ei ole välttämättä sama kuin $\mathsf{P}(B \mid A_i)$, joka voisi olla vaikkapa B:n kansalaisten osuus A_i:ssä. Todennäköisyys, että terroristi on B:n kansalainen, lasketaan kaavalla (4.19). □

Esimerkki. Kybervakoilu. Helsingin Sanomat 15.1.2016, Kaleva 29.12.2020, Helsingin Sanomat 20.7.2021 ja Yle Uutiset 28.1.2022:[44]

> Kybervakoilijoiden ryhmä iskenyt ministeriöihin, yrityksiin ja suurlähetystöihin — jäljet johtavat Venäjälle.
>
> Eduskuntaan tehty kyberhyökkäys on vakava isku suoraan suomalaisen päätöksenteon ytimeen.
>
> Kiinan valtio oli Suomen eduskuntaan syksyllä 2020 kohdistuneen tietomurron takana – – .
>
> Suomalaisia diplomaatteja vakoiltu haittaohjelmalla – – Mahdollisesti on pystytty hyödyntämään jopa puhelimen mikrofonia ja kameraa sekä tallentamaan tieto, joka puhelimen kautta on kulkenut.

Ulkoministeriön sisäiseen verkkoon on murtauduttu. Hyökkäyksen takana voi olla mikä tahansa tietoteknisesti kehittynyt valtio: $S = \{A_1, \ldots, A_n\}$. Aiemman vakoiluaktiviteetin perusteella valtio A_i on syyllinen todennäköisyydellä $\mathsf{P}(A_i)$. Vakoiluohjelman piirteet C viittaavat valtioon A_i todennäköisyydellä $\mathsf{P}(A_i \mid C)$. Todennäköisyys, että kybervakoilija on valtio $B = A_j$ $(j = 1, \ldots, n)$, saadaan kaavasta (4.19). Koska $\mathsf{P}(B \mid A_j \cap C) = 1$ ja $\mathsf{P}(B \mid A_i \cap C) = 0$, jos $i \neq j$, todennäköisyys typistyy $\mathsf{P}(A_j \mid C)$:ksi. Ehdollistaminen ei välttämättä johda monimutkaiseen laskuun, vaikka se muuttaisi todennäköisyyttä. □

Olkoot kiinnostuksen kohteena tapahtumat A, B ja C. Bayesin kaavan (4.15) ehdollinen versio on

$$\begin{aligned} \mathsf{P}(A \mid B \cap C) &= \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)}{\mathsf{P}(B \mid C)} \\ &= \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)}{\sum_{i=1}^{n} \mathsf{P}(B \mid A_i \cap C)\mathsf{P}(A_i \mid C)}. \end{aligned} \tag{4.20}$$

Yllä oletetaan, että $\mathsf{P}(A \cap C) > 0$ ja $\mathsf{P}(B \cap C) > 0$, jotta ehdolliset todennäköisyydet ovat määriteltyjä. Viimeinen muoto olettaa taustalle lisäksi otosavaruuden jaon erillisiin tapahtumiin A_i $(i = 1, \ldots, n)$ ja että niihin liittyvät kaavassa esiintyvät ehdolliset todennäköisyydet ovat määriteltyjä.

Perustelu:

$$\mathsf{P}(A \mid B \cap C) \overset{1.}{=} \frac{\mathsf{P}(A \cap B \cap C)}{\mathsf{P}(B \cap C)} = \frac{\mathsf{P}(B \cap (A \cap C))}{\mathsf{P}(B \cap C)} \overset{3.}{=} \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \cap C)}{\mathsf{P}(B \cap C)}$$

$$= \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)\mathsf{P}(C)}{\mathsf{P}(B \mid C)\mathsf{P}(C)} \overset{5.}{=} \frac{\mathsf{P}(B \mid A \cap C)\mathsf{P}(A \mid C)}{\mathsf{P}(B \mid C)}.$$

Kolmas yhtäsuuruus seuraa tulosäännöstä (4.9) mieltämällä tapahtuma $A \cap C$ yhdeksi tapahtumaksi kuten tulosäännön yleistyksessä (4.13). Neljäs yhtäsuuruus seuraa niinikään tulosäännöstä. Sijoittamalla kaava (4.19) viimeiseen nimittäjään yllä saadaan haettu tulos (4.20). Ensimmäinen, kolmas ja viides yhtäsuuruus havainnollistavat, kuinka sama ehdollinen todennäköisyys voidaan laskea monella tavalla. Sopivin kannattaa valita tilanteen mukaan.

4.9 Simpsonin paradoksi

Kuva 4.19: Simpsonin paradoksi.[2]

Kuvan 4.19 piirtäjä on sivistynyt ja tuntee *Simpsonin paradoksin*. Sen kuvasivat Udny Yule 1903 ja Edward Simpson 1951. Simpsonin esimerkeissä osaaineistoissa on eroja mutta yhdistetyssä aineistossa ei. Ensimmäinen empiirinen esimerkki paradoksista on julkaistu 1934. Paradoksi on jo ennen nimeämistään tuottanut 1800-luvulla jupinan “valhe, emävalhe, tilasto” (Wainer ja Brown 2004). Paradoksin juju selviää esimerkistä.[45]

[2]Kuva on artikkelista A. Edwards (2007): A Cautionary Tale. *Significance*, maaliskuu 2007, 47–48. Kiitän *Significance*-lehteä (`www.significancemagazine.com`) luvasta painaa kuva.

Esimerkki.[46] Kaksi maisteria. B ja B^C ovat juuri valmistuneet maistereiksi. He ovat molemmat opiskelleet aineita C ja C^C. B:n pääaine oli C; B^C:n C^C. Merkitään hyvää arvosanaa A:lla ja huonoa A^C:lla. B:n ja B^C:n suorittamat kurssit arvosanojen mukaan ryhmiteltynä ovat alla:

	C			C^C			yhteensä		
	A	A^C	Σ	A	A^C	Σ	A	A^C	Σ
B	70	20	90	10	0	10	80	20	100
B^C	2	8	10	81	9	90	83	17	100

B on opiskellut enimmäkseen C:tä ja B^C enimmäkseen C^C:ia. Vinkeää todistuksissa on, että

- C:ssä B:n hyvien arvosanojen osuus $70/90 = 7/9$ on suurempi kuin B^C:n $2/10 = 1/5$.
- C^C:ssa B:n hyvien arvosanojen osuus $10/10 = 1$ on suurempi kuin B^C:n $81/90 = 9/10$.
- kokonaisuutena katsoen hyvien arvosanojen osuus $80/100$ B:llä on silti pienempi kuin $83/100$ B^C:lla! □

Jos osajoukoissa tapahtumien todennäköisyyksien suuruusjärjestys on päinvastainen kuin koko joukossa, on kohdattu Simpsonin paradoksi. Kolmen tapahtuman A, B ja C tilanteessa pätevät tällöin joko epäyhtälöt

$$\begin{aligned} \mathsf{P}(A \mid C \cap B) &> \mathsf{P}(A \mid C \cap B^C) \text{ ja} \\ \mathsf{P}(A \mid C^C \cap B) &> \mathsf{P}(A \mid C^C \cap B^C) \text{ mutta} \\ \mathsf{P}(A \mid B) &< \mathsf{P}(A \mid B^C) \end{aligned}$$

tai päinvastaiset epäyhtälöt. Simpsonin paradoksin voi ajatella ilmenevän myös, jos osajoukoissa pätee yhtäsuuruus muttei koko joukossa tai päinvastoin.

Selitys epäyhtälöille edellä on ehdollistettu kokonaistodennäköisyyden laki (4.19). Sen mukaan

$$\mathsf{P}(A \mid B) = \mathsf{P}(A \mid C \cap B)\mathsf{P}(C \mid B) + \mathsf{P}(A \mid C^C \cap B)\mathsf{P}(C^C \mid B)$$

ja

$$\mathsf{P}(A \mid B^C) = \mathsf{P}(A \mid C \cap B^C)\mathsf{P}(C \mid B^C) + \mathsf{P}(A \mid C^C \cap B^C)\mathsf{P}(C^C \mid B^C).$$

Vaikka pätisi

$$\mathsf{P}(A \mid C \cap B) > \mathsf{P}(A \mid C \cap B^C) \text{ ja } \mathsf{P}(A \mid C^C \cap B) > \mathsf{P}(A \mid C^C \cap B^C),$$

niin $\mathsf{P}(C \mid B)$ tai $\mathsf{P}(C^C \mid B)$ voi olla niin pieni, tai $\mathsf{P}(C \mid B^C)$ tai $\mathsf{P}(C^C \mid B^C)$ niin suuri, että

$$\mathsf{P}(A \mid B) < \mathsf{P}(A \mid B^C).$$

Esimerkki. Kaksi maisteria (jatkoa). Sijoitetaan tiedot osuuksista ehdollisen kokonaistodennäköisyyden kaavaan:

$$\begin{aligned}
\mathsf{P}(A \mid B) &= \mathsf{P}(A \mid C \cap B)\mathsf{P}(C \mid B) + \mathsf{P}(A \mid C^C \cap B)\mathsf{P}(C^C \mid B) \\
&= \frac{70}{90} \times \frac{90}{100} + \frac{10}{10} \times \frac{10}{100} \\
&= \frac{80}{100} \\
&< \\
&= \frac{83}{100} \\
&= \frac{2}{10} \times \frac{10}{100} + \frac{81}{90} \times \frac{90}{100} \\
&= \mathsf{P}(A \mid C \cap B^C)\mathsf{P}(C \mid B^C) + \mathsf{P}(A \mid C^C \cap B^C)\mathsf{P}(C^C \mid B^C) \\
&= \mathsf{P}(A \mid B^C).
\end{aligned}$$

B:n todistuksessa aineen C suuri osuus $90/100 = \mathsf{P}(C \mid B)$ painottaa kohtuullista arvosanasuhdetta $70/90$ ja aineen C^C pieni osuus $10/100 = \mathsf{P}(C^C \mid B)$ mitätöi loistavaa arvosanasuhdetta $10/10$. B^C:n todistuksessa aineen C pieni osuus mykistää huonoa arvosanasuhdetta $2/10$ ja aineen C^C suuri osuus $90/100 = \mathsf{P}(C^C \mid B^C)$ rummuttaa hienoa arvosanasuhdetta $81/90$. Seuraus on epäyhtälö yllä. □

Esimerkki. Hyvä miehille, hyvä naisille, huono ihmisille (Baker ja Kramer 2001). Verrataan syöpähoitoja B ja B^C. Merkitään miehiä C:llä, naisia C^C:lla, syövästä paranemista A:lla ja syöpään kuolemista A^C:lla. Sivun 74 taulukon mukaan B on B^C:a parempi hoito sekä miehille että naisille. Ihmisille B vaikuttaa huonommalta hoidolta. □

Esimerkki. Hyvä miehille, hyvä naisille, huono ihmisille (jatkoa). Syöpäesimerkissä B^C-hoito vaikutti paremmalta ihmisille ylipäänsä. Huolellisempi tarkastelu paljastaa, että B on parempi hoito. Huomataan, että syöpähoidot tehoavat

erilailla miehiin ja naisiin ja että heitä on aivan eri osuudet syöpähoidoissa (s:n 74 taulukko). Idea: Standardoidaan hoitoihin osallistuvien sukupuolten osuudet samoiksi ja verrataan hoitojen tehokkuutta standardoiduilla osuuksilla.

Kuva 4.20 havainnollistaa. Siinä on laskettu onnistuneiden hoitojen osuudet, jos hoito toimii sivun 74 taulukon mukaisilla sukupuolikohtaisilla osuuksilla ja naisten (C^C) osuus tutkimuksessa vaihtelee välillä $[0, 1]$.

Punainen suora

$$(1-w)\times\frac{7}{9}+w\times\frac{1}{1}=\frac{7}{9}+\frac{2}{9}w$$

kuvaa paranemisten osuutta hoidolla B ja sininen suora

$$(1-w)\times\frac{2}{10}+w\times\frac{81}{90}=\frac{1}{5}+\frac{7}{10}w$$

hoidolla B^C. Yllä w on naisten osuus hoidossa. Suorat ovat painotettuja keskiarvoja miesten ja naisten paranemisosuuksista kyseisellä hoitomuodolla. Kuvaan on merkitty pisteillä ja katkoviivoilla sivun 74 taulukon mukaisia osuuksia:

$$(1-\frac{10}{100})\times\frac{7}{9}+\frac{10}{100}\times\frac{1}{1}=0.9\times\frac{7}{9}+0.1=0.8$$

(C^C:n osuus 0.1) ja

$$(1-\frac{90}{100})\times\frac{2}{10}+\frac{90}{100}\times\frac{81}{90}=0.1\times 0.2+0.9\times 0.9=0.83$$

(C^C:n osuus 0.9). Punainen suora on sinisen suoran yläpuolella kaikilla naisosuuksilla.

Naisten osuus hoidoissa määrää, kumpi hoidoista vaikuttaa useammin parantavan syövän ihmisillä. Molemmat hoidot toimivat naisilla paremmin kuin miehillä ($10/10 > 70/90$ ja $81/90 > 2/10$), ja erityisen hyvin naisilla toimii hoito B (parantuneiden osuus $10/10 = 1$). Miehillä hoito B^C toimii erityisen huonosti (parantuneiden osuus $2/10 = 0.2$). Jos naisten osuus on hyvin pieni (0.1) hoidossa B ja hyvin suuri (0.9) hoidossa B^C, niin kokonaiskuva hoitojen tehokkuudesta vääristyy ja hoito B^C näyttää hoitoa B paremmalta ($83/100 > 80/100$). Jos naisten ja miesten osuudet hoitoihin osallistumisessa poikkeaisivat vähemmän, hoito B olisi yleensä parempi. Kuvaan on piirretty esimerkkinä tilanne, jossa naisten ja miesten osuudet hoidoissa ovat yhtäsuuret. Hoito B on tällöin selvästi parempi ($0.889 > 0.55$): Hoidolla B parantumisosuus on

$$0.5\times\frac{70}{90}+0.5\times\frac{10}{10}\approx 0.889$$

ja hoidolla B^C

$$0.5 \times \frac{2}{10} + 0.5 \times \frac{81}{90} = 0.55.$$

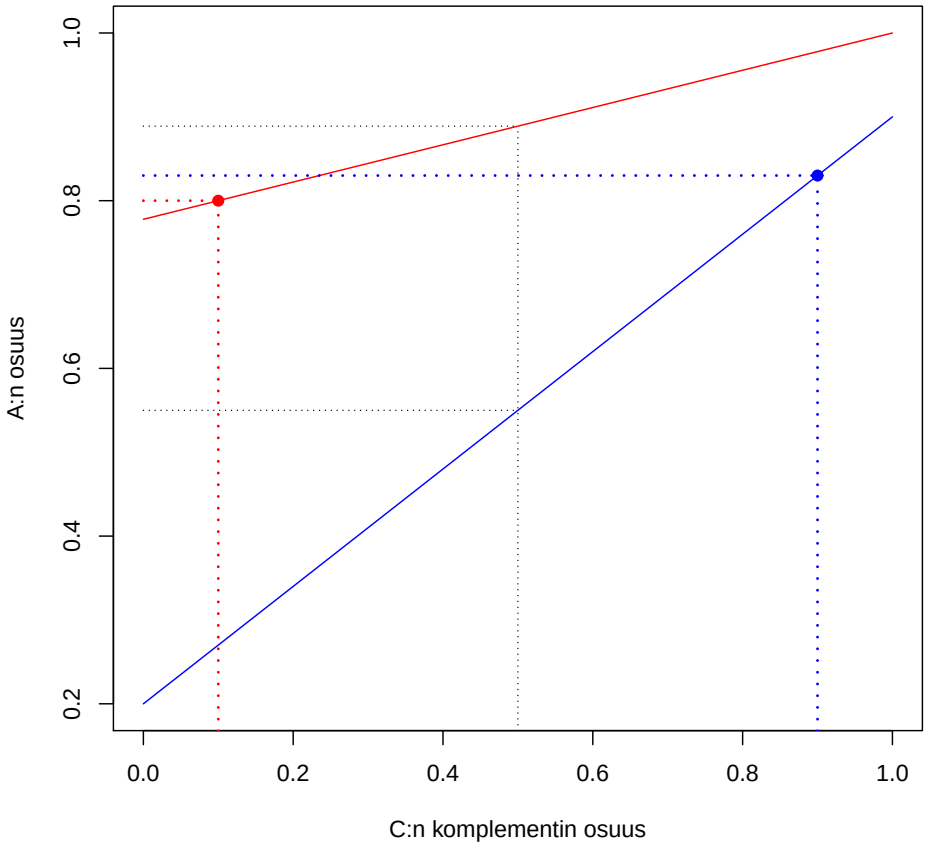
Jos miehet ja naiset voidaan ohjata vapaasti kumpaan tahansa hoitoon, heidät kannattaa ohjata hoitoon B. □

Esimerkki. Hyvä miehille, hyvä naisille, huono ihmisille (jatkoa). Merkitään A:lla ja A^C:lla pitkä- ja lyhytikäisyyttä sekä B:llä ja B^C:lla geenin B omaamista ja puuttumista. Miehiä ja naisia osoitetaan edelleen C:llä ja C^C:lla. Geeni on yleinen miehillä mutta harvinainen naisilla, ja geenin omaavat miehet ja naiset ovat pitkäikäisempiä kuin omaamattomat (s:n 74 taulukko). Väestössä geenin omaamattomat ovat silti pitkäikäisempiä kuin geenin omaavat. Geenin omaamisen osuutta sukupuolittain tai sukupuolien osuutta väestössä ei voi muuttaa. Geeni B on siten ylitsepääsemättömästi hyvä naisille, hyvä miehille mutta huono ihmisille. □

Simpsonin paradoksin tulkinta on tapauskohtainen. Ylipäänsä ehdollistettu (tarkemmin luokiteltu) tieto on informatiivisempaa, ja siihen kannattaa kiinnittää huomiota enemmän kuin ehdollistamattomaan (vähemmän tarkasti luokiteltuun). Silti luokittelematon tieto kuvaa kokonaisuuden.

Esimerkki.[47] Empiirisiä esimerkkejä Simpsonin paradoksista:

- Urheilussa pelaaja voi pelata toista pelaajaa paremmin kauden ensimmäisellä ja toisella kaudella mutta kokonaisuutena huonommin.
- Tupakoijien kuolleisuus on kaikissa ikäryhmissä suurempi kuin tupakoimattomien. Tupakoijien kuolleisuus ylipäänsä on kuitenkin pienempi kuin tupakoimattomien, koska tupakoijat ovat keskimäärin tupakoimattomia nuorempia. (Cochran 1968.)
- Yhdysvalloissa Floridan osavaltiossa 1976–1986 valkoihoiset tuomittiin mustaihoisia useammin kuolemaan mutta aineiston alaryhmissä mustaihoiset. (Agresti 2019.)
- Berkeleyn yliopisto haastettiin oikeuteen sukupuolisyrjinnästä vuoden 1973 opiskelijavalinnan takia. Mieshakijoista oli hyväksytty selvästi suurempi osuus kuin naishakijoista. Asiaa tutkittaessa ilmeni, että monilla laitoksilla naisten hyväksymisprosentit olivat suurempia kuin miesten. Selitys miesten suurempaan hyväksymisprosenttiin ylipäänsä oli, että he pyrkivät oppiaineisiin (esim. matemaattisiin tai teknisiin tieteisiin), joille oli helpompi päästä (hyväksyttyjen osuudella mitattuna) kuin naisten suosimiin oppiaineisiin (esim. psykologia). (Bickel ym. 1978.)



Kuva 4.20: A:n osuuden riippuvuus C^C:n osuudesta.

- Helsingin yliopiston valtiotieteellisessä tiedekunnassa miehet saavat pro gradu -tutkielmista parempia arvosanoja kuin naiset. Ero häviää, jos verrataan arvosanoja oppiaineittain. Ilmiö selittyy näin: Taloustieteessä myön-

netään korkeita arvosanoja ja valtaosa opiskelijoista on miehiä. Sosiaalityössä arvosanat ovat huonompia ja opiskelijoiden enemmistö on naisia. Molemmissa oppiaineissa sukupuolet saavat yhtälailla arvosanoja. (Holm 2013.) □

Luku 5

Kombinatoriikkaa

> Termi todennäköisyyslaskenta on paradoksi. Varmuudesta eroten todennäköisyys on sitä, mitä emme tiedä. Kuinka voimme laskea tuntemattoman?[48]
>
> Henri Poincaré (1854–1912)

Mistä todennäköisyydet tulevat? Monissa yhteyksissä todennäköisyys voidaan päätellä kombinatorisilla laskuilla. Muun muassa myöhemmin esitettävät sovellusten kannalta tärkeät binomi-, multinomi- ja hypergeometriset jakaumat seuraavat tällaisista laskuista. Kombinatoriikkaa tarvitaan tilastotieteessä myös arvioitaessa vastauksia sen tapaisiin kysymyksiin, kuinka monta regressiota (luku 13) voidaan tehdä. Periaatteille alla on paljon käyttöä.

Yhteenlaskuperiaate: Oletetaan, että

- operaatiot A ja B ovat erillisiä eli että niistä vain toinen voidaan suorittaa.
- A voidaan suorittaa n_1:llä ja B n_2:lla tavalla.

Tällöin yhdistetty operaatio "A tai B" voidaan suorittaa $(n_1 + n_2)$:lla tavalla.

Kertolaskuperiaate: Oletetaan, että

- operaatiot A ja B voidaan suorittaa toisistaan riippumattomasti.
- A ja B voidaan suorittaa n_1:llä ja n_2:lla tavalla.

Tällöin yhdistetty operaatio "A ja B" voidaan suorittaa $(n_1 \times n_2)$:lla tavalla.

Esimerkki. Maisteriopinnot. Opiskelija pohtii, miten erikoistua maisterivaiheessa. Linjoja on kaksi. Vaihtoehdon A voi suorittaa 56:lla ja vaihtoehdon B 126:lla

erilaisella kurssikombinaatiolla. (Tarkoitus on, että suoritetaan vain toisen linjan opinnot.) Yhteensä opiskelijalla on $56 + 126 = 182$ erilaista mahdollisuutta suorittaa maisteriopinnot. (Syy kurssikombinaatioiden lukumäärille selviää myöhemmässä esimerkissä.) □

Esimerkki. Maisteriopinnot (jatkoa). Opiskelija on niin innostunut, että pohtii, suorittaisiko (tarkoitetunvastaisesti) molempien linjojen opinnot. Hän voisi suorittaa ne $56 \times 126 = 7\ 056$ erilaisella tavalla. □

Esimerkki. Maisteriopinnot (jatkoa). Opiskelijalla on sivuaineopintoja suorittamatta. Ne voi suorittaa 6 tavalla. Opiskelija voisi suorittaa pääaineensa kahden linjan ja sivuaineensa opinnot $7\ 056 \times 6 = 42\ 336$ tavalla. □

Koostukoon joukko A erilaisista alkioista $a_1, \ldots, a_n$. Ne voidaan järjestää

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 \tag{5.1}$$

erilaiseen *jonoon* eli *permutaatioon*. Merkintä $n!$ luetaan "n:n kertoma".

Perustelu: Jonon ensimmäiseksi alkioksi on mahdollista valita n alkiota, toiseksi alkioksi mahdollisuuksia on $n-1$, kolmanneksi $n-2$ jne. Toiseksi viimeinen alkio joudutaan valitsemaan kahden alkion väliltä. Viimeiseksi alkioksi jää aiemmin valitsematon alkio. Kertolaskuperiaatteen mukaan kaksi ensimmäistä operaatiota — tässä järjestämistä — voidaan tehdä $n \times (n-1)$ tavalla. Kolmas operaatio voidaan ajatella koostuvan ensimmäisestä yhdistetystä operaatiosta, joka voidaan tehdä $[n \times (n-1)]$ tavalla ja toisesta operaatiosta, joka voidaan tehdä $n-2$ tavalla. Kertolaskuperiaate tuottaa mahdollisten järjestettyjen jonojen lukumääräksi $[n \times (n-1)] \times (n-2)$. Neljäs operaatio ajatellaan jälleen koostuvaksi ensimmäisestä yhdistetystä operaatiosta $[n \times (n-1) \times (n-2)]$ vaihtoehdolla ja seuraavasta operaatiosta $(n-3)$ vaihtoehdolla. Näin jatkaen päädytään kaavaan yllä.

Esimerkki. Kertomia. 5:n kertoma on 120:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120.$$

Kertomat suurenevat nopeasti:

$$8! = 8 \times 7 \times \cdots \times 2 \times 1 = 40\ 320 \quad \text{ja}$$
$$14! = 14 \times 13 \times \cdots \times 2 \times 1 = 87\ 178\ 291\ 200.$$

Erikoistapauksena määritellään 0:n kertomaksi 1:

$$0! = 1.$$

Laskujen tulokset voi tarkistaa R:llä käskyillä factorial(5), factorial(8), factorial(14) ja factorial(0). □

Esimerkki. Maisteriopinnot (jatkoa). Linjalla A luennoidaan 8 kurssia. Extrainnokas opiskelija pohtii, kävisikö kaikki. Kuinka monessa järjestyksessä hän voisi suorittaa ne? Järjestyksiä on $8! = 40\ 320$. □

Monesti on tarpeen selvittää, kuinka monta k:n ($k \leq n$) pituista permutaatiota voidaan muodostaa n:stä erilaisesta alkiosta. Lukumäärä voidaan päätellä kuten edellä mutta katkaisemalla valintojen tekeminen k:n alkion järjestämisen jälkeen:

$$\begin{aligned} & n \times (n-1) \times \cdots \times [n-(k-2)] \times [n-(k-1)] \\ &= n \times (n-1) \times \cdots \times (n-k+2) \times (n-k+1) \\ &= \frac{n \times \cdots \times (n-k+1) \times (n-k) \times \cdots \times 1}{(n-k) \times \cdots \times 1} \\ &= \frac{n!}{(n-k)!}. \end{aligned} \tag{5.2}$$

Ensimmäisellä rivillä kerrottavia on k kappaletta.

Esimerkki. Maisteriopinnot (jatkoa). Linjan 8 kurssista 5 täytyy suorittaa maisterin tutkintoa varten. Kuinka monta vaihtoehtoista suoritusjärjestystä opiskelijalla on 5 kurssille?

Järjestyksiä on 6 720:

$$\frac{8!}{(8-5)!} = \frac{8!}{3!} = 8 \times 7 \times 6 \times 5 \times 4 = 6\ 720.$$

Samaan vastaukseen päätyy järkeilemällä, että ensimmäiseksi kurssiksi on 8 vaihtoehtoa, seuraavaksi 7 ja niin edespäin viidenteen kurssiin asti, johon on jäljellä $8 - 4 = 4$ vaihtoehtoa. Kertolaskuperiaatetta soveltamalla vastaus on $8 \times 7 \times 6 \times 5 \times 4$. □

Esimerkki. Maisteriopinnot (jatkoa). Opiskelijaa kiinnostaa linjan 8 kurssista vain 5. Kuinka monessa järjestyksessä opiskelija voi suorittaa häntä kiinnostavat 5 kurssia?

Nyt vaihtoehtoisia kursseja on vain 5. Opiskelija voi tenttiä ne 120 järjestyksessä:

$$\frac{5!}{(5-5)!} = \frac{5!}{0!} = \frac{5!}{1} = 120.$$

Yltä ilmenee yksi syy, miksi oli hyödyllistä määritellä $0! = 1$: Sama laskusääntö $n!/(n-k)!$ pätee myös tilanteessa $n = k$. □

Olkoon joukossa A erilaisia alkioita n kappaletta: $A = \{a_1, \ldots, a_n\}$. Kuinka monta k:n kokoista ($0 \leq k \leq n$) erilaista osajoukkoa voidaan A:sta poimia? Vastaus on

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{5.3}$$

erilaista osajoukkoa. Suure $\binom{n}{k}$ on *binomikerroin*. Se luetaan "n yli k:n".

Tulos päätellään näin: Merkitään tuntematonta osajoukkojen lukumäärää N:llä. Kustakin osajoukosta voidaan muodostaa $k!$ permutaatiota (kaava (5.1)). Kertolaskuperiaatteen mukaan kaikista osajoukoista saadaan $N \times k!$ permutaatiota. Toisaalta n alkiosta voidaan muodostaa k:n pituisia permutaatiota $n!/(n-k)!$ (kaava (5.2)). Täytyy päteä $N \times k! = n!/(n-k)!$. Ratkaisemalla yhtälöstä N saadaan tulos (5.3).

Esimerkki. Maisteriopinnot (jatkoa). Linjalla A luennoidaan 8 ja linjalla B 9 kurssia. Kursseja täytyy suorittaa 5. Kuinka monta erilaista kurssikokonaisuutta linjaa A opiskeleva voi muodostaa 5 kurssista? Entä linjalla B opiskeleva? Sivuainekursseja on tarjolla 4, joista 2 täytyy suorittaa. Kuinka monta 2 kurssin kombinaatiota sivuaineen kursseista voi muodostaa?

Linjalla A voi muodostaa 56 ja linjalla B 126 erilaista kurssikokonaisuutta:

$$\binom{8}{5} = \frac{8!}{5!(8-5)!} = \frac{8!}{5!3!} = \frac{8 \times 7 \times 6}{3 \times 2} = 56 \quad \text{ja}$$
$$\binom{9}{5} = \frac{9!}{5!(9-5)!} = \frac{9!}{5!4!} = \frac{9 \times 8 \times 7 \times 6}{4 \times 3 \times 2} = 126.$$

Sivuaineen 2 kurssin kombinaatioita on 6:

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{4 \times 3}{2} = 6.$$

Laskut voi tarkistaa R-komennoilla `choose(8,5)`, `choose(9,5)` ja `choose(4,2)`. Näitä lukumääriä käytettiin ensimmäisessä ja kolmannessa Maisteriopinnot-esimerkissä. □

Esimerkki. Maisteriopinnot (jatkoa). Linjan B kurssit jakaantuvat 5 teoreettiseen ja 4 empiiriseen kurssiin. Teoreettisia kursseja pitää käydä 3 ja empiirisiä 2. Kuinka monta erilaista kurssikombinaatiota linjan B opiskelija voi suorittaa (5 kurssista)?

Opiskelija voi valita teoreettisia kursseja 10 tavalla:

$$\binom{5}{3} = \frac{5!}{3!2!} = \frac{5 \times 4}{2} = 10.$$

Empiirisissä kursseissa on 6 valintamahdollisuutta:

$$\binom{4}{2} = 6.$$

Kertolaskuperiaatteen mukaan opiskelija voi muodostaa 60 erilaista opintokokonaisuutta B-linjan maisteriopinnoista:

$$\binom{5}{3}\binom{4}{2} = 10 \times 6 = 60.$$

□

Perustellaan, että binomikerroin (5.3) on myös on erilaisten n:n pituisten jonojen lukumäärä kahdenlaisista (vaikkapa oransseista ja vihreistä) alkioista o ja v, joita on k ja $n - k$ kappaletta: Merkitään erilaisten jonojen lukumäärää N:llä. Mikäli voitaisiin erotella o-alkiot toisistaan, olisi erilaisia jonoja $N \times k!$ kappaletta, sillä o-alkiot voidaan järjestää $k!$ eri tavalla yhdessä jonossa (kaava (5.1)). Mikäli voitaisiin erotella myös v-alkiot, olisi erilaisia jonoja $N \times k! \times (n-k)!$ kappaletta, sillä v-alkiot voidaan järjestää $(n-k)!$ eri tavalla yhdessä jonossa. Tällöin pystyttäisiin erottelemaan kaikki alkiot, jolloin erilaisia jonoja on $n!$. Näin ollen täytyy päteä

$$N \times k! \times (n-k)! = n!$$

eli

$$N = \frac{n!}{k! \times (n-k)!} = \binom{n}{k}. \tag{5.4}$$

Päätellään seuraavaksi, kuinka monella tavalla n erilaista alkiota voidaan jakaa k:hon erilaiseen osajoukkoon, joiden koko on $n_1, \ldots, n_k$ ($\sum_{i=1}^{k} n_i = n$).

Kertolaskuperiaatteesta seuraa, että jakojen lukumäärä on

$$\begin{aligned}
&\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-n_1-n_2}{n_3} \times \cdots \times \\
&\binom{n-n_1-n_2-\cdots-n_{k-2}}{n_{k-1}} \times \binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k} \\
=&\frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \times \frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \times \cdots \times \\
&\frac{(n-n_1-\cdots-n_{k-2})!}{n_{k-1}!(n-n_1-\cdots-n_{k-1})!} \times \frac{(n-n_1-\cdots-n_{k-1})!}{n_k!0!} \\
=&\frac{n!}{n_1!n_2!n_3!\ldots n_{k-1}!n_k!}.
\end{aligned}$$

Neljännellä rivillä on sijoitettu $(n-n_1-n_2-\cdots-n_{k-1})-n_k = n_k - n_k = 0$. Peräkkäisissä osamäärissä nimittäjien ja osoittajien vastaavat termit supistuvat. Laskettu lukumäärä on *multinomikerroin*:

$$\binom{n}{n_1,\ldots,n_k} = \frac{n!}{n_1!\ldots n_k!}. \tag{5.5}$$

Multinomikerroin voidaan binomikertoimen tapaan tulkita toisinkin: Multinomikerroin on n:n pituisten permutaatioiden lukumäärä, kun alkioita on k:nlaisia ja kussakin osajoukossa on n_i alkiota ($i = 1,\ldots,k$ ja $\sum_{i=1}^{k} n_i = n$). Perustelu on samanlainen kuin binomikertoimen vastaavalle tulokselle. Samaan tapaan saadaan yhtälö

$$N \times n_1! \times n_2! \times \ldots \times n_k! = n!.$$

(N on permutaatioiden lukumäärä, $N \times n_1!$ on permutaatioiden lukumäärä, jos 1. osajoukon alkiot voitaisiin erotella, $N \times n_1! \times n_2!$, jos myös 2. osajoukon alkiot voitaisiin erotella jne.) Yhtälöstä ratkaistaan

$$N = \frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}. \tag{5.6}$$

Esimerkki. Maisteriopinnot (jatkoa). Linjan B kurssit jakaantuvat 5 teoreettiseen ja 4 empiiriseen kurssiin. Teoreettisia kursseja pitää käydä 3 ja empiirisiä 2. Linjan opiskelija on valinnut, mitkä 3 teoreettista ja 2 empiriistä kurssia ja mitkä 2 kurssia jäljellä olevista sivuaineopinnoistaan hän tenttii. Hän haluaa vaihtelua

ja pohtii, kuinka monessa järjestyksessä hän voisi tenttiä nämä $3 + 2 + 2 = 7$ kurssia. Vastaus “210:ssä järjestyksessä” saadaan multinomikertoimesta (5.6):

$$\binom{7}{3,2,2} = \frac{7!}{3!2!2!} = \frac{7 \times 6 \times 5 \times 4}{4} = 210. \quad \square$$

Binomikerroin on erikoistapaus multinomikertoimesta:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{n}{k, n-k}.$$

Binomikertoimet saadaan Pascalin kolmiosta:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

Ykkösestä poikkeavat luvut ovat kolmannesta rivistä alkaen luvun yllä olevien kahden luvun summa. Luvut ovat binomikertoimia kaaviossa alla:

$$\begin{array}{ccccccccccc}
 & & & & & \binom{0}{0} & & & & & \\
 & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
 & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} \\
 & & & & & \vdots & & & & &
\end{array}$$

Binomikerroin-nimitys tulee siitä, että binomikertoimet ovat kertoimia binomin $(x + y)^n$

aukikirjoitetussa muodossa

$$\begin{aligned}(x+y)^n =& \binom{n}{0}x^n y^0 + \binom{n}{1}x^{n-1}y^1 + \binom{n}{2}x^{n-2}y^2 \\ &+ \cdots + \binom{n}{n-1}x^1 y^{n-1} + \binom{n}{n}x^0 y^n \\ =& \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i.\end{aligned} \tag{5.7}$$

Binomilause (5.7) perustellaan näin: Termi $x^{n-i}y^i$ muodostuu kerrottaessa tulossa

$$(x+y)^n = (x+y) \times (x+y) \times \cdots \times (x+y) \times (x+y)$$

i kappaletta y:itä ja $n-i$ kappaletta x:iä keskenään. Tällainen tulo voidaan muodostaa lausekkeesta yllä $\binom{n}{i}$ tavalla. Tähän päädytään pohtimalla, kuinka monesta järjestyksestä i kappaletta y:itä ja $n-i$ kappaletta x:iä tulo $x^{n-i}y^i$ syntyy. Vastaus on binomikerroin $\binom{n}{i}$. Vaihtoehtoisesti voi ajatella, että poimitaan i kappaletta y:itä $(x+y)$-termeistä ja jäljellejäävistä $n-i$:stä $(x+y)$-termistä x:t. Binomikerroin $\binom{n}{i}$ kertoo, kuinka monta i:n kokoista osajoukkoa eriväristä y:tä voidaan poimia n:stä eriväristä y:stä. Kukin osajoukko vastaa tiettyä valintaa y:itä ja x:iä tulossa yllä. Kaava (5.7) seuraa käymällä jompikumpi argumentti läpi i:n arvoille $0, \ldots, n$.

Esimerkki. Binomi astetta 5.

$$\begin{aligned}(x+y)^5 =& \binom{5}{0}x^5 + \binom{5}{1}x^4y^1 + \binom{5}{2}x^3y^2 + \binom{5}{3}x^2y^3 + \binom{5}{4}x^1y^4 + \binom{5}{5}y^5 \\ =& x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5. \quad \square\end{aligned}$$

Luku 6

Diskreetit ja jatkuvat satunnaismuuttujat

> – – ei mikään näytä niin hienolta kuin pätkä kreikkaa. Kirjaimet henkivät jo sinänsä syvällisyyttä. Huomatkaa vaikka vain *epsilonin* neuvokas ilme! *Fiin* pitäisi tosiaankin olla piispa! Onko terävämpää kaveria ollut kuin *omikron*? Katsokaa nyt tuota *tauta*![49]
>
> Edgar Allan Poe (1809–1849)

6.1 Todennäköisyysjakauma ja kertymäfunktio

Satunnaismuuttuja (jakso 4.1) voi olla *diskreetti* tai *jatkuva*.[50] Satunnaismuuttujan arvojen todennäköisyyksiä kuvataan *todennäköisyysjakaumalla* (*probability distribution function*). Diskreettiä todennäköisyysjakaumaa kutsutaan *pistetodennäköisyysfunktioksi* (*probability mass function*); jatkuvaa *tiheysfunktioksi* (*probability density function*, *density function*).

Diskreetillä satunnaismuuttujalla X on äärellinen (k) määrä lukuarvoja, $x_1 < \cdots < x_k$, joita se voi saada.[51] Pistetodennäköisyysfunktio kuvaa todennäköisyydet, joilla se saa kunkin arvon x_i ($i = 1, \ldots, k$):

$$\mathsf{P}(X = x_i) = \pi_i.$$

Lisäksi pätee $0 \leq \pi_i \leq 1$ ja $\sum_{i=1}^{k} \pi_i = 1$.

Diskreetin satunnaismuuttujan *kertymäfunktio* (*cumulative distribution func-*

tion) on

$$F(x) = \mathsf{P}(X \leq x) = \sum_{i=1}^{\mathsf{int}[x]} \pi_i.$$

Summassa $\mathsf{int}[x]$ on argumentin x kokonaislukuosa (esim. $\mathsf{int}[5.6] = 5$).

Kertymäfunktion yllä ominaisuuksia ovat:

1. Kertymäfunktio saa arvoja nollan ja yhden välillä ($0 \leq F(x) \leq 1$).
2. Kertymäfunktio on kasvava (ei-vähenevä) funktio ($F(a) \leq F(b)$, jos $a < b$, jossa a ja b ovat reaalilukuja).
3. Kertymäfunktion arvo suppenee nollaan, kun kertymäfunktion argumentti menee kohti miinus ääretöntä ($\lim_{x \to -\infty} F(x) = 0$).
4. Kertymäfunktion arvo suppenee yhteen, kun kertymäfunktion argumentti menee kohti ääretöntä ($\lim_{x \to \infty} F(x) = 1$).
5. Kertymäfunktio on porrasfunktio niin, että $F(x_i) - F(x_{i-1}) = \pi_i$.

Kertymäfunktion arvo kasvaa siis portaittaisesti nollasta yhteen argumentin x kasvaessa miinus äärettömästä äärettömään. Jaksossa 7.1 on esimerkkejä diskreettien satunnaismuuttujien pistetodennäköisyys- ja kertymäfunktioista.

Jatkuva satunnaismuuttuja saa arvoja jatkuvasti reaalilukujen joukossa. Joukko voi olla rajattu tietylle välille tai tietyille väleille.[52] Tiheysfunktio $f(x)$ kuvaa, kuinka satunnaismuuttujan X arvojen todennäköisyydet jakautuvat. Tiheysfunktion ominaisuuksia ovat:

1. Tiheysfunktio saa vain ei-negatiivisia arvoja ($f(x) \geq 0,\ -\infty < x < \infty$).
2. Tiheysfunktion kuvaajan ja x-akselin välinen pinta-ala on 1.
3. Todennäköisyys, että satunnaismuuttujan toteuma osuu välille $[a, b]$ on tiheysfunktion kuvaajan ja x-akselin välin $[a, b]$ rajaaman alueen pinta-ala.

Tiheysfunktio ei kuvaa yksittäisten lukuarvojen x todennäköisyyksiä. Kunkin lukuarvon todennäköisyys on 0, sillä mahdollisia lukuarvoja on äärettömästi.

Jatkuvan satunnaismuuttujan X kertymäfunktio toteuttaa diskreetin tilanteen tapaan yhtälön

$$F(x) = \mathsf{P}(X \leq x).$$

Sillä on diskreetin satunnaismuuttujan kertymäfunktion ominaisuudet 1–4 muttei 5:dettä. Tällaisen kertymäfunktion arvo pisteessä x on pinta-ala, jonka rajaavat tiheysfunktion kuvaaja, x-akseli ja pisteen x kohdalle piirretty pystysuora viiva. Jaksossa 7.2 on esimerkkejä jatkuvien satunnaismuuttujien tiheys- ja kertymäfunktioista.

Jakaumaa kuvataan usein *kvantiileilla* (*quantile*) eli *fraktiileilla* (*fractile*) ξ_q. Jos jakaumassa ei ole katkoksia tai hyppyjä, jakauman q. kvantiili ξ_q toteuttaa yhtälöt

$$F(\xi_q) = \mathsf{P}(X \leq \xi_q) = q, \tag{6.1}$$

jossa $0 < q < 1$. Jos kertymäfunktio on aidosti kasvava, q. kvantiili on yksikäsitteinen. *Kvantiilifunktio* (*quantile function*) kuvaa kertymäfunktion arvot q vastaaviksi satunnaismuuttujan arvoiksi ξ_q. Kertymä- ja kvantiilifunktio ovat siis toistensa käänteisfunktioita.

Edellä ξ on kreikkalaisten aakkosten kirjain “ksii”. Tapa merkitä jakaumaa kuvaavat parametrit (jakso 6.2) kreikkalaisilla kirjaimilla juontaa muun muassa Studentin (1908) ja Fisherin (1922a) artikkeleihin (Aldrich 2003, 2005).[53]

Jatkuva-arvoinen satunnaismuuttuja saa (tavanomaisissa tilanteissa) todennäköisyydellä q pienemmän tai yhtäsuuren arvon kuin q. kvantiili ξ_q. *Kvartiileja* (*quartile*) ovat 0.25., 0.50. ja 0.75. kvantiili (1., 2. ja 3. kvartiili); *desiilejä* (*decile*) kvantiilit $0.10, 0.20, \ldots, 0.90$ $(1., 2., \ldots, 9.$ desiili$)$.[54] Jos osuus q kuuluu joukkoon $\{1\ \%, 2\ \%, \ldots, 99\ \%\}$, kvantiilia kutsutaan *prosentiiliksi* (*percentile*) $(1., 2., \ldots, 99.$ prosentiili$)$.

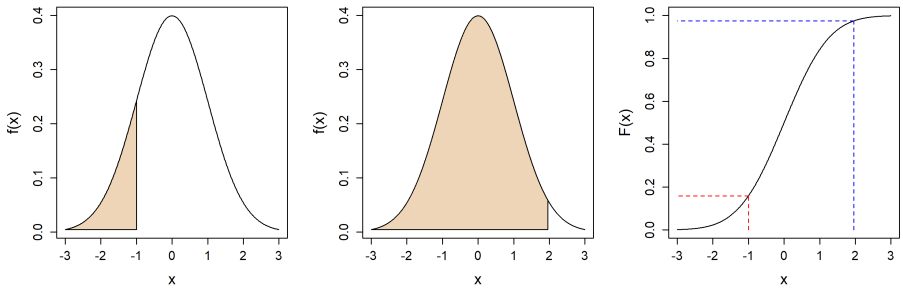
Esimerkki.[55] Kuvassa 6.1 osoitetaan värillä tiheysfunktion ja x-akselin väliin jäävä pinta-ala eli todennäköisyys, kun $x = -1$ tai $x = 1.960$. Kuvassa oikealla on tiheysfunktiota vastaava kertymäfunktio. Värikatkoviivoilla osoitetaan x:n arvoihin liittyvät kertymäfunktion arvot 0.159 ja 0.975. Niitä vastaavat kvantiilit ovat -1 ja 1.960. (Tiheysfunktio on jaksossa 7.2.1 määriteltävä standardinormaalijakauma.) □

Jatkuvan satunnaismuuttujan kertymäfunktio ja siitä seuraavat todennäköisyydet määritellään integraalien avulla (kun ne ovat olemassa). Jatkuvan satunnaismuuttujan X kertymäfunktio on

$$F(x) = \mathsf{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,dt, \quad -\infty < x < \infty.$$

Tiheysfunktion kuvaajan ja x-akselin välinen pinta-ala on 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$



Kuva 6.1: Tiheysfunktio, sen rajaama pinta-ala eli todennäköisyys ja kertymäfunktio.

Todennäköisyys, että jatkuvan satunnaismuuttujan toteuma osuu välille $[a, b]$ on

$$\mathsf{P}(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Integraalit lasketaan paloittain, mikäli satunnaismuuttujan jakaumassa on hyppäyksiä tai katkoksia.

6.2 Satunnaismuuttujan sijainti- ja vaihtelumittoja

Satunnaismuuttujan todennäköisyysjakauma voi olla itsessään kiinnostava ja tärkeä. Monesti kiinnostus kohdistuu kuitenkin *populaation* numeerisesti mitattavissa olevaan yksittäiseen piirteeseen tai piirteisiin, joita kuvaa yksi tai useampi *parametri*. Tilastotieteessä populaatiolla tarkoitetaan asiaintilaa, josta ollaan kiinnostuneita ja jota kuvataan tilastollisella mallilla. Nyt esillä olevassa yhteydessä populaatio voidaan samaistaa kiinnostuksen kohteena olevan satunnaisilmiön todennäköisyysjakaumaksi. Populaation käsitteeseen palataan luvussa 8. Populaatiota kuvaavan jakauman odotusarvo, usein μ (“myy”), ja varians-

si, usein σ^2 (“sigma toiseen”), ovat jakauman keskeisiä piirteitä ja esimerkkejä parametreista.

Sijaintimitoilla eli *keskiluvuilla* (*measures of central tendency*) kuvataan jakauman sijaintia. Tärkein sijaintimitta on *odotusarvo* (*expected value*)

$$\mathsf{E}(X) = \mu.$$

Odotusarvo on arvo, jonka satunnaismuuttuja saa keskimäärin. Toinen kuvaus on, että odotusarvo on jakauman painopiste. Usein alaindeksillä ilmaistaan, minkä satunnaismuuttujan odotusarvo on kyseessä (esim. μ_X).

Diskreetin satunnaismuuttujan odotusarvo on

$$\mathsf{E}(X) = \sum_{i=1}^{k} x_i \pi_i.$$

Odotusarvo on satunnaismuuttujan arvojen painotettu keskiarvo painoina niiden todennäköisyydet. Jatkuvan satunnaismuuttujan tilanteessa satunnaismuuttujalla on ääretön määrä mahdollisia arvoja. Tällöin odotusarvo lasketaan intuitiivisesti samaan tapaan painoina tiheysfunktion arvot.

Merkintä $\mathsf{E}(\cdot)$ tarkoittaa, että sulkujen sisällä olevan muuttujan odotusarvon laskemiseksi tarpeelliset operaatiot suoritetaan. E:tä kutsutaan odotusarvo-operaattoriksi. Jos odotusarvo-operaattori E kohdistetaan vakioon (c), tulos on vakio:

$$\mathsf{E}(c) = c.$$

Vakion odotusarvo on vakio itse.

Odotusarvo on usein jakauman kiinnostavin piirre. Odotusarvo kuvaa, mitä keskimäärin tapahtuu tai miten asiat keskimäärin ovat. Odotusarvo ei ole aivan osuva nimitys, koska satunnaismuuttuja ei välttämättä voi saada odotusarvon mukaista arvoa.

Esimerkki. Jalkojen lukumäärä.[56] Useimmilla suomalaisilla on kaksi jalkaa. Joillain on yksi jalka ja muutamilla ei yhtään. Jalkojen lukumäärä on diskreetti satunnaismuuttuja, joka voi saada arvot 0, 1 tai 2. Jalkojen lukumäärän odotusarvo on suomalaisten jalkojen lukumäärän keskiarvo. Se on alle kaksi. Kellään ei ole odotusarvon mukaista lukumäärää jalkoja. Useimmilla suomalaisilla on jalkoja enemmän kuin suomalaisilla on jalkoja keskimäärin! □

Jatkuva-arvoisen satunnaismuuttujan X *mediaani* (*median*) M on arvo, joka toteuttaa yhtälöt

$$\mathsf{P}(X < M) = \mathsf{P}(X > M) = \frac{1}{2}.$$

Diskreetin satunnaismuuttujan tilanteessa voidaan muodostaa kuvitteellinen otos, jossa kukin arvo toistuu todennäköisyyttään vastaavalla määrällä. Mediaani voidaan laskea tästä otoksesta jaksossa 8.2 kuvattavalla tavalla.[57] Mediaania pidetään monesti odotusarvoa mielekkäämpänä sijaintimittana, jos tutkittava jakauma on vino (jakso 6.4).

Jakauman *tyyppiarvo* eli *moodi* (*mode*) on satunnaismuuttujan todennäköisin arvo. Joidenkin mielissä tyyppiarvo poimii parhaiten keskiluvun idean. Jakaumalla voi olla monta tyyppiarvoa. Jakaumalla on tällöin ainakin kaksi huippua.

Kellokäyräntapaisten symmetristen jakaumien tilanteessa keskiluvut ovat yksi sama luku. Tyyppiarvo ja mediaani voivat olla keskiarvoa informatiivisempia tunnuslukuja eritoten silloin, kun otosjakauma on hyvin vino. Mediaani ja tyyppiarvo ovat käyttökelpoisia myös, kun aineistot ovat järjestysasteikollisia. Tyyppiarvo on tulkittavissa jopa luokka-aineiston tilanteessa. (Mitta-asteikot nimettiin jaksossa 2.2.)

Vaihtelumitoista eli *hajontaluvuista* (*measures of variation*) yleisimpiä on *varianssi* (*variance*)

$$\mathsf{V}(X) = \sigma^2 = \mathsf{E}(X - \mu)^2.$$

Se mittaa, kuinka suuri satunnaismuuttujan ja sen odotusarvon erotuksen neliö on keskimäärin. *Keskihajonta* (*standard deviation*) on varianssin neliöjuuri:

$$\mathsf{SD}(X) = \sigma = \sqrt{\mathsf{E}(X - \mu)^2}.$$

Keskihajonta on samassa mittayksikössä kuin satunnaismuuttuja. Sen lukuarvon merkitys on siksi helpompi hahmottaa kuin varianssin. Satunnaisvaihtelua mitataan useimmiten varianssilla tai sen neliöjuurella keskihajonnalla.

Varianssi ei ole ehkä yhtä ilmeisen mielenkiintoinen suure kuin odotusarvo. Varianssi on silti keskeinen tilastotieteellinen käsite. Odotusarvoa ei tarvitsisi pohtia, jos satunnaisvaihtelua ei olisi. Tilastollista päättelyä ylipäänsä ei tarvittaisi.

Varianssilla on merkitystä paitsi tilastollisessa analyysissa ja vaativissa asiantuntijatehtävissä myös yhteiskunnassa ylipäänsä ja käytännön elämässä. Useimmiten pientä varianssia pidetään toivottavana, mutta joskus varianssin suuruudella nähdään arvoa. Varianssi voi vaikuttaa yhteiskunnallisiin päätöksiin ja ihmisten käyttäytymiseen.

Esimerkki. Bruttokansantuotteen vakaata kasvua ja pientä varianssia pidetään yleensä tärkeänä. Toimet suhdannevaihteluiden tasaamiseksi voivat olla taloudellisesti mittavia. □

Esimerkki. Valuuttakurssien vaihtelu luo epävarmuutta kansainväliseen kauppaan ja voi vähentää sitä. Valuuttakurssien ennustettavuutta eli pientä varianssia yleensä toivotaan. Yksi syy Suomen liittymiselle euroalueeseen 2002 oli valuuttakurssiriskin poistuminen kaupankäynnistä euroalueella. □

Esimerkki. Varianssi luo liikevoittomahdollisuuksia. Arvopaperi tammikuu 2021 (s:t 20–22):

> Uniperilla on – – isot kaasuvarastot, joita se hyödyntää kaupankäynnissään. Kun kaasu viime talvena halpeni, Uniper osti kaasua päivän hintaan ja toimitti sitä pitkien kiinteähintaisten kaasusopimusten asiakkaille saaden ylimääräistä marginaalia. "Ylituoton voi ottaa aina kun se on mahdollista, vähimmäistuotto tulee sopimushinnan perusteella. Sama toimii kaasun varastoinnissa. Kun viime talvi oli lämmin, Uniper osti kaasua varastot täyteen ja myi sen seuraavaksi talveksi pitkillä sopimuksilla hyvään hintaan", [toimitusjohtaja Markus] Rauramo kertoo. – – Pohjoismaisen sähkömarkkinan hintavaihtelun lisäksi tulokseen vaikuttaa Uniperin onnistuminen kaasu- ja sähkökaupoissa. Kaasun tukkukaupassa hinnan volatiliteetti [varianssi] on Fortumille hyvä asia. *Kaasun hinnalla on vähemmän merkitystä kuin sillä, että se liikkuu.*

(Kursivointi lisätty.) Hintavaihtelun voi nähdä mahdollisuutena ylimääräisille liikevoitoille.[58] □

Esimerkki. Osto-option arvo. Osto-option omaajalla on oikeus ostaa yhtiön osake ennaltamäärättyyn hintaan tiettynä ajankohtana tulevaisuudessa. Jos osake on tuolloin kalliimpi kuin hinta, jolla optiolla voi osakkeen ostaa, optiolla on arvoa. Muulloin optiosta ei ole hyötyä ja option ostaja on maksanut turhasta. Osakkeen hinnan suuri varianssi kasvattaa todennäköisyyttä, että osakkeen hinta nousee suuremmaksi kuin option määrittelemä hinta, jolla option omaaja voi osakkeen ostaa. Mitä suurempaa osakkeen hinnan vaihtelu on, sitä todennäköisempi kallis osake jossain vaiheessa on ja sitä enemmän optiosta kannattaa maksaa. □

Esimerkki. Yhteiskuntaluokalla on taipumus periytyä. Moni pitäisi suotavana, että liikkuvuus yhteiskuntaluokasta toiseen olisi suurempaa eli että siirryttäessä sukupolvesta seuraavaan yhteiskuntaluokan varianssi olisi nykyistä suurempi. □

Esimerkki. Sääennusteen epävarmuus.[59] Meteorologi Joanna Rinne bloggasi 7.6. 2019:

> Ensi viikon ennuste: Lumisade vs. megahelle — kumpi voittaa? – – ennusteen epävarmuus on suurin, jonka muistan nähneeni — realistisia vaihtoehtoa ovat sekä kalseus, jopa lumisade, että tämänhetkisiäkin kuumemmat helteet! – – Ensi viikon lopulla Suomen sääherruudesta kamppailee kaksi ilmamassa-aluetta, tietokone-ennusteet eivät vielä tiedä kumpi näistä pääsee meille. Toisen mukaan meille tulisi kalseaa kesäsäätä ja Lapissa jopa lumisadetta, toisen mukaan tämänhetkisiäkin kuumemmat helteet. – – En muista meteorologin urani aikana

> montaa kertaa nähneeni näin suurta vastakkainasettelua: joko erittäin viileää, jopa kylmää, tai erittäin kuumaa. Vielä ei ole mitään keinoa kertoa, kumpi näistä toteutuu, vai toteutuuko kumpikaan.

Suuri satunnaisvaihtelu lämpötilassa voi hankaloittaa ihmisten ja yritysten suunnitelmia suuresti, ja siitä on tarpeen varoittaa. □

Varianssi voi olla yhteiskuntaa perustavanlaatuisesti kuvaava suure. Varianssilla voi olla yhteiskuntaan syvälle käyviä vaikutuksia.

Esimerkki.[60] Dingin ja Savanin (2020) psykologisissa kokeissa altistus erilaisten suureiden suuremmalle varianssille sai ihmiset vaatimaan kovempia tuomioita epäeettisestä toiminnasta. Vastaavasti ihmiset, jotka kuvasivat maailman epävarmempana, kannattivat useammin kuolemantuomiota. □

Vaihtelumittoja on muitakin kuin tässä esitetyt kaksi (harjoitustehtävä). Luvussa 7 on numeerisia esimerkkejä odotusarvon ja varianssin laskusta.

Jatkuvan satunnaismuuttujan odotusarvo ja varianssi määritellään integraaleina

$$\mathsf{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$

ja

$$\mathsf{V}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

(kun ne ovat olemassa).

6.3 Satunnaismuuttujien lineaarimuunnosten odotusarvo ja varianssi sekä standardointi

Oletetaan, että X_i on satunnaismuuttuja, jonka odotusarvo on μ_i ja että c ja a_i ovat vakioita ($i = 1, \ldots, n$). Tällöin:

$$\begin{aligned} \mathsf{E}(cX_i) &= c\,\mathsf{E}(X_i) = c\mu_i, \\ \mathsf{E}(c + X_i) &= c + \mathsf{E}(X_i) = c + \mu_i, \\ \mathsf{E}(X_i + X_j) &= \mathsf{E}(X_i) + \mathsf{E}(X_j) = \mu_i + \mu_j. \end{aligned}$$

Tulokset ponnahtavat odotusarvon määritelmästä. Yhtälöistä seuraa, että satunnaismuuttujien lineaarimuunnoksen $c + \sum_{i=1}^{n} a_i X_i$ odotusarvo on

$$\mathsf{E}\left(c + \sum_{i=1}^{n} a_i X_i\right) = c + \sum_{i=1}^{n} a_i \mu_i$$

(harjoitustehtävä). Tulos ei edellytä satunnaismuuttujien X_i riippumattomuutta.

Olkoot X_i:t riippumattomia satunnaismuuttujia, joiden varianssi on σ_i^2 ($i = 1, \ldots, n$). Tällöin:

$$\begin{aligned} \mathsf{V}(cX_i) &= c^2\,\mathsf{V}(X_i) = c^2\sigma_i^2, \\ \mathsf{V}(c + X_i) &= \mathsf{V}(X_i) = \sigma_i^2, \\ \mathsf{V}(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n} \mathsf{V}(X_i) = \sum_{i=1}^{n} \sigma_i^2. \end{aligned}$$

Ominaisuudet seuraavat varianssin määritelmästä. Yhtälöistä saadaan riippumattomien satunnaismuuttujien lineaarimuunnoksen varianssiksi

$$\mathsf{V}\left(c + \sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

(harjoitustehtävä).

Huom! Riippumattomien satunnaismuuttujien lineaarimuunnoksen keskihajonta ei noudata yllä olevan kaltaista yksinkertaista kaavaa.

Kahden ei-riippumattoman satunnaismuuttujan summan varianssi on

$$\mathsf{V}(X_1 + X_2) = \mathsf{V}(X_1) + \mathsf{V}(X_2) + 2\mathsf{C}(X_1, X_2).$$

$\mathsf{C}(X_1, X_2)$ määritellään jaksossa 6.5.

Tärkeä lineaarimuunnos on standardointi. Olkoot satunnaismuuttujan X odotusarvo μ ja varianssi σ^2. Standardoitu satunnaismuuttuja on

$$\frac{X - \mu}{\sigma}.$$

Sen odotusarvo on 0 ja varianssi 1:

$$\mathsf{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\mathsf{E}(X-\mu) = \frac{1}{\sigma}[\mathsf{E}(X)-\mu] = \frac{1}{\sigma}(\mu-\mu) = 0$$

ja

$$\mathsf{V}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathsf{V}(X-\mu) = \frac{1}{\sigma^2}\mathsf{V}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

Lukematon määrä tilastollisia menetelmiä ja testejä perustuu standardoinnille.

6.4 Vinous

Odotusarvo

$$\mathsf{E}\left(\frac{X-\mu}{\sigma}\right)^3 = \frac{\mathsf{E}(X-\mu)^3}{\sigma^3}$$

mittaa jakauman *vinoutta* (*skewness*). Jos satunnaismuuttujan arvot jakautuvat peilikuvamaisesti odotusarvon ympärille, jakauma on symmetrinen ja vinous on 0. Jos satunnaismuuttuja saa odotusarvoaan paljon suurempia arvoja ($x-\mu > 0$) todennäköisemmin kuin vastaavia pienempiä arvoja ($\mu - x$), odotusarvo yllä tapaa olla positiivinen. Jakauma on tällöin vino oikealle. Jos odotusarvo yllä on negatiivinen, jakauma on vino vasemmalle. Vinous voi olla 0, vaikka jakauma ei olisi symmetrinen odotusarvon ympärillä. Keskihajonta σ nimittäjässä poistaa satunnaismuuttujan mittayksikön vaikutuksen.

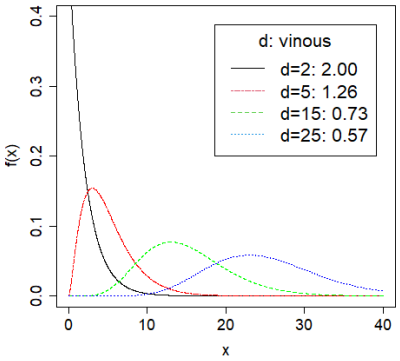
Esimerkki. χ^2- ja F-jakaumien vinous. Kuviin 6.2 ja 6.3 on piirretty oikealle vinoja $\chi^2(d)$- ($d = 2, 5, 15, 25$) ja $\mathsf{F}(d, 100)$-jakaumia ($d = 5, 10, 15, 50$) ja kirjattu kunkin vinous. Jakaumien symmetrisoituessa vinous pienenee. Jakaumat kuvataan tarkemmin jaksoissa 7.2.2 ja 7.2.4. □

6.5 Satunnaismuuttujien korrelaatio

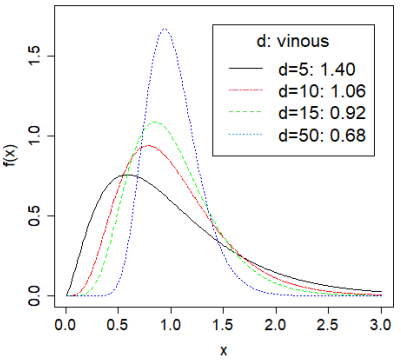
Useimmiten tutkitaan monia muuttujia. Niiden välisten suhteiden mittaaminen ja kuvaaminen on oleellinen osa tilastotiedettä.

Tunnetuimpia tilastotieteellisiä käsitteitä on *korrelaatio* (*correlation*). Kahden satunnaismuuttujan X_1 ja X_2 välisen lineaarisen yhteyden vahvuutta mitataan usein *korrelaatiokertoimella* (*correlation coefficient*), tai lyhyemmin korrelaatiolla, ρ:

$$\rho = \frac{\mathsf{C}(X_1, X_2)}{\sqrt{\mathsf{V}(X_1)\mathsf{V}(X_2)}}.$$



Kuva 6.2: $\chi^2(d)$-jakauman vinous ($d = 2, 5, 15, 25$).



Kuva 6.3: $\mathsf{F}(d, 100)$-jakauman vinous ($d = 5, 10, 15, 50$).

Tässä

$$\mathsf{C}(X_1, X_2) = \mathsf{E}(X_1 - \mu_1)(X_2 - \mu_2) = \mathsf{E}(X_1 X_2) - \mu_1 \mu_2.$$

Suuretta $\mathsf{C}(X_1, X_2)$ kutsutaan muuttujien X_1 ja X_2 *kovarianssiksi* (*covariance*). Alaindekseillä osoitetaan joskus, minkä muuttujien korrelaatiosta on kyse (esim. ρ_{X_1,X_2}).

Kovarianssi poimii X_1:n ja X_2:n lineaarisen yhteyden etumerkin ja suuruuden. Keskihajonnat $[\mathsf{V}(X_1)]^{1/2}$ ja $[\mathsf{V}(X_2)]^{1/2}$ skaalavat kovarianssin niin, että korrelaatiokerroin saa arvoja välillä $[-1, 1]$. Arvo 0 vastaa tilannetta, että lineaarista riippuvuutta ei ole. Positiivisen korrelaation tilanteessa satunnaismuuttujat X_1 ja X_2 tapaavat vaihdella samansuuntaisesti; negatiivisen korrelaation tilanteessa vastakkaissuuntaisesti. Ääritapauksissa $\rho = 1$ tai $\rho = -1$ lineaarinen riippuvuus on täydellistä. Muuttujia sitoo tällöin ei-satunnainen lineaarinen yhtälö ($X_2 = a + bX_1$, jossa a ja $b \neq 0$ ovat vakiota).

Satunnaismuuttujat voivat olla riippuvia, vaikka ne eivät korreloisi eli niiden korrelaatio olisi 0 (jakso 8.2). Korrelaatio on vain lineaarisen riippuvuuden mitta. Kuva 7.14 jaksossa 7.5 havainnollistaa korrelaatiota, ja kuva 8.4 jaksossa 8.2 selventää, millaista riippuvuutta korrelaatio mittaa ja millaista ei.

Luku 7

Todennäköisyysjakaumia

> Ne, jotka rakastuvat käytäntöön vailla teoreettista tietoutta, ovat kuin merimies, joka astuu laivaan vailla peräsintä tai kompassia eikä voi koskaan olla varma, mihin laiva on menossa. Käytännön tulee aina perustua teoriaan.[61]
>
> Leonardo da Vinci (1452–1519)

Todennäköisyysjakaumien avulla voidaan laskea todennäköisyys mitä erilaisimmille tapahtumille. Oleellista on taito valita sopiva todennäköisyysjakauma kuhunkin tilanteeseen. Todennäköisyysjakauma on tilastollisen päättelyn tärkein työkalu. Luvussa kuvataan yleisimpiä todennäköisyysjakaumia diskreeteille ja jatkuva-arvoisille satunnaismuuttujille sekä niiden todennäköisyysjakaumien välisiä yhteyksiä.

Havaintojen sanotaan joskus noudattavan tiettyä jakaumaa, vaikka havainnot ovat satunnaismuuttujien toteumia eivätkä satunnaismuuttujia. Tällaisia ilmaisuja käytetään paikoin myös tässä luentomateriaalissa.

7.1 Diskreettejä jakaumia

7.1.1 Bernoulli-jakauma

Bernoulli-kokeessa on kaksi tulosvaihtoehtoa: A tapahtuu tai ei. *Bernoulli-satunnaismuuttuja* (X) saadaan, kun tulosvaihtoehtoihin liitetään luvut 1 (tapahtuu) ja 0 (ei tapahdu). Bernoulli-satunnaismuuttujan jakauman määrittelee parametri π, joka on todennäköisyys 1:lle eli tapahtumiselle. Jos X noudattaa Bernoulli-jakaumaa parametrilla π, merkitään $X \sim \mathsf{B}(\pi)$.

Bernoulli-jakautuneen satunnaismuuttujan odotusarvo ja varianssi ovat π ja $\pi(1-\pi)$:

$$\mathsf{E}(X) = \pi \times 1 + (1-\pi) \times 0 = \pi \tag{7.1}$$

ja

$$\begin{aligned} \mathsf{V}(X) = \mathsf{E}(X-\mu)^2 = \mathsf{E}(X-\pi)^2 &= \pi \times (1-\pi)^2 + (1-\pi) \times (0-\pi)^2 \\ &= \pi(1-\pi)^2 + \pi^2(1-\pi) = \pi(1-\pi)(1-\pi+\pi) = \pi(1-\pi). \end{aligned} \tag{7.2}$$

Yllä on varianssin määritelmässä $\mathsf{E}(X-\mu)^2$ merkitty odotusarvoa μ:llä. Bernoulli-jakautunut satunnaismuuttuja on yksinkertaisin mahdollinen satunnaismuuttuja. Tapahtumaa tai tapahtumattomuutta kuvataan monesti ilmaisuilla "onnistuminen" tai "epäonnistuminen", "voitto" tai "häviö", "sairastuminen" tai "pysyminen terveenä" jne.

Esimerkki. Lantin heitto (jatkoa). Kuvataan "kruuna" ykköseksi ja "klaava" nollaksi. Lantin heitto -satunnaismuuttuja X on Bernoulli-jakautunut parametrilla $\pi = 0.5$. Satunnaismuuttujan X odotusarvo ja varianssi ovat 0.5 ja $0.5^2 = 0.25$.

Tapahtumaa, joka vastaisi 0.5:ttä, ei ole. Odotusarvoa vastaava tapahtuma ei ole välttämättä "yleinen" tai edes mahdollinen. □

Bernoulli-satunnaismuuttujan pistetodennäköisyysfunktio on

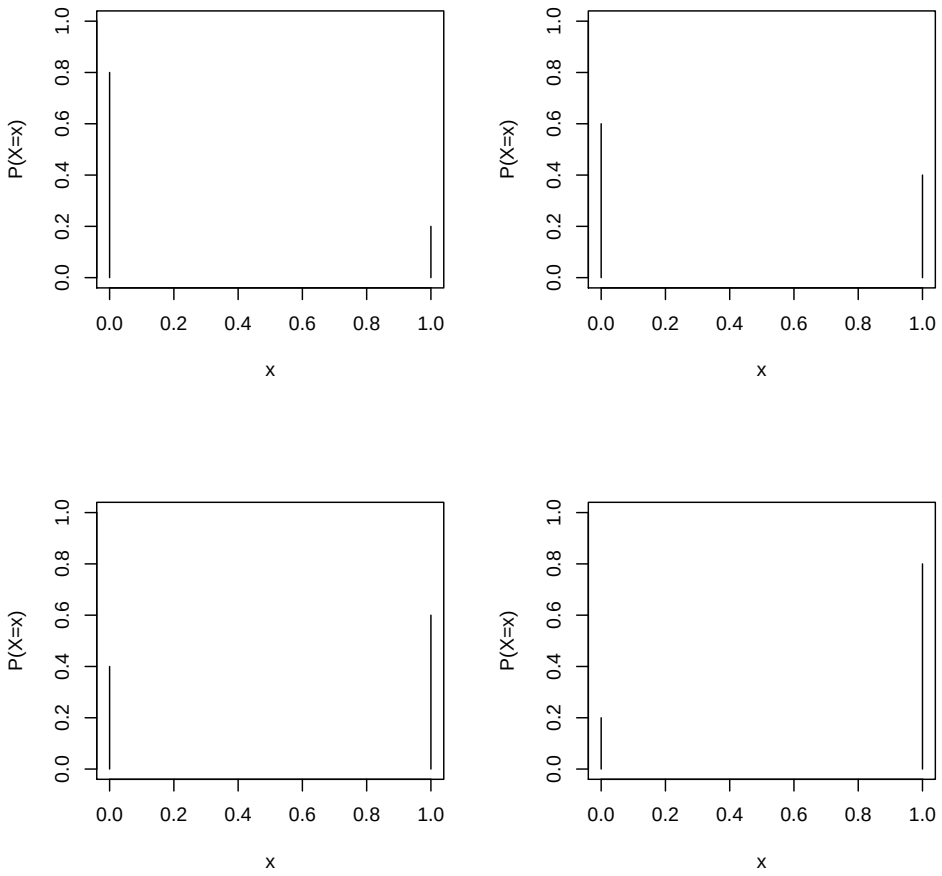
$$\mathsf{P}(X=x) = \pi^x(1-\pi)^{1-x} = \begin{cases} \pi, & \text{jos } x = 1, \\ 1-\pi, & \text{jos } x = 0. \end{cases}$$

Ensimmäisen yhtäsuuruusmerkin jälkeinen muoto on avuksi binomijakaumaa johdettaessa (jakso 7.1.3). Kuva 7.1 havainnollistaa Bernoulli-jakaumaa π:n arvoilla 0.2, 0.4, 0.6 ja 0.8.

7.1.2 Diskreetti tasainen jakauma

Satunnaismuuttuja X noudattaa *diskreettiä tasaista jakaumaa*, jos sen pistetodennäköisyysfunktio on

$$\mathsf{P}(X=x) = \begin{cases} \dfrac{1}{k}, & \text{jos } x = x_1, \ldots, x_k, \\ 0, & \text{muutoin.} \end{cases}$$



Kuva 7.1: Bernoulli-jakautuneen satunnaismuuttujan pistetodennäköisyysfunktioita π:n arvoilla 0.2, 0.4, 0.6 ja 0.8.

Mikäli luvut $x_1, \ldots, x_k$ ovat peräkkäisiä kokonaislukuja, kätevät esitysmuodot odotusarvolle ja varianssille ovat

$$\mathsf{E}(X) = (x_1 + x_k)/2$$

ja

$$\mathsf{V}(X) = [(x_k - x_1 + 1)^2 - 1]/12$$

(esim. Tijms 2012, 313).

Lukujen $x_1, \ldots, x_k$ ollessa peräkkäisiä kokonaislukuja odotusarvo on

$$\mathsf{E}(X) = \sum_{i=0}^{k-1} \frac{x_1 + i}{k} = x_1 + \frac{1}{k} \sum_{i=0}^{k-1} i = x_1 + \frac{1}{k} \frac{k(k-1)}{2} = x_1 + \frac{k-1}{2} = \frac{x_1 + x_1 + k - 1}{2}$$
$$= \frac{x_1 + x_k}{2}.$$

Yllä on hyödynnetty kaavaa $\sum_{i=1}^{k} i = k(k+1)/2$. Sen todistus: Merkitään $S = \sum_{i=1}^{k} i$. Selvästikin pätee

$$S = 1 + 2 + \cdots + (k-1) + k \quad \text{ja} \quad S = k + (k-1) + \cdots + 2 + 1.$$

Lasketaan yhtälöt puolittain yhteen. Saadaan

$$2S = (k+1) + (k+1) + \cdots + (k+1) + (k+1) = k(k+1).$$

Ratkaistaan S:

$$S = \frac{k(k+1)}{2}.$$

Esimerkki. Nopan heitto (jatkoa). Nopan silmäluku (X) on diskreetisti tasaisesti jakautunut satunnaismuuttuja. Kunkin silmäluvun todennäköisyys on $1/6$. Silmäluvun odotusarvo ja varianssi ovat

$$\mathsf{E}(X) = \sum_{i=1}^{6} i \times \frac{1}{6} = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5 \quad \text{ja}$$

$$\mathsf{V}(X) = \sum_{i=1}^{6} (i - 3.5)^2 \times \frac{1}{6} = (1 - 3.5)^2 \times \frac{1}{6} + \cdots + (6 - 3.5)^2 \times \frac{1}{6} = \frac{35}{12} \approx 2.92.$$

Ne voi laskea kätevämmin näin: $(1+6)/2 = 3.5$ ja $[(6-1+1)^2 - 1]/12 = 35/12$. □

7.1.3 Binomijakauma

On tehty tai havaittu n riippumatonta samanlaista Bernoulli-koetta (kussakin tapahtumatodennäköisyys on π). Tällaista yhdistettyä koetta kutsutaan *binomikokeeksi* (*binomial experiment*). Merkitään tapahtumien lukumäärää binomikokeessa y:llä ($0 \leq y \leq n$). Sen pistetodennäköisyys on

$$\mathsf{P}(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}. \tag{7.3}$$

Jakaumaa kutsutaan *binomijakaumaksi* (otoskoolla n ja parametrilla π) ja sitä merkitään $\mathsf{Bin}(n, \pi)$.

Selitys pistetodennäköisyydelle (7.3): Kaikkien n:n pituisten tapahtumajonojen, joissa on y tapahtumaa, todennäköisyys on $\pi^y(1-\pi)^{n-y}$. Esimerkiksi jos kolme ensimmäistä koetta onnistuvat, seuraava epäonnistuu ja kaksi viimeistä koetta onnistuu ja epäonnistuu ja onnistumisia on yhteensä y, havaitun tapahtumajonon todennäköisyys on

$$\pi\pi\pi(1-\pi) \times \cdots \times \pi(1-\pi) = \pi^y(1-\pi)^{n-y}.$$

Muoto oikealla saadaan kokoamalla tulon termit. Järjestyksestä riippumatta muoto oikealla pätee, jos onnistumisia on y kappaletta. (Vrt. uhkapelurin virhepäätelmä jaksossa 4.5.) Vaihtoehtoisia järjestyksiä y:lle onnistumiselle ja $n-y$:lle epäonnistumiselle on binomikertoimen $\binom{n}{y}$ mukainen määrä. Kukin järjestys on erillinen. Todennäköisyys $\mathsf{P}(Y = y)$ saadaan erillisyyden perusteella laskemalla kaikkien mahdollisten jonojen, joissa on y onnistumista, todennäköisyydet yhteen:

$$\mathsf{P}(Y = y) = \pi^y(1-\pi)^{n-y} + \cdots + \pi^y(1-\pi)^{n-y} = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

Merkitään binomijakauman taustalla olevia yksittäisiä Bernoulli-satunnaismuuttujia X_i:llä (saa arvon 1 tai 0). Tällöin $Y = \sum_{i=1}^{n} X_i$. Nyt voidaan päätellä binomijakautuneen satunnaismuuttujan odotusarvo ja varianssi $n\pi$ ja $n\pi(1-\pi)$:

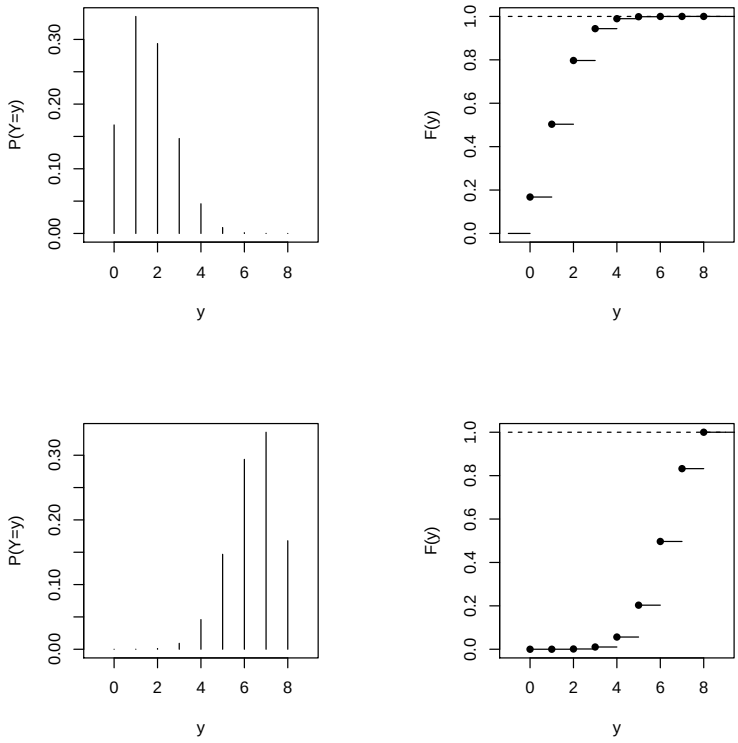
$$\mathsf{E}(Y) = \mathsf{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathsf{E}(X_i) = \sum_{i=1}^{n} \pi = n\pi \tag{7.4}$$

ja

$$\mathsf{V}(Y) = \mathsf{V}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathsf{V}(X_i) = \sum_{i=1}^{n} \pi(1-\pi) = n\pi(1-\pi). \tag{7.5}$$

Ylle on sijoitettu Bernoulli-satunnaismuuttujan odotusarvo π ja varianssi $\pi(1-\pi)$ kaavoista (7.1) ja (7.2). Varianssin laskussa on hyödynnetty oletusta satunnaismuuttujien X_i riippumattomuudesta.

Binomijakauma on epäsymmetrinen, paitsi jos $\pi = 0.5$. Kuva 7.2 havainnollistaa binomijakauman pistetodennäköisyys- ($\mathsf{P}(Y = y)$) ja kertymäfunktiota ($F(y)$) π:n arvoilla 0.2 ja 0.8, kun $n = 8$.[62] Pistetodennäköisyysfunktiot ovat toistensa peilikuvia, koska ne määrittävät parametriarvot sijaitsevat yhtä kaukana 0.5:stä.



Kuva 7.2: Binomijakautuneen satunnaismuuttujan pistetodennäköisyys- ja kertymäfunktiot π:n arvoilla 0.2 ja 0.8, kun $n = 8$.

Havaintojen poimiminen ei saa vaikuttaa tapahtuman todennäköisyyteen myöhemmin, jotta satunnaismuuttuja voisi olla (likimain) binomijakautunut. Tällaisia ovat tilanteet, joissa kukin poimittu alkio palautetaan takaisin perusjoukkoon tai perusjoukko on äärettömän (tai hyvin) suuri, jolloin alkion poiminta ei vaikuta mitään (tai vaikuttaa mitättömästi) seuraavan Bernoulli-kokeen tuloksen todennäköisyyteen.

Binomijakauma on sovellusten kannalta tärkeimpiä jakaumia. Yksi syy on, että binomijakauma kuvaa luontevasti monia ilmiöitä. Toinen syy on, että monesti kiinnostuksen kohde on osuus Y/n. Sen jakauma on skaalausta vaille sama kuin Y:n, kun n on kiinteä. Y on vain jaettu vakiolla osuuden laskemiseksi. Myös jakauman yksinkertaisuus on ilmeinen syy sen suosiolle. Siksi sillä approksimoidaan monia tilanteita, jotka ovat lähes mutteivät täsmälleen kuvattavissa binomijakaumalla.

Olkoon satunnaismuuttuja Y binomijakautunut: $Y \sim \mathsf{Bin}(n, \pi)$. R laskee binomitodennäköisyyden tapahtumille $Y = y$ ja $Y \leq y$ käskyillä

`dbinom(y,n,`π`)` ja `pbinom(y,n,`π`)`,

kun niihin sijoittaa sopivat arvot.

Esimerkki. Kortin peluu (jatkoa). Vedetään hyvinsekoitetusta korttipakasta sattumanvaraisesti kortti, katsotaan se, palautetaan kortti pakkaan ja sekoitetaan pakka. Toistetaan tämä 5 kertaa. Lasketaan todennäköisyydet, että kuninkaita tulee 0, 1, 2, 3, 4 tai 5 kappaletta sekä todennäköisyys, että ainakin 1 nostetuista korteista on ollut kuningas.

Kukin kortin nosto on Bernoulli-koe, jossa tapahtuman "kuningas" todennäköisyys on $4/52 = 1/13$. Lasketaan kysytyt binomijakauman pistetodennäköisyydet (Y on "kuninkaiden lukumäärä"):

$$\mathsf{P}(Y=0) = \binom{5}{0}\left(\frac{1}{13}\right)^0\left(1-\frac{1}{13}\right)^{5-0} = \left(\frac{12}{13}\right)^5 = 0.6701769,$$

$$\mathsf{P}(Y=1) = \binom{5}{1}\left(\frac{1}{13}\right)^1\left(1-\frac{1}{13}\right)^{5-1} = 5\times\frac{1}{13}\times\left(\frac{12}{13}\right)^4 = 0.2792404,$$

$$\mathsf{P}(Y=2) = \binom{5}{2}\left(\frac{1}{13}\right)^2\left(1-\frac{1}{13}\right)^{5-2} = 10\times\left(\frac{1}{13}\right)^2\times\left(\frac{12}{13}\right)^3 = 0.04654006,$$

$$\mathsf{P}(Y=3) = \binom{5}{3}\left(\frac{1}{13}\right)^3\left(1-\frac{1}{13}\right)^{5-3} = 10\times\left(\frac{1}{13}\right)^3\times\left(\frac{12}{13}\right)^2 = 0.003878339,$$

$$\mathsf{P}(Y=4)=\binom{5}{4}\left(\frac{1}{13}\right)^4\left(1-\frac{1}{13}\right)^{5-4}=5\times\left(\frac{1}{13}\right)^4\times\frac{12}{13}=0.0001615974 \quad \text{ja}$$
$$\mathsf{P}(Y=5)=\binom{5}{5}\left(\frac{1}{13}\right)^5\left(1-\frac{1}{13}\right)^{5-5}=\left(\frac{1}{13}\right)^5=0.000002693291.$$

Todennäköisyydet on saatu sijoittamalla $\pi = 1/13$ ja $n = 5$ kaavaan (7.3) ja laskemalla ne R-komennoilla `dbinom(0,5,1/13)`, `dbinom(1,5,1/13)` jne.

Todennäköisyys saada ainakin 1 kuningas on todennäköisyys saada 1 tai enemmän kuninkaita. Erillisten tapahtumien yhteenlaskusäännön (4.5) perusteella todennäköisyydet näille tapahtumille voidaan laskea yhteen:

$$0.279 + 0.047 + 0.004 + 0.000 + 0.000 = 0.330.$$

Tulos 0.330 saataisiin helpommin laskemalla se vastatapahtuman todennäköisyyden ($\mathsf{P}(Y = 0)$) avulla:

$$1 - \mathsf{P}(Y = 0) = 1 - 0.6701769 \approx 0.330.$$

Todetaan lisäksi, että odotusarvo ja varianssi kuninkaiden lukumäärälle ovat noin 0.385 ja 0.355 (kaavat (7.4) ja (7.5)):

$$\mathsf{E}(Y)=5\times\frac{1}{13}=\frac{5}{13}\approx 0.385 \quad \text{ja}$$
$$\mathsf{V}(Y)=5\times\frac{1}{13}\times\frac{12}{13}=\frac{60}{169}\approx 0.355.$$

□

Esimerkki. Asumisriidat lapsista käräjäoikeuksissa. Oikeuspoliittinen tutkimuslaitos (nykyinen Kriminologian ja oikeuspolitiikan instituutti) tutki käräjäoikeuksien päätöksiä lapsen asumisesta ajalla 14.11.2005–13.2.2006 (529 havaintoa). Tutkitaan päätöksiä, joissa lapset määrättiin asumaan vain jommankumman vanhemman luona. Niissä lapsi osoitettiin asumaan 35:ssä isän ja 83:ssa äidin luona (118 havaintoa).[63] Vastaavat prosenttiosuudet ovat noin 29.7 ja 70.3. Oletetaan, että käräjäoikeudet ylipäänsä määräisivät lapset asumaan eri sukupuolta olevien vanhempien luona yhtä todennäköisesti eli prosenttiosuuksilla 50 ja 50 ja että päätökset olisivat riippumattomia. Mikä olisi odotusarvo lapsille, jotka oikeus osoittaa asumaan isän luona? Entä äidin luona? Mikä olisi todennäköisyys, että 118:n suuruisessa satunnaisessa aineistossa 35:ssä tai vähäisemmässä lukumäärässä lapset määrätään asumaan isän luona?

Kukin päätös voidaan mieltää Bernoulli-kokeena ja päätösten lukumäärä ajatella binomijakautuneeksi. Päätökset ajatellaan poimituiksi äärettömästä määrästä potentiaalisia päätöksiä, joita käräjäoikeudet olisivat voineet tehdä.

Binomijakauman mukaan odotusarvo sekä isän että äidin luokse osoitettavien lasten lukumäärälle on $118 \times 0.5 = 59$ (kaava (7.4)). Kysytty todennäköisyys saadaan laskemalla $\mathsf{Bin}(118, 0.5)$-jakauman kertymäfunktion arvo pisteessä 35. R-käsky

```
pbinom(35,118,0.5)
```

antaa täksi todennäköisyydeksi noin 0.000006. Jos $\pi = 0.5$, niin havaittu lukumäärä 35 on poikkeuksellisen pieni. □

7.1.4 Multinomijakauma

On tehty tai havaittu n riippumatonta koetta. Kussakin kokeessa on c tulosvaihtoehtoa tai luokkaa, jonka todennäköisyys on π_i ($\sum_{i=1}^{c} \pi_i = 1$). Satunnaismuuttuja koostuu frekvensseistä eli lukumääristä N_i kussakin luokassa eli on moniulotteinen ($\sum_{i=1}^{c} N_i = n$). Havaitun tapahtumajonon todennäköisyys on

$$\pi_1^{n_1} \pi_2^{n_2} \dots \pi_c^{n_c},$$

jossa n_i:t ovat lukumääriä, joilla i. tulosvaihtoehto on toteutunut. Todennäköisyys on sama riippumatta järjestyksestä, jossa tulosvaihtoehdot ovat toteutuneet, jos n_i:t ovat samat. Tapahtumajonoja on yhteensä

$$\frac{n!}{n_1! n_2! \dots n_c!}$$

(kaava (5.6)). Ne ovat erillisiä, joten todennäköisyys n_1:lle havainnolle luokassa 1, n_2:lle havainnolle luokassa 2 jne. on

$$\mathsf{P}(N_1 = n_1, N_2 = n_2, \dots, N_c = n_c) = \frac{n!}{n_1! n_2! \dots n_c!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_c^{n_c} \qquad (7.6)$$

($\sum_{i=1}^{c} N_i = n$). Jakaumaa kutsutaan *multinomijakaumaksi* (otoskoolla n ja parametreilla $\pi_1, \dots, \pi_c$), ja sitä merkitään $\mathsf{Mul}(n, \pi_1, \dots, \pi_c)$. Jakauma määräytyy $(c-1)$:stä π_i-parametrista; viimeinen voidaan ratkaista rajoituksesta $\sum_{i=1}^{c} \pi_i = 1$.

Luokan i lukumäärän odotusarvo ja varianssi ovat

$$\mathsf{E}(N_i) = n\pi_i$$

ja

$$\mathsf{V}(N_i) = n\pi_i(1 - \pi_i).$$

Ne seuraavat binomijakaumasta: Luokan i lukumäärä on binomijakautunut karkeistamalla multinomijakauman luokkajako kahdeksi: i. luokaksi ja muiksi luokiksi.

Pearsonin χ^2-testi (jakso 12.2) on käytetyimpiä tilastollisia testejä. Sitä käytettäessä aineisto voidaan monesti ajatella syntyneeksi multinomijakaumasta. Se on relevantti lukuisissa muissakin yhteyksissä kuten kyselytutkimuksissa.

Esimerkki. Puoluekannatus. Tehdään otantatutkimus äänestysikäisten suomalaisten puoluekannoista. On päätetty selvittää 1 000:lta riippumattomasti ja sattumanvaraisesti poimitulta äänestysikäiseltä (kullakin yhtäsuuri todennäköisyys tulla poimituksi otokseen) kannattavatko he vasemmistoa, keskustaa vai oikeistoa (V, K ja O). Kukin haastattelu on koe, jossa on kolme tulosvaihtoehtoa. Kunkin haastattelun jälkeen haastateltu palautetaan perusjoukkoon ja hänet saatetaan haastatella uudestaan, jos hän tulee poimituksi myöhemmin.[64]

Otantatutkimus tuottaa kolmesta luokasta koostuvan taulukon alla. Haastattelujen järjestyksestä riippumatta kaikki tulosvaihtoehtojonot, joissa olisi yhtä monta havaintoa kutakin puoluekantaa, olisivat yhtä todennäköisiä.

V	K	O
320	350	330

Oletetaan, että kunkin luokan todennäköisyys on 1/3 eli että vasemmistolla, keskustalla ja oikeistolla on yhtä suuri kannatus. (Poliittinen kanta noudattaa diskreettiä tasaista jakaumaa.) Havaittujen kannatuslukumäärien todennäköisyys on tällöin kaavan (7.6) mukaan noin 0.0004. Se on laskettu R:n komennoilla

```
x <- c(320, 350, 330)
prob <- c(1/3, 1/3, 1/3)
dmultinom(x,size=1000,prob)
```

Tällaisen yksittäisen luokkafrekvenssikombinaation todennäköisyys on tyypillisesti hyvin pieni. Mielenkiintoisempi on esimerkiksi kysymys, mikä on todennäköisyys havaita vasemmistolle 32 %:n tai sitä pienempi kannatus, jos todellisuudessa vasemmiston kannatus on 33.33 %. Tähän kysymykseen vastaus voidaan laskea binomijakauman kertymäfunktiosta ajatellen kyselyä binomikokeena, jossa on kaksi vaihtoehtoa “vasemmisto” ja “muut” vasemmiston todennäköisyyden ollessa 1/3. Kysytty todennäköisyys on noin 0.19 (laskettu R:n käskyllä `pbinom(320,1000,1/3)`). Ei ole poikkeuksellista havaita 1 000 henkilöä kattavassa otantatutkimuksessa 32 %:n kannatus puolueelle, jonka kannatus on todellisuudessa 33.33 %. □

7.1.5 Hypergeometrinen jakauma

Hypergeometrinen jakauma on erityisen relevantti tilanteissa, joissa perusjoukon koko on pienehkö ja havainnot poimitaan perusjoukosta palauttamatta niitä siihen poiminnan jälkeen. Kukin poiminta on Bernoulli-koe muttei riippumaton sellainen. Jakaumalla on myös muuta käyttöä. Jakauma tavataan perustella pallojen poimimis -analogialla.

Pussissa on l liilaa ja m mustaa palloa. Poimitaan pussista yksitellen sattumanvaraisesti n palloa palauttamatta niitä pussiin. Hypergeometrisen jakauman pistetodennäköisyydet

$$\mathsf{P}(Y = y) = \frac{\binom{l}{y}\binom{m}{n-y}}{\binom{l+m}{n}} \tag{7.7}$$

ovat todennäköisyyksiä saada y liilaa palloa. Yllä $0 \leq y \leq l$ ja $0 \leq n - y \leq m$; muulloin $\mathsf{P}(Y = y) = 0$.

Jakauman (7.7) johto: Kaikki poimitut pallokombinaatiot ovat yhtä todennäköisiä. Erilaisia n pallon kombinaatioita on $\binom{l+m}{n}$. Liiloista palloista voidaan poimia y palloa $\binom{l}{y}$ tavalla. Mustista palloista saa $n - y$ pallon erilaisia kombinaatiota $\binom{m}{n-y}$. Kombinaatiot liiloista ja mustista palloista voidaan muodostaa toisistaan riippumattomasti, joten kertolaskuperiaatteesta seuraa, että erilaisia tapoja muodostaa y liilan ja $n - y$ mustan pallon kombinaatiota on $\binom{l}{y}\binom{m}{n-y}$. Suhteuttamalla se klassisen todennäköisyyden määritelmän mukaisesti kaikkien mahdollisten kombinaatioiden lukumäärään saadaan y liilan pallon poimimistodennäköisyydeksi kaava (7.7).

Merkintä $Y \sim \mathsf{HG}(l, m, n)$ tarkoittaa, että satunnaismuuttujan Y pistetodennäköisyydet ovat kaavan (7.7) määräämät. Hypergeometrisesti jakautuneen satunnaismuuttujan odotusarvo ja varianssi ovat

$$\mathsf{E}(Y) = n \times \frac{l}{l+m} = n\pi \tag{7.8}$$

ja

$$\mathsf{V}(Y) = n \times \frac{l}{l+m} \times \frac{m}{l+m} \times \frac{l+m-n}{l+m-1} = n\pi(1-\pi) \times \frac{l+m-n}{l+m-1}. \tag{7.9}$$

(esim. Lindgren 1976, 169–170). Yllä π on todennäköisyys saada liila pallo ensimmäisellä nostolla.

Osoitetaan, että todennäköisyys saada liila pallo myös i:llä nostolla on π $(i = 1, \ldots, n)$ ja että $\mathsf{E}(Y) = n\pi$. Määritellään satunnaismuuttuja X_i:

$$X_i = \begin{cases} 1, & \text{jos pallo on liila } i.\text{ nostolla,} \\ 0, & \text{jos pallo on musta } i.\text{ nostolla.} \end{cases}$$

Pallot (n kappaletta) voidaan poimia

$$(l+m)(l+m-1)\cdots[l+m-(n-1)] = (l+m)(l+m-1)\cdots(l+m-n+1)$$

järjestyksessä (ajatellen palloja yksilöinä; kaava (5.2)). Kaikki permutaatiot ovat yhtä todennäköisiä. Permutaatioita, joissa $i.$ pallo on liila, on

$$l(l+m-1)\cdots[l+m-(n-1)] = l(l+m-1)\cdots(l+m-n+1).$$

Yllä $i.$ liilaksi palloksi on l vaihtoehtoa. Loput $n-1$ palloa voidaan järjestää $(l+m-1)\cdots(l+m-n+1)$ tavalla. Todennäköisyys, että $i.$ pallo on liila, on permutaatioiden lukumäärä, joissa ehto toteutuu suhteessa kaikkien permutaatioiden lukumäärään:

$$\mathsf{P}(X_i = 1) = \frac{l(l+m-1)\cdots(l+m-n+1)}{(l+m)(l+m-1)\cdots(l+m-n+1)} = \frac{l}{l+m} = \pi.$$

Liilojen pallojen lukumäärä n:ssä nostossa on $Y = \sum_{i=1}^{n} X_i$. Sen odotusarvo on

$$\mathsf{E}(Y) = \mathsf{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathsf{E}(X_i) = \sum_{i=1}^{n} \pi = n\pi.$$

Satunnaismuuttujat X_i eivät ole riippumattomia, mutta se ei vaikuta odotusarvo-operaattorin käyttöön.

Vastaava summan varianssin laskusääntö edellyttää summattavien satunnaismuuttujien riippumattomuuden. Liilojen pallojen lukumäärän varianssin lasku on monimutkaisempi ja sivuutetaan.

Esimerkki.[65] Liilojen ja mustien pallojen poimiminen. Pussissa on sekaisin 3 liilaa ja 7 mustaa palloa. Poimitaan 3 palloa pussista palauttamatta niitä pussiin nostojen jälkeen. Määritellään satunnaismuuttuja X_i, joka saa arvon 1 tai 0, jos $i.$ pussista nostettu pallo on liila tai musta.

Todennäköisyys, että 1. pussista nostettu pallo on liila, on 3/10.

$$\mathsf{P}(X_1) = \frac{3}{3+7} = \frac{3}{10}.$$

Osoitetaan, että todennäköisyys 3. pussista nostetun pallon liiluudelle on myös 3/10. Liila pallo voi tulla kolmantena $2 \times 2 \times 1 = 4$:llä eri tavalla (1. ja 2. pallolle 2 vaihtoehtoa ja 3. pallolle 1). Kokonaistodennäköisyyden lakia (4.14) soveltamalla saadaan todettu tulos:

$$\mathsf{P}(X_3) = \mathsf{P}(\text{“001”}) + \mathsf{P}(\text{“011”}) + \mathsf{P}(\text{“101”}) + \mathsf{P}(\text{“111”})$$

$$= \frac{7}{10} \times \frac{6}{9} \times \frac{3}{8} + \frac{7}{10} \times \frac{3}{9} \times \frac{2}{8} + \frac{3}{10} \times \frac{7}{9} \times \frac{2}{8} + \frac{3}{10} \times \frac{2}{9} \times \frac{1}{8} = \frac{3}{10}.$$

Esimerkiksi “001” tarkoittaa, että on nostettu ensin kaksi mustaa ja kolmanneksi liila pallo.

Huom! Todennäköisyys, ehdolla aiempien nostojen tulokset, ylipäänsä poikkeaa edellä lasketusta ehdollistamattomasta todennäköisyydestä. Jos on nostettu 3. ensimmäisellä nostolla liila pallo, todennäköisyys, ehdolla aiemmat nostot, nostaa seuraavaksi liila pallo on 0.

Tulos yleistyy: Jos palloja on n, liilan pallon nostamisen todennäköisyys on sama kaikilla i:n arvoilla, $i = 1, \ldots, n$. □

$\mathsf{HG}(l, m, n)$-jakauman pistetodennäköisyyden ja kertymäfunktion arvon lukuarvolla y voi laskea R:llä komennoilla

`dhyper(y,l,m,n)` ja `phyper(y,l,m,n)`.

Komentoihin sijoitetaan l:n, m:n ja n:n arvot.

Esimerkki. Kortin peluu (jatkoa). Vedetään hyvinsekoitetusta korttipakasta sattumanvaraisesti kortti ja katsotaan se, muttei palauteta sitä pakkaan. Toistetaan tämä 5 kertaa (jaetaan pelaajalle “käsi”). Lasketaan todennäköisyydet, että kuninkaita tulee 0, 1, 2, 3, tai 4 kappaletta (5 kappaletta ei voida saada, koska kortteja ei palauteta pakkaan) sekä todennäköisyys, että ainakin 1 nostetuista korteista on kuningas.

Kortin nostojen tulos on hypergeometrisesti jakautunut, koska kortteja ei palauteta pakkaan nostojen jälkeen. Pistetodennäköisyydet ovat:

$$\mathsf{P}(Y = 0) = \frac{\binom{4}{0}\binom{48}{5}}{\binom{52}{5}} \approx 0.658842, \quad \mathsf{P}(Y = 1) = \frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}} \approx 0.2994736,$$

$$\mathsf{P}(Y = 2) = \frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} \approx 0.0399298, \quad \mathsf{P}(Y = 3) = \frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} \approx 0.0017361$$

ja

$$\mathsf{P}(Y = 4) = \frac{\binom{4}{4}\binom{48}{1}}{\binom{52}{5}} \approx 0.0000185.$$

Todennäköisyydet voidaan laskea R:n `choose(,)`-komennon avulla. Esimerkiksi todennäköisyys 0.658842 on laskettu käskyllä

```
choose(4,0)*choose(48,5)/choose(52,5)
```

Saman tuloksen saa R:n hypergeometrisia pistetodennäköisyyksiä tuottavalla käskyllä

```
dhyper(0,4,48,5)
```

Siinä 0 on tapahtumien lukumäärä, jolle pistetodennäköisyys lasketaan, 4 on kuninkaiden lukumäärä pakassa (edellä liilat pallot pussissa), 48 on muiden korttien lukumäärä pakassa (edellä mustat pallot pussissa) ja 5 on pakasta vedettävien korttien lukumäärä (poimittujen pallojen lukumäärä edellä). Käskyssä ei tarvitse määritellä muista korteista nostettavien korttien lukumäärää, koska se määräytyy nostettavien kuninkaiden ja yhteensä nostettavien korttien lukumääristä.

Todennäköisyys, että ainakin 1 nostetuista korteista on kuningas, on noin 0.341:

$$\begin{aligned} &0.2994736 + 0.03992982 + 0.001736079 + 0.00001846893 \\ &= 0.341158 \\ &= 1 - 0.658842. \end{aligned}$$

Todennäköisyys on laskettu ensin tapahtumien erillisyyteen (kaava (4.5)) perustuen ja lopuksi lyhyemmin vastatapahtuman todennäköisyyden kautta.

Todennäköisyydet poikkeavat vähän aiemmasta kortin peluu -esimerkistä, jossa kukin vedetty kortti palautettiin pakkaan ja kuninkaiden lukumäärä noudatti binomijakaumaa. Todennäköisyys jäädä ilman kuningasta on tässä pienempi ($0.659 < 0.670$), koska kortteja pakasta vedettäessä on todennäköisempää, että saadaan muu kortti kuin kuningas. Se kasvattaa todennäköisyyttä saada myöhemmin kuningas. Myöskään todennäköisyys saada 1 tai enemmän kuninkaita ei poikkea juurikaan esimerkeissä ($0.341 > 0.330$), vaikka binomijakautuneessa tilanteessa on mahdollista saada 5 kuningasta. Sen todennäköisyys on merkityksettömän pieni. Odotusarvo kuninkaiden lukumäärälle on sama kuin binomijakautuneessa versiossa (kaavat (7.4) ja (7.8)). □

Esimerkki.[66] Toisiaan tuntemattomat Antti ja Anna tutustuvat miljoonakaupungissa — vaikkapa pääkaupunkiseudulla. Molemmilla on 500 tuttua kaupungissa. Oletetaan, että molempien tutut voidaan ajatella yksinkertaisiksi satunnaisotoksiksi (jakso 8.3) kaupunkilaisista. Mikä on todennäköisyys, että Antilla ja Annalla on ainakin yksi yhteinen tuttu?

Hahmotetaan tehtävä pallojen poimimis -analogian avulla. Pussissa on 500 liilaa (Antin tuttua) ja 999 500 mustaa palloa (muita kaupunkilaisia). Poimitaan pussista 500 palloa (Annan tutut). Todennäköisyys saada ainakin 1 liila pallo (yhteinen tuttu) on 1 miinus todennäköisyys, ettei saada yhtään liilaa palloa:

$$1 - \frac{\binom{500}{0}\binom{999\,500}{500}}{\binom{1\,000\,000}{500}} \approx 0.221.$$

Todennäköisyys Antin ja Annan yhteisille tutuille 0.221 on melko suuri.

Kombinaatioiden lukumäärä osamäärässä on niin valtaisa, että R ei pysty suorittamaan komentoa `choose(500,0)*choose(999500,500)/choose(1000000,500)`. Todennäköisyys (0.2212965) on laskettu komennolla `1-dhyper(0,500,999500,500)`. □

7.1.6 Poisson-jakauma

Poisson-jakautuneen satunnaismuuttujan $Y \sim \mathsf{Poi}(\mu)$ pistetodennäköisyysfunktio on

$$\mathsf{P}(Y = y) = \begin{cases} e^{-\mu}\dfrac{\mu^y}{y!}, & y = 0, 1, 2, \ldots \\ 0, & \text{muulloin.} \end{cases} \tag{7.10}$$

Yllä $\mu > 0$ ja e on Neperin luku (≈ 2.71828). Tällaisen satunnaismuuttujan odotusarvo ja varianssi ovat molemmat μ:

$$\mathsf{E}(Y) = \mathsf{V}(Y) = \mu. \tag{7.11}$$

(esim. Blitzstein ja Hwang 2015, 161–162). Poisson-jakautunut satunnaismuuttuja saa kokonaislukuarvoja. $\mathsf{Poi}(\mu)$-jakauman tyyppiarvo on $\mathsf{int}[\mu]$, jos μ ei ole kokonaisluku ($\mathsf{int}[x]$ on argumentin x kokonaislukuosa). Jos on, tyyppiarvoja on kaksi: $\mu - 1$ ja μ.

Poisson-jakautuneen satunnaismuuttujan varianssi kasvaa odotusarvon kanssa. Se on intuitiivista: Poisson-jakautunut satunnaismuuttuja ei voi saada nollaa pienempiä arvoja. Alarajalla on merkitystä, jos odotusarvo on pieni. Jos odotusarvo on suuri, satunnaismuuttujalla on enemmän tilaa vaihdella.

Jakauma on paljon käytetty, koska sillä voidaan approksimoida monia tilanteita. Taustalla on usein ajatus lukuisista kokeista pienellä todennäköisyydellä, niin että lopputuloksena havaitaan kokonaislukuarvoinen satunnaismuuttuja, jonka odotusarvo on μ. Pienen todennäköisyyden voi tulkita intensiteettinä,

jolla tapahtumia tulee esimerkiksi aikayksikössä, ja tarkasteltavana on tapahtumien lukumäärä aikavälillä. Lukumäärä voi yksinkertaisimmillaan noudattaa binomijakaumaa, jolloin $\mu = n\pi$ (tästä lisää lisämateriaalissa alla). Tapahtumilla voi olla eri todennäköisyydet ja ne voivat jossain määrin riippua toisistaan, ja silti Poisson-jakauma pätee. Tuloksen johto sivuutetaan matemaattisesti vaativana.

Jakauman soveltamista avittaa tulos riippumattomien Poisson-jakautuneiden satunnaismuuttujien summalle: Jos $Y_1 \sim \mathsf{Poi}(\mu_1)$ ja $Y_2 \sim \mathsf{Poi}(\mu_2)$ ja Y_1 ja Y_2 ovat riippumattomia, niin

$$Y_1 + Y_2 \sim \mathsf{Poi}(\mu_1 + \mu_2). \tag{7.12}$$

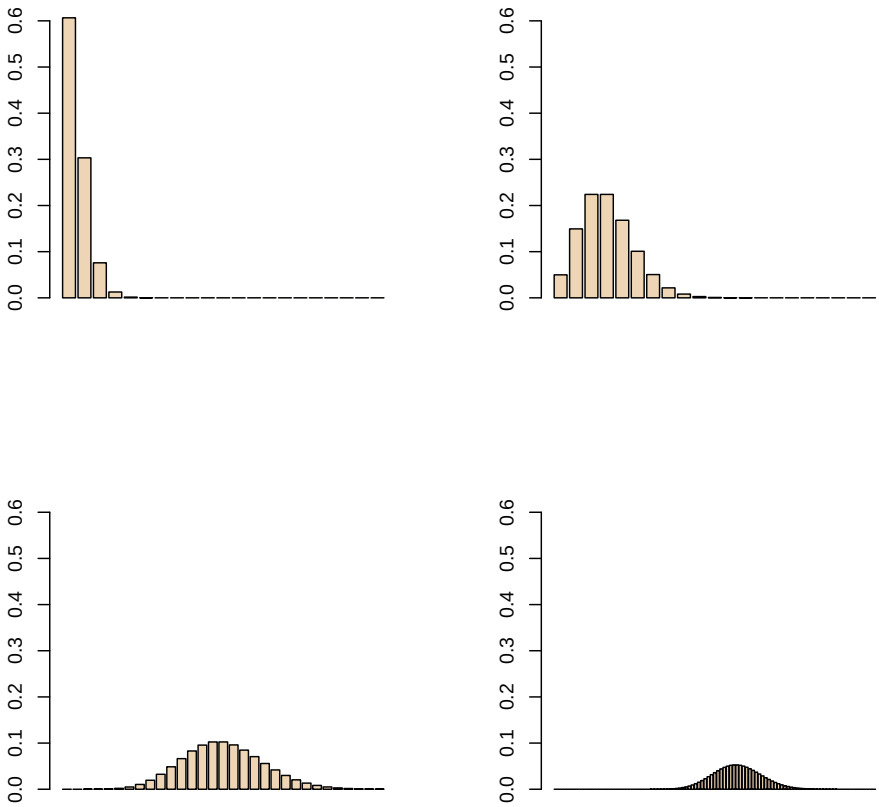
Kuva 7.3 havainnollistaa Poisson-jakaumaa μ:n arvoilla 0.5, 3, 15 ja 57. Jakauma on hyvin vino oikealle, kun odotusarvo on pieni, mutta symmetrisoituu odotusarvon suuretessa. Tyyppiarvo on 0, kun $\mu = 0.5$. Kokonaislukuarvoilla $\mu = 3$ sekä $\mu = 15$ jakaumissa näkyy molemmissa kaksi tyyppiarvoa (i:n arvoilla 2 ja 3 sekä 14 ja 15).

Esimerkki.[67] Liikenteessä liikkuu paljon ihmisiä, ja jokaisella on pieni todennäköisyys joutua liikenneonnettomuuteen. Lähtökohta viittaa Poisson-jakauman mahdollisuuteen. Liikenneonnettomuudet eivät silti ole aivan luonteva Poisson-jakauman sovellus, koska niissä monesti loukkaantuu monta ihmistä kerralla mutta Poisson-jakauman taustalla ajatellaan olevan yksittäisiä tapahtumia. Jalankulkijoiden ja pyöräilijöiden liikennekuolemat tapahtunevat yksi kerrallaan. Ne voisivat noudattaa Poisson-jakaumaa.

Seitsemän jalankulkijaa tai pyöräilijää on kuollut liikenneonnettomuuksissa Tampereella noin kolmen vuoden aikana 1/2017–11/2019. Esimerkinomaisesti ilmiötä enemmän tutkimatta oletetaan, että jalankulkijoiden ja polkupyöräilijöiden vuosittaiset kuolemat noudattavat $\mathsf{Poi}(2.3)$-jakaumaa ($7/3 \approx 2.3$). Liikennekuolemien pistetodennäköisyyksiä on kuvassa 7.4. Se on piirretty R-käskyllä `barplot(dpois(0:10, 2.3), names=0:10, ylim=c(0,0.3))` (väri lisätty). Pistetodennäköisyydet voidaan laskea sijoittamalla $\mu = 2.3$ ja y:lle yksi kerrallaan arvot $0, 1, \ldots, 12$ kaavaan (7.10). Pistetodennäköisyydet saadaan yhdellä iskulla käskyllä `dpois(0:10, 2.3)`:

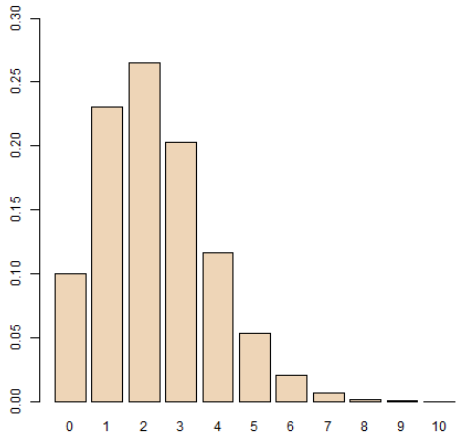
```
##  [1] 0.1002588437 0.2305953406 0.2651846416 0.2033082253 0.1169022295
##  [6] 0.0537750256 0.0206137598 0.0067730925 0.0019472641 0.0004976342
## [11] 0.0001144559
```

Vastaavat kertymäfunktion arvot laskee käsky `ppois(0:10, 2.3)`:



Kuva 7.3: Poisson-jakautuneen satunnaismuuttujan pistetodennäköisyysfunktioita μ:n arvoilla 0.5, 3, 15 ja 57.

```
## [1] 0.1002588 0.3308542 0.5960388 0.7993471 0.9162493 0.9700243 0.9906381
## [8] 0.9974112 0.9993584 0.9998561 0.9999705
```



Kuva 7.4: $\mathsf{Poi}(2.3)$-jakauman pistetodennäköisyyksiä.

Mediaani sijaitsee 1:n ja 2:n välissä, tyyppiarvo on 2 ja odotusarvo on 2.3. Todennäköisyys, että liikennekuolemia ei tapahdu lainkaan, on 0.10. Todennäköisyys, että niitä tulee yli 5 on $\mathsf{P}(Y > 5) = 1 - \mathsf{P}(Y \leq 5) = 1 - 0.9700243 = 0.0299757$ eli noin 0.03. (Relevantit pistetodennäköisyys- ja kertymäfunktion arvot on punattu.) □

7.2 Jatkuvia jakaumia

7.2.1 Normaalijakauma

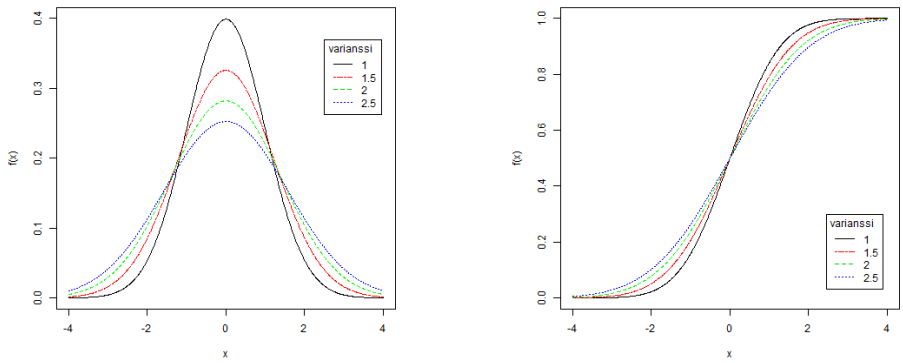
Tärkein jatkuva-arvoisen satunnaismuuttujan (X) jakauma on *normaalijakauma* $\mathsf{N}(\mu, \sigma^2)$. Sen sijainnin ja muodon määräävät odotusarvo μ ja varianssi σ^2:

$$\mathsf{E}(X) = \mu \quad \text{ja } \mathsf{V}(X) = \sigma^2.$$

Normaalijakautuneen satunnaismuuttujan tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Jos $\mu = 0$ ja $\sigma^2 = 1$, puhutaan *standardinormaalijakaumasta* $\mathsf{N}(0,1)$. Se on taulukoitu lukuisissa oppikirjoissa. Lukemattomat tilastolliset tunnusluvut voidaan muokata noudattamaan standardinormaalijakaumaa suurilla havaintomäärillä. Standardinormaalijakautunutta satunnaismuuttujaa merkitään usein Z:lla. Kuva 7.5 havainnollistaa normaalijakauman levittymistä varianssin σ^2 kasvaessa. Väli $(-4, 4)$ oleellisesti kattaa standardinormaalijakautuneen satunnaismuuttujan vaihtelun.



Kuva 7.5: $\mathsf{N}(0, \sigma^2)$-jakautuneen satunnaismuuttujan tiheys- ja kertymäfunktioita ($\sigma^2 = 1, 1.5, 2, 2.5$).

Esimerkki. $\mathsf{N}(0, 2.5)$-jakauman 0.01. kvantiili -3.678 on pienempi kuin standardinormaalijakauman 0.01. kvantiili -2.326. Kvantiilit on laskettu R-komennoilla `qnorm(0.01, 0, sqrt(2.5))` ja `qnorm(0.01)`. □

Normaalijakaumalla on kätevä ominaisuus, että lineaarikombinaatio riippumattomista normaalijakautuneista satunnaismuuttujista (X_1 ja X_2) on normaalijakautunut:

$$a_1 X_1 + a_2 X_2 \sim \mathsf{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2). \tag{7.13}$$

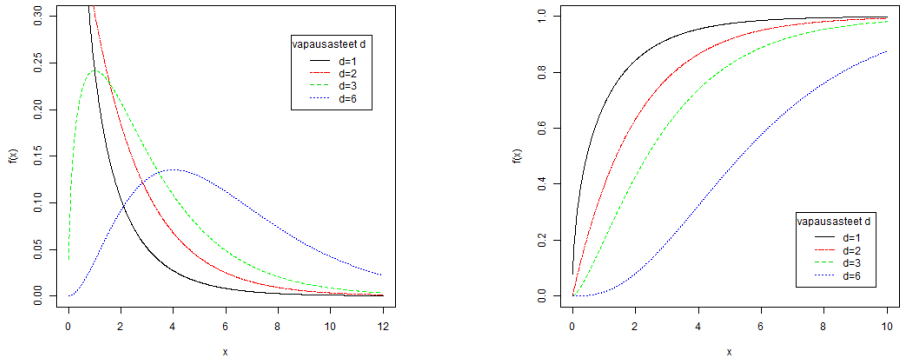
Tässä a_1 ja a_2 ovat vakioita.

7.2.2 χ^2-jakauma

Satunnaismuuttuja

$$X = \sum_{i=1}^{d} Z_i^2 \sim \chi^2(d)$$

noudattaa *χ^2-jakaumaa* d vapausasteella. Kaavassa Z_i:t ovat riippumattomia standardinormaalijakautuneita satunnaismuuttujia. Kuva 7.6 havainnollistaa jakaumaa. Pienillä vapausasteilla jakauma on vino oikealle mutta symmetrisoituu vapausasteiden kasvaessa. $\chi^2(d)$-jakautuneen satunnaismuuttujan odotusarvo ja varianssi ovat d ja $2d$. Jakaumaa käytetään muun muassa kuuluisan χ^2-testin yhteydessä sekä varianssin suuruutta testattaessa (jaksot 12.2 ja 12.5.1). Seuraavat kaksi jakaumaa voidaan määritellä standardinormaali- ja χ^2-jakautuneiden satunnaismuuttujien avulla.



Kuva 7.6: $\chi^2(d)$-jakautuneen satunnaismuuttujan tiheys- ja kertymäfunktioita ($d = 1, 2, 3, 6$).

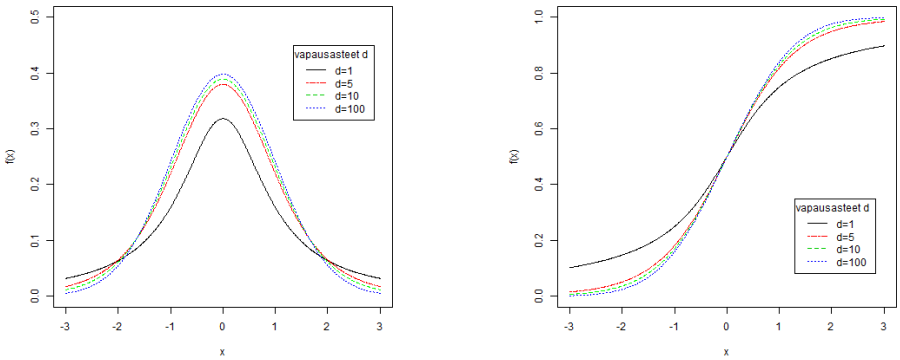
7.2.3 Studentin t-jakauma

Satunnaismuuttuja

$$\frac{Z}{\sqrt{X/d}} \sim \mathsf{t}(d)$$

seuraa t-*jakaumaa* d vapausasteella. Yllä Z on standardinormaalijakautunut, X on d vapausasteella χ^2-jakautunut satunnaismuuttuja ja Z ja X ovat riippumattomia. Kuvassa 7.7 on $\mathsf{t}(d)$-jakauman tiheys- ja kertymäfunktioita vapausasteen d eri arvoilla. Jakauman hännät ovat paksummat kuin standardinormaalijakaumalla ja erityisesti, jos vapausasteita on vähän. Vapausasteiden kasvaessa kohti ääretöntä jakauma yhtyy standardinormaalijakaumaan. Jakaumalla on paljon käyttöä tilastollisessa päättelyssä.

Esimerkki. $\mathsf{t}(1)$-jakauman 0.01. kvantiili -31.82052 on tavattomasti pienempi kuin standardinormaalijakauman 0.01. kvantiili -2.326. Vapausasteiden kasvaessa kvantiilit lähenevät: $\mathsf{t}(10)$-, $\mathsf{t}(50)$- ja $\mathsf{t}(250)$-jakaumien 0.01. kvantiilit ovat -2.764, -2.403 ja -2.341. Kvantiilit t-jakaumalle on laskettu käskyillä qt(0.01,1), qt(0.01,10), qt(0.01,50) ja qt(0.01,250). □



Kuva 7.7: $\mathsf{t}(d)$-jakautuneen satunnaismuuttujan tiheys- ja kertymäfunktioita ($d = 1, 5, 10, 100$).

7.2.4 F-jakauma

Riippumattomista $\chi^2(d_1)$- ja $\chi^2(d_2)$-jakautuneista satunnaismuuttujista X_1 ja X_2 ja niiden vapausasteista muodostettu osamäärä

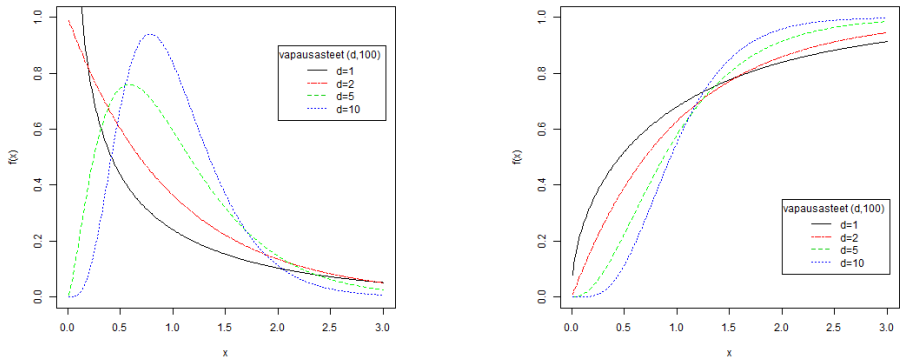
$$\frac{X_1/d_1}{X_2/d_2} \sim \mathsf{F}(d_1, d_2)$$

on F-*jakautunut* vapausasteilla d_1 ja d_2. Eri vapausasteet voivat tuottaa hyvin erimuotoisia jakaumia. $\mathsf{F}(1, d_2)$-jakautunut satunnaismuuttuja on $\mathsf{t}(d_2)$-jakautuneen satunnaismuuttujan neliö:

$$\frac{X_1/1}{X_2/d_2} = \frac{Z^2}{X_2/d_2} = \left[\frac{Z}{\sqrt{X_2/d_2}}\right]^2 \sim [\mathsf{t}(d_2)]^2.$$

Yllä X_1 on standardinormaalijakautuneen satunnaismuuttujan Z neliö ja X_2 on siitä riippumattomasti $\chi^2(d_2)$-jakautunut satunnaismuuttuja.

Kuva 7.8 havainnollistaa $\mathsf{F}(d, 100)$-jakaumaa d:n arvoilla 1, 2, 5 ja 10. Kuvasta löytyy neliöidyn $\mathsf{t}(100)$-jakautuneen satunnaismuuttujan tiheys- ja kertymäfunktioiden kuvaajat ($d_1 = d = 1$ ja $d_2 = 100$). F-jakaumaa käytetään muun muassa variansseja testattaessa sekä varianssi- ja regressioanalyysissa (jaksot 12.5 ja luku 13).



Kuva 7.8: $\mathsf{F}(d, 100)$-jakautuneen satunnaismuuttujan tiheys- ja kertymäfunktioita ($d = 1, 2, 5, 10$).

7.3 Keskeinen raja-arvolause

Olkoot $X_1, X_2, \ldots, X_n$ riippumattomia satunnaismuuttujia, joilla on odotusarvo $\mathsf{E}(X_i) = \mu$ ja varianssi $\mathsf{V}(X_i) = \sigma^2 > 0$. Muodostetaan keskiarvo $\bar{X} =$

$n^{-1}\sum_{i=1}^{n} X_i$. Sen odotusarvo ja varianssi ovat μ ja σ^2/n:

$$\mathsf{E}(\bar{X}) = \mathsf{E}\left(n^{-1}\sum_{i=1}^{n} X_i\right) = n^{-1}\sum_{i=1}^{n}\mathsf{E}(X_i) = n^{-1}n\mu = \mu$$

ja

$$\mathsf{V}(\bar{X}) = \mathsf{V}\left(\sum_{i=1}^{n} n^{-1}X_i\right) = n^{-2}\mathsf{V}\left(\sum_{i=1}^{n} X_i\right) = n^{-2}\sum_{i=1}^{n}\mathsf{V}(X_i) = n^{-2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Standardoidun satunnaismuuttujan

$$Z_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

odotusarvo on 0 ja varianssi 1:

$$\begin{aligned}\mathsf{E}(Z_n) &= \mathsf{E}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma/\sqrt{n}}\mathsf{E}(\bar{X}-\mu) = \frac{1}{\sigma/\sqrt{n}}[\mathsf{E}(\bar{X})-\mu] = \frac{1}{\sigma/\sqrt{n}}(\mu-\mu)\\ &= 0\end{aligned}$$

ja

$$\mathsf{V}(Z_n) = \mathsf{V}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma^2/n}\mathsf{V}(\bar{X}-\mu) = \frac{1}{\sigma^2/n}\mathsf{V}(\bar{X}) = \frac{\sigma^2/n}{\sigma^2/n} = 1.$$

Normaalijakauman suuri merkitys tilastotieteessä johtuu paljolti *keskeisestä raja-arvolauseesta* (*Central limit theorem*). Sen mukaan X_i:den lukumäärän n kasvaessa kohti ääretöntä Z_n:n jakaumaksi muodostuu standardinormaalijakauma. Suurilla n:n arvoilla pätee likimäärin

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \quad \text{eli} \quad \bar{X} \sim \mathsf{N}(\mu, \sigma^2/n). \tag{7.14}$$

"Keskeinen" tarkoittaa tässä perustavaa laatua olevaa. Keskiarvon $\bar{X}$ normaalisuus pätee pitkälti riippumatta satunnaismuuttujien X_i, joista keskiarvo lasketaan, jakaumasta.[68] Se on vaikuttava tulos ja selittää normaalijakauman tärkeyttä tilastotieteessä.

Henri Poincaré kuvasi keskeisen raja-arvolauseen merkitystä kieli poskella (Ross 2017, 310):

> Jokainen uskoo siihen: Empiirikot ajattelevat sen olevan matemaattinen välttämättömyys. Matemaatikot pitävät sitä empiirisenä tosiasiana.

Keskeinen raja-arvolause kohoaa sellaisiin korkeuksiin Galtonin mielessä, että suomennos ei ehkä yltäisi samaan:[69]

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency Error" [keskeinen raja-arvolause]. The Law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason.

Mahtavuudestaan huolimatta aivan kaikkiin satunnaisilmiöihin ei keskeinen raja-arvolause pure, eivätkä kaikki satunnaismuuttujat ole keskiarvoja. Jo Francis Edgeworth (1922) toppuutteli, että normaalijakauma ei ole aivan kaikkialla läsnä ja moitti normaalijakauman 1800-luvun propagoijaa Adolphe Quételet'tä sen näkemisestä sielläkin, missä se ei lymyä (Stigler 1986, 203).

7.4 Jakaumien yhteydet

7.4.1 Binomijakauma ja hypergeometrinen jakauma

Binomijakautuneen ja hypergeometrisesti jakautuneen satunnaismuuttujan odotusarvo on sama. Muuttujien varianssit eroavat *äärellisen perusjoukon korjaustekijän*

$$\frac{l+m-n}{l+m-1} \leq 1$$

verran (yhtälö (7.9)). *Otantaosuuden* (*sampling fraction*) $n/(l+m)$ supetessa kohti nollaa eli perusjoukon koon $l+m$ kasvaessa kohti ääretöntä korjaustekijä suppenee kohti yhtä. Varianssien ero häviää tällöin.

Karkea peukalosääntö on, että varianssien ja jakaumien erolla ei ole merkitystä, jos

$$\frac{n}{l+m} < 0.1$$

(vrt. yhtälö (7.8)). Seber (2013, 3) puoltaa tiukempaa sääntöä $n/(l+m) < 0.05$.

Esimerkki. Kortin peluu (jatkoa). Poimittaessa 5 korttia 52 kortin pakasta otantaosuus on $5/52 \approx 0.096 < 0.1$. Otantaosuus alittaa peukalosäännön ohjeen juuri ja juuri. Esimerkeissä edellä binomijakaumasta ja hypergeometrisesta jakaumasta lasketut pistetodennäköisyydet kuninkaiden saamisen todennäköisyyksille eivät eronneet paljoa toisistaan. □

7.4.2 Binomijakauma ja Poisson-jakauma

Binomijakaumaa voidaan approksimoida Poisson-jakaumalla, kun toistojen lukumäärä n on suuri ja todennäköisyys π on pieni:

$$\mathsf{P}(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \approx e^{-n\pi}\frac{(n\pi)^y}{y!}.$$

Approksimaation avulla binomitodennäköisyys voidaan laskea, vaikka n tai π eivät olisi tiedossa, mutta Poisson-jakauman parametri $\mu = n\pi$ yllä tunnetaan tai voidaan arvioida.

Peukalosääntöjä[70] approksimaation käyttökelpoisuudelle: $\pi \leq 0.1$ ja $n \geq 40$, $n\pi < 5$ ja $n > 50$, $\pi \leq 0.05$ ja $n \geq 20$ tai $n\pi < 10$ ja $n \geq 100$ (approksimaation pitäisi olla "erinomainen"). Esimerkkejä approksimaation toimivuudesta löytyy monista oppikirjoista.[71] Taulukko alla on Tijmisin (2012, 112) kirjasta. Vasemmanpuolimmaisessa sarakkeessa on tapahtumien lukumäärä (y). Siitä oikealla on lukumäärään liittyvä pistetodennäköisyys binomijakaumasta laskettuna, kun π toteuttaa yhtälön $n\pi = 1$. Oikeanpuolimmaisessa sarakkeessa on $\mathsf{Poi}(1)$-jakaumasta laskettu pistetodenäköisyys lukumäärälle. Esimerkiksi jos $n = 100$ ja $\pi = 0.01$, niin 0:lle tapahtumalle binomijakauman pistetodennäköisyys on 0.3660 ja Poisson-jakauman 0.3679. Jakaumien pistetodennäköisyydet ovat varsin samanlaisia jo, kun $n = 25$. Jos Poisson-jakauman parametri (μ) olisi suurempi, yhteensopivuus ei välttämättä olisi näin hyvä (esim. Armitage ym. 2002, 76).

	binomitodennäköisyys ($n\pi = 1$)				$\mathsf{Poi}(1)$
y	$n = 25$	$n = 100$	$n = 500$	$n = 1000$	
0	0.3604	0.3660	0.3675	0.3677	0.3679
1	0.3754	0.3697	0.3682	0.3681	0.3679
2	0.1877	0.1849	0.1841	0.1840	0.1839
3	0.0600	0.0610	0.0613	0.0613	0.0613
4	0.0137	0.0149	0.0153	0.0153	0.0153
5	0.0024	0.0029	0.0030	0.0030	0.0031

Noudattakoon satunnaismuuttuja Y binomijakaumaa $\mathsf{Bin}(n, \pi)$:

$$\mathsf{P}(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y},$$

jossa $0 < \pi < 1$.

Voidaan osoittaa, että

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{-n} \approx 2.71828.$$

Yhtäsuuruuksista seuraa, että

$$\begin{aligned}\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n &= \lim_{n\to\infty}\left(1+\frac{1}{n/a}\right)^n \\ &= \lim_{n\to\infty}\left[\left(1+\frac{1}{n/a}\right)^{n/a}\right]^a \\ &= e^a.\end{aligned}$$

Yllä $a \neq$ on vakio. (Jos $a = 0$, niin $\lim_{n\to\infty}(1+0/n)^n = 1 = e^0$.)

Merkitään $\mu \equiv n\pi$. Ilmaistaan binomitodennäköisyys μ:n avulla ja annetaan havaintojen lukumäärän n kasvaa kohti ääretöntä siten, että tulo $\mu = n\pi$ on vakio (todennäköisyys π lähenee nollaa samalla vauhdilla kuin n kasvaa kohti ääretöntä):

$$\begin{aligned}\lim_{\substack{n\to\infty\\ n\pi=\mu}}\binom{n}{y}\pi^y(1-\pi)^{n-y} &= \lim_{\substack{n\to\infty\\ n\pi=\mu}}\binom{n}{y}\left(\frac{n\pi}{n}\right)^y\left(1-\frac{n\pi}{n}\right)^{n-y} \\ &= \lim_{n\to\infty}\binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y} \\ &= \lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^{n-y}\frac{\mu^y}{n^y}\frac{n!}{y!(n-y)!} \\ &= \lim_{n\to\infty}\left(1+\frac{-\mu}{n}\right)^{n-y}\frac{\mu^y}{y!}\frac{n(n-1)\times\cdots\times(n-y-1)}{n^y} \\ &= e^{-\mu}\frac{\mu^y}{y!}.\end{aligned}$$

Tulos on Poisson-pistetodennäköisyys.

Esimerkki.[72] Oikean (tai lähes oikean) lottorivin todennäköisyys määräytyy hypergeometrisesta jakaumasta. Lottoajia on paljon, kunkin mahdollisuus veikata oikein (tai lähes oikein) lottorivi on hyvin pieni ja riippumaton toisista lottoajista. Oikeiden (tai lähes oikeiden) lottorivien lukumäärä on siten binomijakautunut sopivalla ajanjaksolla (esim. muutama vuosi). Jollei täytettyjen lottorivien lukumäärä ole tiedossa, ei binomijakauma ole käytettävissä. Kun binomijakauman taustalla olevien Bernoulli-kokeiden lukumäärä on suuri ja kussakin kokeessa tapahtuman (tässä lähes oikean lottorivin) todennäköisyys on pieni, voidaan binomijakaumaa approksimoida Poisson-jakaumalla. Approksimaation käyttöön ei tarvita tietoa Bernoulli-kokeiden lukumäärästä.

Ontariolaisessa pikkukaupungissa asuva Bob Edmonds lottosi aina ainakin yhdellä samalla lottorivillä. Hän tapasi antaa lottokupongin myyjän tarkistaa,

ovatko hänen lottorivinsä voittaneet. Edmonds osti kaksi lottoriviä 27.7.2001. Toinen lottoriveistä voitti 250 000 dollaria. Kupongin myyjä ei kertonut sitä Edmondsille vaan lunasti voiton itse. Edmonds ymmärsi myöhemmin, mitä oli tapahtunut, ja haastoi Ontarion veikkausyhtiön oikeuteen. Kolmen ja puolen vuoden oikeustaistelun jälkeen yhtiö suostui maksamaan Edmondsille 200 000 dollaria ehdolla, että hän ei paljasta julkisuuteen, mitä oli tapahtunut. Tällaisistä väärinkäytöksistä nousi kohu. Julkisuuden paineessa veikkausyhtiö maksoi Edmondsille koko voittosumman ja pyysi anteeksi tapahtunutta — vain päiviä ennen kuin Edmonds kuoli syöpään 2.4.2007.

Kanadan kansallinen TV-yhtiö CBC pyysi tilastotieteen professori Jeffrey Rosenthalia tutkimaan, onko lottovoittojen jakamisessa tapahtunut vastaavia huijauksia, joissa lottomyyjät olisivat lunastaneet itselleen asiakkaittensa voittoja. TV-yhtiön selvitysten mukaan Ontarion alueella oli vuosina 1999–2006 jaettu 5 713 suurta (yli 50 000 dollarin) voittoa. Niiden lunastajista (vähintään) 200 (3.5 %) oli lottokuponkien vähittäismyyjiä. Heitä oli Ontariossa noin 60 000 ja he lottosivat keskimäärin puolitoistakertaisella summalla keskiverto-ontariolaiseen verrattuna. Koska Ontariossa oli noin 8 900 000 täysi-ikäistä asukasta, Rosenthal arvioi, että loton vähittäismyyjien suurten lottovoittojen lukumäärän voisi odottaa olevan noin 57:

$$5713 \times \frac{60000 \times 1.5}{8900000} \approx 57.$$

Se on luonteva estimaatti eli arvio Poisson-jakauman $\mathsf{Poi}(\mu)$ odotusarvolle eli Poisson-jakauman määrittävälle parametrille μ. Rosenthal päätteli, että approksimatiivisesti $Y \sim \mathsf{Poi}(57)$, jossa satunnaismuuttuja Y on vähittäismyyjien saamien suurten lottovoittojen lukumäärä. Tällöin

$$P(Y \geq 200) = 1 - P(Y < 200) = 1 - \sum_{i=0}^{199} \frac{57^i e^{-57}}{i!}.$$

R-komento `1-ppois(199,57)` laskee 0:ksi erotuksen 1:n ja Poisson-jakauman $\mathsf{Poi}(57)$ kertymäfunktion arvon pisteessä 199 välillä. Rosenthalin mukaan todennäköisyys on alle yhden suhde triljoonaan triljoonaan triljoonaan triljoonaan. Tehdyillä oletuksilla todennäköisyys, että loton vähittäismyyjistä vähintään 200 voittaa suuren lottovoiton, on oleellisesti nolla.

Rosenthal teki muitakin vastaavia analyysejä. CBC:n väärinkäytöksistä kertova dokumenttiohjelma lähetettiin 25.10.2006, ja ne olivat pääuutisia seuraavina päivinä Kanadassa. Ontarion veikkausyhtiö yritti kiistää väärinkäytökset

mutta joutui myöntämään ne ja muuttamaan käytäntöjään: Loton vähittäismyyjien kuponkien tarkastuskoneiden tulee nykyään olla asiakkaiden nähtävissä ja niiden pitää hälyttää voitoista äänimerkillä, asiakkaiden täytyy allekirjoittaa lottokuponkinsa eivätkä loton vähittäismyyjät saa enää ostaa lottokuponkeja omasta myymälästään. Yhtiön toimitusjohtaja erotettiin. Vastaavat tutkimukset käynnistettiin Brittiläisessä Kolumbiassa. Myös siellä havaittiin väärinkäytöksiä, ja Brittiläisen Kolumbian veikkausyhtiön toimitusjohtaja erotettiin. Tutkimukset levisivät muualle Kanadaan ja Yhdysvaltoihin. Suurin paljastunut huijaus oli 12 500 000 dollarin voiton väärä lunastus. Rosenthalin analyysit johtivat toimitusjohtajien erottamisen lisäksi useisiin vankilatuomioihin ja miljoonien dollarien maksatukseen ihmisille, joilta oli huijattu voitto. Huijauksia on paljastunut myöhemmin lisää ulkomailla ja Suomessa.[73] □

7.4.3 Binomijakauma ja normaalijakauma

Binomijakautunut satunnaismuuttuja Y on summa Bernoulli-jakautuneista satunnaismuuttujista X_i: $Y = \sum_{i=1}^{n} X_i$. Muodostetaan keskiarvo $\bar{X} = \sum_{i=1}^{n} X_i/n$. Keskeisen raja-arvolauseen (7.14) ja Bernoulli-jakautuneen satunnaismuuttujan varianssin (7.2) perusteella suurilla n:n arvoilla

$$\bar{X} \sim \mathsf{N}(\pi, \pi(1-\pi)/n).$$

Koska $\bar{X}$ on vakiolla (n) jaettu binomijakautunut Y, on binomijakaumankin muoto suurilla havaintomäärillä normaalijakauman

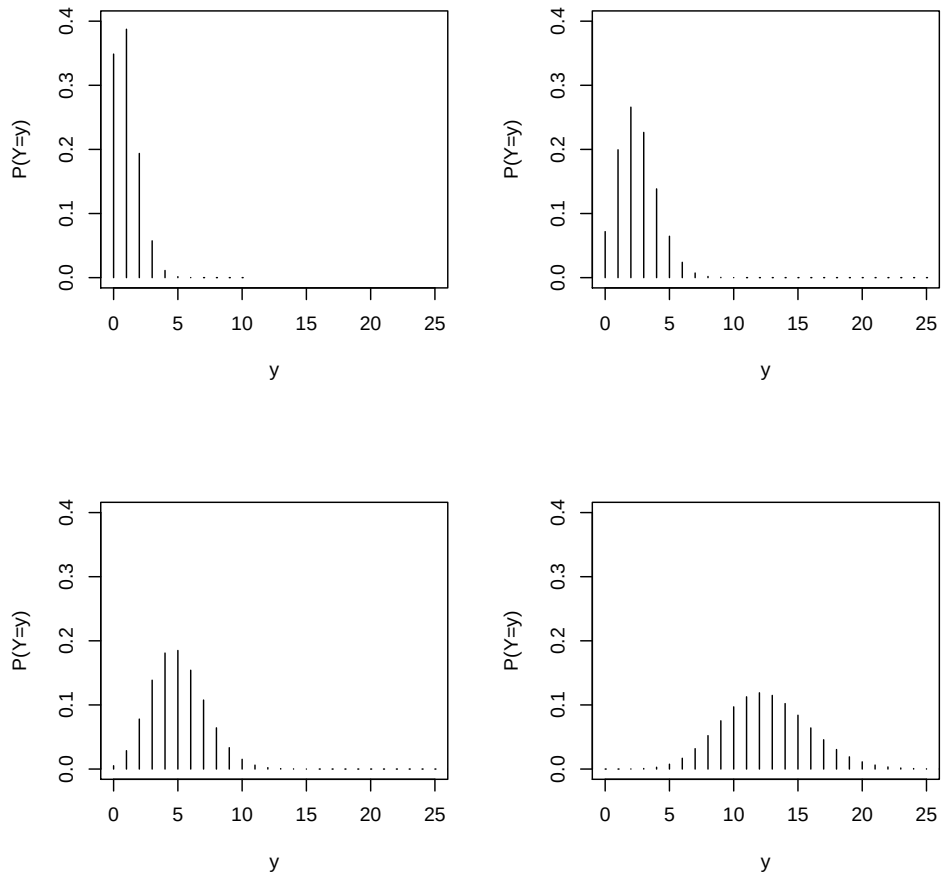
$$\mathsf{N}(n\pi, n\pi(1-\pi))$$

tapainen. Kuva 7.9 havainnollistaa binomijakauman ($\pi = 0.1$) muotoutumista normaalijakaumaksi n:n kasvaessa.

Approksimaation toimivuudelle on esitetty peukalosääntöjä kuten $\sqrt{n\pi(1-\pi)} > 3$ tai $n\pi > 5$ ja $n(1-\pi) > 5$. Säännön toteutuminen ei takaa, että approksimaatio olisi riittävän hyvä. Riittävyys riippuu käyttötarkoituksesta.

Esimerkki. Huoltoriidat käräjäoikeuksissa (jatkoa). Lapsista 35 osoitettiin asumaan isän ja 83 äidin luona (118 havaintoa). Edellä pohdittiin todennäköisyyttä tällaiselle tai pienemmälle isien voittojen lukumäärälle, jos isän ja äidin voittotodennäköisyys on sama. $\mathsf{Bin}(118, 0.5)$-jakauman kertymäfunktion arvoksi pisteessä 35 laskettiin noin 0.000006. Lasketaan approksimaatio tälle todennäköisyydelle normaalijakauma-approksimaatiolla.

Merkitään Y:llä isien luokse asumaan osoitettavien lasten lukumäärää. Tehtävän tilanteessa $\mathsf{E}(Y) = n\pi = 118 \times 0.5 = 59$ ja $\mathsf{V}(Y) = n\pi(1-\pi) =$



Kuva 7.9: Binomijakautuneen satunnaismuuttujan pistetodennäköisyysfunktioita n:n arvoilla 10, 25, 50 ja 125, kun $\pi = 0.1$.

$118 \times 0.5 \times 0.5 = 29.5$. Normaalijakauma-approksimaation mukaan todennäköisyys, että satunnaisessa 118:n havainnon aineistossa lapset osoitetaan isälle

asumaan päätöksistä 35:ssä tai pienemmässä lukumäärässä on

$$\begin{aligned}\mathsf{P}(Y \leq 35) &= \mathsf{P}\left(\frac{Y-59}{\sqrt{29.5}} \leq \frac{35-59}{\sqrt{29.5}}\right) \approx \mathsf{P}(Z \leq \frac{35-59}{\sqrt{29.5}}) = \Phi\left(\frac{35-59}{\sqrt{29.5}}\right) \\ &= \Phi\left(-4.418758\right) \approx 0.000005.\end{aligned}$$

Yllä Z on standardinormaalisti jakautunut satunnaismuuttuja ja $\Phi(z)$ on standardinormaalijakauman kertymäfunktion arvo pisteessä z. Kertymäfunktion arvo on laskettu R:n käskyllä

```
pnorm(-4.418758)
```

Todennäköisyys on approksimaation mukaan noin 0.000005. Se on lähes sama kuin todellinen todennäköisyys 0.000006.

Todennäköisyys voitaisiin laskea myös osuudesta lasketun standardoidun suureen

$$\frac{\bar{y}-\pi}{\sqrt{\pi(1-\pi)/n}} = \frac{0.2966102-0.5}{\sqrt{0.5\times 0.5/118}} = -4.418758$$

avulla. Yllä $\bar{y} = 35/118 = 0.2966102 = \hat{\pi}$ on havaittu osuus. □

Approksimaatio helpotti aiemmin suuresti binomijakauman kertymätodennäköisyyksien laskua. Nykyään niitä on helppo laskea tilasto-ohjelmistoilla. Approksimaatio edelleen konkretisoi binomijakauman muotoutumisen suurilla toistokoemäärillä, helpottaa tunnuslukujen likimääräisten jakaumien johtamista ja käyttämistä sekä erityisesti luottamusvälien laskua.

7.4.4 Galtonin kone

Francis Galton (1877) kuvasi sittemmin Galtonin koneena (*Galton's machine*, *Galton board*, *Galton box* tai *Quincunx*) tunnetun mekaanisen laitteen. Sillä voidaan konkreettisella tavalla havainnollistaa binomi- ja normaalijakauman yhteys. Kuvissa 7.10–7.13 on Galtonille 1873 tehty kone, Galtonin luonnos koneestaan (1889, 63), Pearsonin luonnos koneen yleistyksestä 1895 ja uusi Galtonin kone.[74]

Suppilosta tippuu kuulia, jotka päätyvät koneen alaosassa oleviin numeroituihin $(0, 1, \ldots, n)$ laareihin $(n+1$ kappaletta). Välissä on pyramidin muodossa n riviä pinnejä (*i.* rivillä i pinniä, $i = 1, \ldots, n$) niin, että 1. pinni on suppilon suun keskipisteen alapuolella.

Kuula, joka on päätynyt laariin "y" $(k = 0, \ldots, n)$, on pompannut y kertaa oikealle todennäköisyydellä π ja $n - y$ kertaa vasemmalle todennäköisyydellä $1 - \pi$. Kunkin laariin "y" johtavan polun todennäköisyys on $\pi^y(1-\pi)^{n-y}$.

Jos $\pi = 1 - \pi = 0.5$ — kuten alla oletetaan — niin jokaisen polun mihin tahansa laariin todennäköisyys on $(0.5)^n$. Keskimmäisiin laareihin päätyy kuitenkin enemmän kuulia kuin laitimmaisiin, koska edellisiin johtaa useampia polkuja — kaikista laitimmaisiin vain yksi.

Kuinka monta polkua johtaa laariin "y"? Tarkastellaan oransseja ja vihreitä alkioita ("o" ja "v"), joita on y ja $n-y$ kappaletta (yhteensä n alkiota). Tuloksen (5.4) mukaan ne voidaan järjestää binomikertoimen $\binom{n}{y}$ mukaiseen määrään erilaisia n alkiota sisältävään jonoon. Korvataan o- ja v-alkiot pompuilla oikealle ja vasemmalle. Polku laariin "y" koostuu y pompusta oikealle ja $n-y$ pompusta vasemmalle. Päättelyn edellä mukaan tällaisia järjestyksiä eli polkuja laariin "y" on $\binom{n}{y}$ kappaletta.

Polut ovat erillisiä, ja jokainen polku on yhtä todennäköinen. Todennäköisyys, että kuula päätyy laariin "y", saadaan summaamalla kaikkien laariin "y" johtavien polkujen todennäköisyys eli kertomalla yhden polun todennäköisyys polkujen lukumäärällä $\binom{n}{y}$ $(y = 0, \ldots, n)$:

$$\begin{aligned} \mathsf{P}(\text{kuula päätyy laariin "}y\text{"}) &= \pi^y(1-\pi)^{n-y} + \ldots + \pi^y(1-\pi)^{n-y} \\ &= \binom{n}{y}\pi^y(1-\pi)^{n-y}. \end{aligned}$$

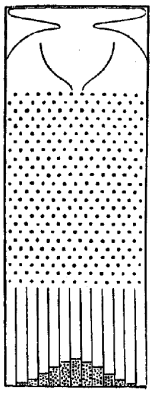
Kuulan päätyminen laariin "y" noudattaa binomijakaumaa. Lisäämällä luonteva oletus $\pi = 0.5$ todennäköisyydeksi saadaan

$$\begin{aligned} \mathsf{P}(\text{kuula päätyy laariin "}y\text{"}) &= \binom{n}{y}(0.5)^y(1-0.5)^{n-y} \\ &= \binom{n}{y}(0.5)^n. \end{aligned}$$

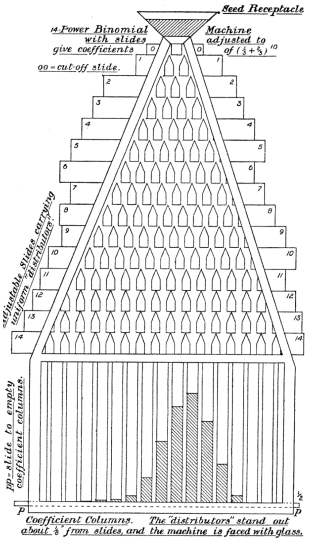
Galtonin kone osoittaa konkreettisesti, kuinka kuulien lukumäärän kasvaessa laareihin muodostuu normaalijakauman kuvio, vaikka jakauma on binomijakauma (kuvat 7.10–7.11, 7.13). Normaalijakauma approksimoi binomijakaumaa.

Useimmat Galtonin koneet havainnollistavat tilannetta $\pi = 0.5$ (kuula pomppaa samalla todennäköisyydellä vasemmalle tai oikealle). Pearsonin (1895) versiossa pinnirivien paikkaa siirtämällä voidaan havainnollistaa binomijakaumaa muillakin π:n arvoilla (kuva 7.12; myös Yule 1911, 294). Kuvassa 7.12 kuulan todennäköisyys pompata vasemmalle on $\pi = 1/3$ ja oikealle $1-\pi = 2/3$. Proschan ja Shaw (2016, 156–159) kuvaavat monimutkaisempia Galtonin koneita.

Kuva 7.10: Ensimmäinen Galtonin kone 1873.



Kuva 7.11: Galtonin (1889, 63) luonnos koneestaan.



Kuva 7.12: Pearsonin luonnos 1895.

Kuva 7.13: Galtonin kone 2016.

7.4.5 Poisson-jakauma ja normaalijakauma

Koska binomijakaumaa voi approksimoida Poisson- tai normaalijakaumalla, ei ole ehkä yllättävää, että Poisson- ja normaalijakauman välillä on kytkös. Tuloksen (7.12) mukaan $\mathsf{Poi}(n)$-jakautunut satunnaismuuttuja Y voidaan ilmaista n:n $\mathsf{Poi}(1)$-satunnaismuuttujan Y_i summana: $Y = \sum_{i=1}^{n} Y_i$. Muodostetaan keskiarvo $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. Summattavien odotusarvo ja varianssi ovat 1. Keskeisestä raja-arvolauseesta (7.14) seuraa, että likimain pätee

$$\bar{Y} \sim \mathsf{N}(1, 1/n) \quad \text{eli} \quad Y \sim \mathsf{N}(n, n).$$

Tuloksen mukaan Poisson-jakauman tulisi normalisoitua jakauman odotusarvon kasvaessa. Kuvan 7.3 mukaan niin näyttää käyvän.

7.5 Yhteisjakauma ja korrelaatio

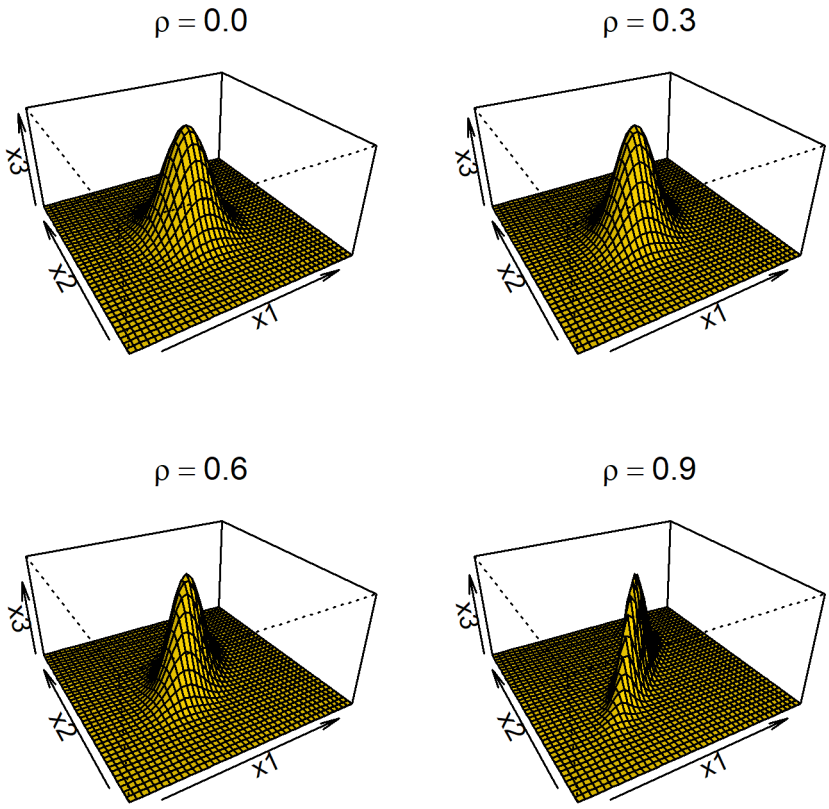
Kahden satunnaismuuttujan X_1 ja X_2 *yhteisjakauma* (*joint distribution*) määrää diskreettien satunnaismuuttujien tilanteessa lukuparin (x_1, x_2) todennäköisyyden $\mathsf{P}(X_1 = x_1 \text{ ja } X_2 = x_2)$ ja jatkuvien satunnaismuuttujien tilanteessa todennäköisyyden, että satunnaismuuttujat saavat arvon lukuparin (x_1, x_2) ympäristössä (esim. pisteiden $x_1 \pm 0.01$ ja $x_2 \pm 0.01$ rajaamalla alueella). Jälkimmäisessä tilanteessa yhteisjakaumaa kutsutaan *yhteistiheysfunktioksi*.

Kuva 7.14 havainnollistaa yhteistiheysfunktiota ja korrelaation vaikutusta siihen.[75] Akseleilla x_1 ja x_2 on satunnaismuuttujien X_1 ja X_2 arvot. Akselin x_3 suuntaan mitataan todennäköisyyttä. Mitä korkeammalle yhteistiheysfunktion arvoja (x_3) kuvaava pinta kurottaa pisteessä (x_1, x_2), sitä suurempia todennäköisyyksiä liittyy pisteen ympäristöön.

Korrelaation ollessa 0, satunnaismuuttujat jakautuvat kuvassa toisistaan riippumattomasti. Korrelaation suuretessa satunnaismuuttujien arvot pyrkivät keskittymään yhä tiiviimmin positiivisen kulmakertoimen omaavan suoran ympärille: toisen satunnaismuuttujan ollessa suuri tai pieni myös toinen tapaa olla suuri tai pieni.

Jakaumat kuvassa 7.14 ovat esimerkkejä *binormaalijakaumasta*. Sen keskeisiä piirteitä on, että siihen liittyvät satunnaismuuttujat ovat normaalijakautuneita (tarkempi kuvaus sivuutetaan).

Binormaalijakaumalla on paljon käyttöä. Tässä sillä havainnollistettiin ennen kaikkea korrelaatiota. Jatkossa binormaalijakaumaa sivutaan korrelaation suuruutta estimoitaessa ja testatessa.



Kuva 7.14: Korrelaatio ja binormaalijakauma.

Luku 8

Otannan teoriaa ja empiriaa

> Lusikallinen riittää hyvin sekoitetun keiton maistamiseen.[76]
>
> George Gallup (1901–1984)

> Lasku- ja makkarakoneet ovat siinä samanlaisia, että tuloksen arvo riippuu siitä, mitä koneeseen työnnetään.[77]
>
> Wilhelm Johannsen (1857–1927)

Miten selvittää, millainen on 1 000 000 000 000 000 000 alkiosta koostuva populaatio? Vastaus on yksinkertaisuudessaan hämmästyttävä: Poimitaan populaatiosta satunnaisotos ja tutkitaan se. Lähentelee ihmettä, että 100:n kokoisella otoksella voi päästä populaation jostain piirteestä jyvälle varsin hyvin ja 1 000:n kokoisella jo kerrassaan mainiosti. Otanta on maailman tehokkain yleispätevä tiedonkeruumenetelmä.

8.1 Käsitteitä

Havainto tai aineisto on tilastotieteen keskeisin käsite (jakso 2.1). Ilman aineistoa — teoreettista tai empiiristä — ei ole tilastotiedettäkään. Tilastotieteellisestä näkökulmasta aineistossa on usein oleellisinta, miten se on muodostettu. Menettelyä, jolla aineisto kootaan, kutsutaan *otannaksi* (*sampling*). Otannalla saatua aineistoa kutsutaan *otokseksi* (*sample*). Otos on osajoukko *perusjoukosta* eli *populaatiosta* (*population*), josta se on kerätty. Populaation määrittelevät kaikki sen alkiot. Myös satunnaismuuttujan todennäköisyysjakaumaa voidaan ajatella populaationa. Perusjoukon käsite on vilahdellut edellä Bernoulli-kokei-

den yhteydessä, kun on pohdittu, palautetaanko poimittu alkio perusjoukkoon ennen seuraavaa Bernoulli-koetta.

Populaatio on yleensä varsinainen kiinnostuksen kohde. *Kokonaistutkimus* (*complete enumeration*) eli populaation tutkiminen kokonaisuudessaan voi olla vaikeaa tai mahdotonta vaikkapa populaation suuren koon takia. Populaatio voi olla konkreettinen tai abstrakti. Jos populaatio on pieni ja kokonaistutkimus mahdollinen, ei otantaa tarvita.

Esimerkki. Suuri populaatio I. Tieto suomalaisten hyvinvoinnista on terveyspoliittisesti tärkeää. Kaikkien ruokailutottumuksia, päivittäistä liikuntaa, verenpainetta, kolesteroliarvoja jne. ei voida tutkia. Ne voidaan selvittää otoksen alkioista. Otoksen kooksi riittää murto-osa suomalaisia. □

Esimerkki. Pieni populaatio. Kiinnostava populaatio voi olla tarkkaan rajattu ryhmä kuten yliopiston pääsykokeeseen osallistuneet. Jos kiinnostuksen kohde on osallistuneiden ikä- tai (henkilötunnuksen mukainen) sukupuolijakauma, se on helposti selvitettävissä. Otantaa ei tarvita. Jos kiinnostuksen kohde on heidän keskipituutensa, älykkyysosamääränsä tai tietyn geenin omaaminen, otantaa tarvittanee. Kaikkia ei voitane tutkia, mutta otos voitaisiin. □

Esimerkki. Suuri populaatio II. Vaikka suuresta populaatiosta olisi periaatteessa mahdollista selvittää kiinnostava tieto kaikista populaation alkioista, otanta voi olla perusteltua. Yrityksen työntekijät voivat olla helposti tavoitettavissa s-postitse. Jos tarkoitus on seurata heidän hyvinvointiaan neljännesvuosittain, voi olla järkevää lähettää työhyvinvointikysely vain otokselle henkilökunnasta. Kyselyn täyttäminen on poissa muusta työajasta ja henkilökunnan halukkuus vastata kyselyihin voisi vaarantua, jos kaikki saisivat kyselyn joka kvartaali. □

Otoksesta pyritään tekemään päätelmiä populaatiosta. *Tilastollinen päättely* (*statistical inference*) on keskeistä tilastotieteessä. Tilastollisessa päättelyssä arvioidaan aineiston avulla populaatiota ja arvioiden luotettavuutta. Jotta päätelmät olisivat tilastotieteellisesti perusteltuja ja niiden luotettavuus mitattavissa, tulee otanta olla suoritettu tarkoituksenmukaisella tavalla ja olla kohdistettavissa tarkoitettuun *kohdepopulaatioon* (*target population*).

Esimerkki. Abstrakti populaatio I. Urheiluhullut, rakkausrunojen kirjoittajat tai musikaalisesti lahjakkaat voivat olla kiinnostavia populaatioita. Tällaisia populaatioita voi olla vaikea määrittää, eikä otantaa voi kohdistaa niihin yksiselitteisesti. Tällöin otannalla ei välttämättä saada toivottua tietoa. □

Jaksossa 6.2 todettiin, että kiinnostus kohdistuu usein satunnaisilmiön piirteeseen tai piirteisiin, joita kuvaa yksi tai useampi parametri. Otannalla pyritään tyypillisesti selvittämään parametrien suuruus.

Koostukoon otos satunnaismuuttujista $X_1, \ldots, X_n$. Niistä voidaan laskea *tunnusluku* (*statistic*)

$$t(X_1, \ldots, X_n).$$

Tunnusluku on funktio otoksen satunnaismuuttujista tai havainnoista. Tyypillisesti tunnusluvulla pyritään arvioimaan eli estimoimaan populaation parametria. Tällöin on kyse parametrin *estimaattorista* tai sen tietyssä otoksessa saamasta numeerisesta arvosta *estimaatista*

$$t(x_1, \ldots, x_n).$$

Yllä x_i:t ovat X_i:den toteumia.

Monesti puhutaan *piste-estimaattorista* tai *piste-estimaatista* (*point estimator*, *point estimate*; luku 9). Tällöin korostetaan, että yhdellä luvulla pyritään arvioimaan populaation parametria eikä yritetä arvioida, millä välillä parametri on. (Väliestimointi piilutaan luvussa 10.)

Parametrin ja sen estimaattorin ero on keskeinen. Edellinen on kiinteä luku, joka kuvaa populaatiota eli maailmaa, jota tutkitaan. Jälkimmäinen on satunnaismuuttuja, jonka toteuma (estimaatti) havaitaan. Estimaatti on havaittu numeerinen arvio maailmasta.

Esimerkki. Abstrakti populaatio II. Populaatioksi voidaan ajatella ääretön määrä silmälukuja kuvitteellisista nopan heitoista tai silmälukujen todennäköisyysjakauma. Abstraktius ei ole ongelma otannan kannalta. Äärellinen määrä riippumattomia heittoja kyseisellä nopalla ovat ikään kuin otos äärettömän suuresta silmälukujen populaatiosta. Otos kohdentuu tarkasti populaatioon, ja sillä voidaan selvittää populaation ominaisuuksia, kuten ovatko silmäluvut yhtä todennäköisiä kyseisellä nopalla. Parametreja ovat silmälukujen todennäköisyydet. Niitä voidaan estimoida heittämällä noppaa monta kertaa riippumattomasti ja laskemalla kunkin silmäluvun osuus heitoista. □

Populaatiota kuvaavia parametreja tavataan merkitä kreikkalaisilla kirjaimilla ja niiden estimaattoreita sijoittamalla "$\wedge$" ("hattu") niiden päälle. Merkitään estimoitavaa parametria θ:lla ("theeta") ja sen estimaattoria $\hat{\theta}$:lla ("theeta hattu"). Estimaatti saa harvoin saman arvon kuin estimoitava parametri. Niiden erotus on *otanta-* tai *estimointivirhe* (*sampling error*, *estimation error*)

$$\hat{\theta} - \theta.$$

Sen suuruuden ja ominaisuuksien tutkiminen on tärkeää tilastotieteessä. Otantavirhe pyritään saamaan mahdollisimman pieneksi.

Estimaattoria ja estimaattia merkitään monesti samalla symbolilla ($\hat{\pi}$ esimerkissä alla), ja niin tehdään pääsääntöisesti tässäkin tekstissä. Monesti myös kirjoitetaan tai puhutaan estimaatista estimaattorin sijaan, vaikka tarkoitetaan jälkimmäistä. Silti on tärkeää hahmottaa estimaattorin ja estimaatin eli satunnaismuuttujan ja sen toteuman ero ($\bar{X}$ ja $\bar{x}$ esimerkissä alla).

Esimerkki. Puoluekannatus (jatkoa). Demokratia edellyttää äänestysikäisten suomalaisten poliittisten kantojen selvittämistä. Se tehdään vaaleilla. Vaaleja järjestetään harvoin, muun muassa koska ne ovat suuritöisiä ja kalliita. Vaalien välissä poliittisia kantoja tutkaillaan vaivattomammin ja pienemmillä kustannuksilla kyselyillä, joissa tiedustellaan äänestysikäisiltä suomalaisilta heidän puoluekantaansa. Kiinnostuksen kohteena oleva populaatio on äänestysikäiset suomalaiset.

Otos on haastatellut suomalaiset. Tutkittava parametri on tiettyä puoluetta kannattavien suomalaisten osuus $\pi \in [0, 1]$. Kukin otokseen tuleva havainto X_i on Bernoulli-jakautunut satunnaismuuttuja, joka voi saada arvon 1 (kannattaa puoluetta) tai 0 (ei kannata puoluetta). Otoksesta voidaan laskea kannatusosuuden estimaattori keskiarvo $\hat{\pi} = \bar{X} = \sum_{i=1}^{n} X_i/n$. Kun otos on kerätty ja saatu havainnot x_i, voidaan laskea osuuden estimaatti $\hat{\pi} = \bar{x} = \sum_{i=1}^{n} x_i/n$.

Jos populaatiossa V-puolueen kannatus on 33.33 % mutta otoksessa 32 %, otantavirhe on 1.33 %-yksikköä. □

Esimerkki.[78] Otannan tulos 100 %. Anttila (1966) tutki piilorikollisuutta helsinkiläisiltä ja rovaniemeläisiltä asevelvollisuuskutsuntoihin tulleilta syksyllä 1962. Helsinkiläisistä vastaajista 85.5 % (1 257) kertoi syyllistyneensä johonkin lomakkeessa osoitetuista rikoksista. Lomakkeessa ei kysytty pahoinpitely- tai liikennerikkomuksista. Anttila arvioi kaikkien syyllistyneen rikokseen kohdepopulaatiossa:

> Luultavaa myös on, että jos tiedustelu olisi koskenut kaikkia rikoksia – – tuskin kukaan heistä olisi osoittautunut täysin lainkuuliaiseksi. Nuorisorikollisuutta voidaan siten nyky-yhteiskunnassa pitää tilastollisesti normaalina.

□

8.2 Tärkeitä otossuureita

Populaatiosuureet kuten kertymäfunktio, odotusarvo, todennäköisyys tai keskihajonta ovat yleensä tilastotieteellisen analyysin mielenkiinnon kohteena. Populaatiosuureita ei tunneta, mutta otannan avulla ne voidaan estimoida. Intuitiivinen (ei ainoa) tapa estimoida niitä on laskea vastaava *otossuure* eli *otostun-*

nusluku (*sample statistic*). Alla oletetaan, että on käytettävissä riippumattomat havainnot $x_1, \ldots, x_n$. Harjoitustehtävissä opastetaan, miten jakson suureita lasketaan R:llä.

Otospistetodennäköisyysfunktio (*empirical probability mass function*)

$$\hat{f}(x_i) = \frac{1}{n}(\text{lukumäärä havaintoja, joiden arvo on } x_i), \quad i = 1, \ldots, m,$$

on diskreetin satunnaismuuttujan pistetodennäköisfunktion arvon pisteessä x_i otosvastine. Siinä m on satunnaismuuttujan mahdollisten arvojen lukumäärä.

Jos satunnaismuuttuja on välimatka-asteikollinen ja voi saada lukuisia arvoja, otoskoko ei yleensä ole niin suuri, että otospistetodennäköisyydet muodostaisivat säännöllisen todennäköisyysjakauman kuvion. Tällöin monesti piirretään *histogrammi* (kuvat 8.1, 8.2 ja 8.3): Jaetaan satunnaismuuttujan arvojen vaihteluväli tasavälisiin luokkiin, lasketaan havaintojen lukumäärät luokissa ja piirretään lukumääristä pylväskuvio, jossa pylväät sivuavat toisiaan. Karkea luokittelunmukaisen pistetodennäköisyysfunktion estimaatti saadaan jakamalla havaintojen lukumäärät luokissa nh:lla, jossa n on havaintojen lukumäärä ja h luokkien leveys. Luokiteltu aineisto tapaa muodostaa alkuperäistä säännöllisemmän kuvion otospistetodennäköisyyksistä. Samalla tavoin saadaan karkea tiheysfunktion estimaatti, jos satunnaismuuttuja on jatkuva-arvoinen. Luokkakoon voi valita silmämääräisesti kokeilemalla.[79]

Luokkakoon valintaan on selkeitä sääntöjä ja kaavoja (esim. Keen 2018, 183–184). R:n `hist`-käsky soveltaa yhtä tällaista kaavaa oletusarvoisesti.

Otoskertymäfunktio (*empirical cumulative distribution function*)

$$\hat{F}(x) = \frac{\text{lukumäärä havaintoja } x_i, \text{ jotka ovat korkeintaan yhtäsuuria kuin } x}{n}$$

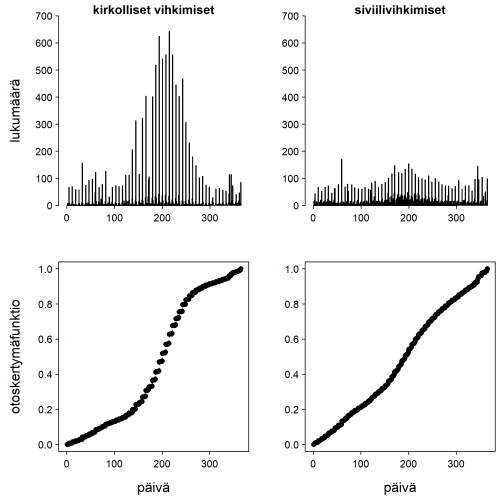
estimoi kertymäfunktion arvoa pisteessä x. Järjestetään havainnot suuruusjärjestykseen $x_{(1)}, \ldots, x_{(n)}$, jossa $x_{(i)}$ on otoksen suuruusjärjestyksessä i. havainto $(i = 1, \ldots, n)$. Muotoilu

$$\hat{F}(x) = \begin{cases} 0, & \text{jos } x < x_{(1)}, \\ i/n, & \text{jos } x_{(i)} \leq x < x_{(i+1)}, \ i = 1, \ldots, n-1 \text{ ja} \\ 1, & \text{jos } x \geq x_{(n)}, \end{cases}$$

korostaa, että otoskertymäfunktio saa arvoja välillä $[0, 1]$ ja että se on porrasfunktio. Kunkin portaan korkeus on $1/n$, paitsi jos aineistossa on *tasahavaintoja*

(*ties*) eli havaintoja, joiden numeerinen arvo on sama. Tällöin portaan korkeus on t/n, jossa t on tasahavaintojen lukumäärä. Esimerkit alla ja jaksossa 15.2 (kuva 15.2) havainnollistavat.

Esimerkki. Vihkimiset 2013 I.[80] Kuvassa 8.1 on kirkollisten vihkimisten ($n = 12\,410$) ja siviilivihkimisten ($n = 8\,878$) päivittäiset lukumäärät ja otoskertymäfunktiot. Histogrammissa on jokainen päivä eroteltu omana luokkanaan. Havaintoja on niin paljon, että kertymäfunktioiden porrasmaisuus ei pistä kuvista silmään. Kirkollisia vihkimisiä on erityisesti kesällä; siviilivihkimiset jakaantuvat tasaisemmin. Kirkollisten vihkimisten otoskertymäfunktion kuvaaja jyrkkenee kesäpäivien aikaan. Myös siviilivihkimisten otoskertymäfunktion kuvaaja reagoi kesäpäiviin mutta loivemmin. □



Kuva 8.1: Kirkollisten vihkimisten ja siviilivihkimisten 2013 päivittäiset (1–365) lukumäärät (yllä) ja otoskertymäfunktiot (alla).

Otoskvantiili (*sample quantile*) lasketaan painotettuna keskiarvona kahdesta suuruusjärjestyksessä peräkkäisestä havainnoista (vrt. 0.5. otoskvantiilin eli otosmediaanin kaava (8.1) alla). Idea on laskea jatkuva-arvoinen otoskvantiili kaavan (6.1) tapaan, vaikka kertymäfunktio on porrasmainen. Monimutkaisen ei-kovin-informatiivisen kaavan sijaan tässä esitetään yksinkertainen R-komento, jolla q.

otoskvantiili lasketaan: `quantile(data,q)` (`data` on R:ään luettu dataksi nimetty aineisto satunnaismuuttujan toteumista). Harjoitustehtävässä palataan otoskvantiilin laskutapoihin.

Odotusarvon luonteva estimaatti on *otoskeskiarvo* (*sample mean*)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Mikäli x_i saa vain arvoja 0 tai 1, niin $\bar{x}$ on välillä $[0, 1]$ ja on usein estimaatti todennäköisyydelle.

Otosmediaani (*sample median*) on

$$m = \begin{cases} x_{(k+1)}, & \text{jos } n = 2k+1, \\ \dfrac{x_{(k)} + x_{(k+1)}}{2}, & \text{jos } n = 2k. \end{cases} \tag{8.1}$$

Yllä $x_{(i)}$ on otoksen suuruusjärjestyksessä *i.* havainto ($i = 1, \ldots, n$ ja k on positiivinen kokonaisluku). Otosmediaani on järjestetyn aineiston keskimmäinen havainto, jos havaintoja on pariton määrä. Muulloin otosmediaani on keskimmäisten havaintojen keskiarvo.

Otoksen tyyppiarvo (*sample mode*) on aineistossa useimmin esiintyvä havainto. Tyyppiarvoja on useita, jos esimerkiksi kahta arvoa on eniten ja yhtä monta aineistossa. Jos aineisto on otos jatkuva-arvoisesta satunnaismuuttujasta, havainnot ovat väistämättä luokiteltuja mittaustarkkuuden puitteissa mutta silti usein yksikään arvo ei ole muita yleisempi. Tällöin aineisto voidaan luokitella karkeammin ja nimetä tyyppiarvoksi luokka (luokkakeskus), johon osuu eniten havaintoja. Erilaisilla luokitteluilla saatetaan päätyä aivan erilaisiin tyyppiarvoihin.

Esimerkki. Opiskelijoiden ansiot. Aineisto (kuvitteellinen) koostuu kymmenen opiskelijan ansioista euroissa vuoden aikana: 0, 0, 0, 0, 5 000, 6 000, 7 000, 8 000, 9 000, 10 000 ja 50 000. Ansioiden tyyppiarvo, mediaani ja keskiarvo ovat 0, 6 000 ja 8 636.36 euroa. Keskiluvut ovat hyvin erilaisia mutta kuvaavat omalla tavallaan osuvasti aineistoa. □

Paljon käytetty kaava *otosvarianssille* on

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Otosvarianssi on sitä suurempi, mitä enemmän havainnot x_i poikkeavat havaintojen keskiarvosta $\bar{x}$. Vastaava keskihajonnan estimaatti, *otoskeskihajonta*, on

otosvarianssin neliöjuuri

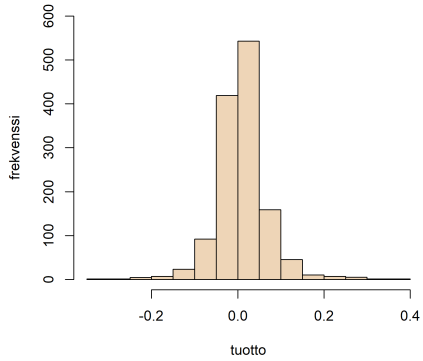
$$s = \sqrt{s^2}.$$

Otosvinous saadaan samaan tapaan:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}.$$

Havaintojen x_i suuret positiiviset poikkeamat keskiarvosta ($x_i - \bar{x} > 0$) kasvattavat vinoutta, mutta vastakkaismerkkiset poikkeamat pienentävät sitä. Poikkeamat kumoavat toisensa ja otosvinous on nolla, jos havainnot jakautuvat täysin symmetrisesti keskiarvon ympärille. Kuten populaatiovastineensa (jakso 6.4), otosvinous voi olla 0, vaikka jakauma ei olisi symmetrinen.

Esimerkki. Osakkeiden tuotto. Kuvassa 8.2 on histogrammi suomalaisten osakkeiden kuukausituotosta (kurssi-indeksin logaritmin kuukausimuutoksesta) 31.10.1912–31.8.2022.[81] R:n `hist`-komento on automaattisesti ryhmitellyt tuotot 0.05:n levyisiin pylväisiin. Jakauma on keskittynyt karkeasti nollan ympärille ja on lievästi vino oikealle. Tuoton tyyppiarvo, mediaani ja keskiarvo ovat 0.000, 0.008 ja 0.010. Keskihajonta ja vinous ovat 0.058 ja 0.308. Aineistoa tutkitaan lisää jaksossa 15.1.1 ja harjoitustehtävissä. □



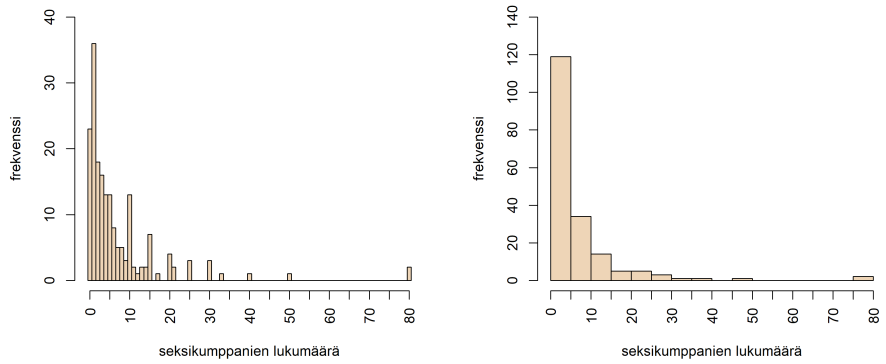
Kuva 8.2: Suomalaisten osakkeiden tuoton 31.10.1912–31.8.2022 jakauma.

Esimerkki. Seksikumppanien lukumäärä. Kuva 8.3 havainnollistaa seksikumppanien lukumäärän otosjakaumaa. Aineisto on 189:n 18–24-vuotiaan naisen vas-

taukset vuoden 2007 Finsex-tutkimuksen kysymykseen numero 74: “Kuinka monen henkilön kanssa olette ollut sukupuoliyhteydessä elämänne aikana?”[82]

Kuvan ensimmäisessä osassa on piirretty pylväs vastausvaihtoehtojen $(0, 1, 2, \ldots)$ kohdalle, johon on tullut vastauksia. Pylvään korkeus on vastausvaihtoehdon ilmoittaneiden vastaajien lukumäärä. Jakauma on hyvin vino oikealle. Se heijastuu tyyppiarvon, mediaanin ja keskiarvon suuruusjärjestyksessä 1 (36 vastausta), 3 ja 6.9. Mediaani 3 kuvannee parhaiten tyypillistä lukumäärää tässä aineistossa. Vinous on 4.0, ja keskihajonta on 11.0.

Kuvan toisessa osassa on vastaava kuvio aineistosta, jossa vastaukset on ryhmitelty luokkiin $0-5, 6-10, \ldots, 75-80)$ ja pylväät peittävät luokkarajat. Ryhmittely tuottaa selkeämmän ja säännöllisemmin käyttäytyvän kuvion. Toisaalta ryhmittelyssä häviää tieto muista poikkeavan 0-luokan koosta. □



Kuva 8.3: Alle 25-vuotiaiden naisten seksikumppanien lukumäärä.

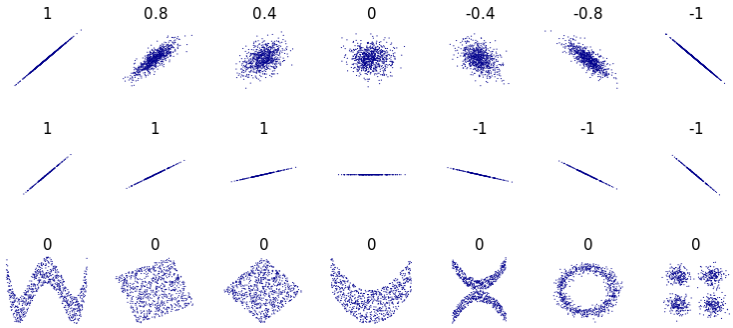
Korrelaatio voidaan estimoida *otoskorrelaatiokertoimella*

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{8.2}$$

Sen lasku edellyttää, että on käytössä havaintoparit $(x_1, y_1), \ldots, (x_n, y_n)$. Korrelaatiokertoimen (jakso 6.5) tapaan otoskorrelaatiokerroin rajoittuu välille $[0, 1]$.

Vakion lisääminen muuttujaan tai sen kertominen vakiolla ei muuta otoskorrelaatiokertoimen arvoa (harjoitustehtävä).

Kahden muuttujan havaintopareista muodostettua kuviota kutsutaan *sirontakuvioksi* (*scatter plot*). Sirontakuviota on sanottu kaikista monikäyttöisimmäksi ja hyödyllisimmäksi visualisoinniksi (Friendly ja Wainer 2021, 8). Kuvan 8.4 korrelaatioiden arvoja havainnollistavat sirontakuviot tukevat näkemystä. Ylimmällä rivillä korrelaatio muuttuu 1:stä −1:hteen. Keskimmäinen rivi osoittaa, että muuttujien x ja y välistä suhdetta kuvaavan suoran kulmakerroin ei riipu yksin korrelaation suuruudesta (vrt. yhtälö (13.5)). (Keskimmäisessä tilanteessa korrelaatiokerrointa ei ole määritelty. Pohdi, miksi!) Kolmannella rivillä korrelaatio on nolla, vaikka muuttujat eivät selvästikään ole riippumattomia.



Kuva 8.4: Otoskorrelaatioiden havainnollistuksia.[1]

Jakson estimaatteja vastaavat estimaattorit saadaan korvaamalla kaavoissa x_i- ja y_i-havainnot X_i- ja Y_i-satunnaismuuttujilla (ml. keskiarvojen kaavoissa). Monesti etuliite "otos-" tiputetaan ja puhutaan keskiarvosta, mediaanista, korrelaatiosta jne., vaikka tarkoitetaan otosvastinetta. Niin tehdään ajoittain tässäkin tekstissä. Korrelaatiokerrointa (8.2) kutsutaan joskus Pearsonin korrelaatiokertoimeksi keksijänsä Karl Pearsonin mukaan.

[1]Kuva: Denis Boigelot. Lähde: Wikimedia Commons (`https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg`; haettu 2.3.2020).

8.3 Todennäköisyys- eli satunnaisotanta

Otannan pitää olla *todennäköisyysotantaa* (*probability sampling*) eli *satunnaisotantaa* (*random sampling*), jotta otoksesta voidaan tehdä tilastotieteellisiä päätelmiä. Todennäköisyysotannassa populaation kunkin alkion todennäköisyys tulla otokseen eli *sisältymistodennäköisyys* (*inclusion probability*) tunnetaan, ja se on nollaa suurempi kaikille alkioille. Tällöin otantavirheen suuruutta voidaan arvioida tilastotieteellisesti. Satunnaisotannalla kerättyä otosta kutsutaan *satunnaisotokseksi* (*random sample*).

Todennäköisyysotanta on pelkistetyimmillään *yksinkertaista satunnaisotantaa* (*simple random sampling*). Siinä kaikkien populaation alkioiden sisältymistodennäköisyys on sama:

$$\mathsf{P}(\text{populaation tietty alkio tulee otokseen}) = \frac{n}{N}.$$

Yllä n on otoksen ja N on populaation koko.

Yksinkertaisella satunnaisotannalla saatua otosta kuvataan joskus harhattomaksi. Sillä tarkoitetaan sitä, että poimittuun otokseen tulee keskimäärin populaationkaltainen joukko alkioita. Toisinaan puhutaan epäselvemmin ja vanhahtavasti "edustavasta otoksesta". Yksittäinen pieni satunnaisotos saattaa kuitenkin olla koostumukseltaan poikkeuksellinen. Se on yksi syy pyrkiä riittävän suuriin otoksiin. Aina se ei ole mahdollista, tai muusta syystä yksinkertainen satunnaisotanta ei ole erityisen sopiva otantamenetelmä. Toisaalta "edustava otos" ei ole välttämättä tavoiteltava. Otos, jossa kutakin väestöryhmää on sama osuus kuin väestössä, ei ole aina optimaalinen. (Tillé 2020, 70–71.)

Esimerkki. Ositettu otanta I. Ositetussa otannassa populaatio jaetaan homogeenisiin ryhmiin, joista kustakin poimitaan otos yksinkertaisella satunnaisotannalla. Näin voidaan taata, että otoksessa ovat kaikki ryhmät halutussa suhteessa edustettuina. Haluttu suhde voi olla ryhmien osuudet populaatiossa, jolloin otos edustaa mahdollisimman hyvin populaatiota ryhmäkoostumukseltaan. □

Esimerkki. Ositettu otanta II. Ositetulla otannalla voidaan hakea populaation osajoukoista tietoisesti epäsuhtainen määrä havaintoja. Pienen osajoukon alkioille asetetaan muita suurempi todennäköisyys tulla otokseen. Näin siihen saadaan pieneen osajoukkoon kuuluvia riittävästi osajoukkojen vertailua varten. Yksinkertaisella satunnaisotannalla koko populaatiosta ei välttämättä saataisi. □

Jakaumia opiskeltaessa tutustuttiin kahteen otantamenetelmään: Otanta takaisinpanolla (binomijakauma) ja ilman (hypergeometrinen jakauma). Molemmis-

sa sovellettiin yksinkertaista satunnaisotantaa. Otanta ilman takaisinpanoa on parempi otantamenetelmä kuin otanta takaisinpanolla. Jälkimmäisessä populaation alkio voi tulla useampaan kertaan eli yliedustetuksi otokseen.

Esimerkki. Kortin peluu (jatkoa). Yritetään päätellä kuninkaiden osuus korttipakassa poimimalla 13 korttia pakasta ilman takaisinpanoa ja takaisinpanolla. Jälkimmäisellä menetelmällä on pieni todennäköisyys saada otos, jossa on $5-13$ kuningasta. Sellainen otos johtaisi täysin väärään päätelmään kuninkaiden osuudesta pakassa. Otannassa ilman takaisinpanoa ei tätä virhemahdollisuutta ole. □

Monesti populaatio on siinä määrin suuri, ettei ole merkitystä, käytetäänkö otantaa takaisinpanolla vai ilman (jakso 7.4.1) tai oletetaanko populaatio äärettömän kokoiseksi vai ei. Laskut voidaan tällöin tehdä helpommin olettaen otanta takaisinpanolla äärettömästä populaatiosta, vaikkei siitä olisi kysymys. Seuraus yksinkertaistavista oletuksista lienee yleensä huomaamaton kupru päättelyssä (pieni luottamusvälin leveneminen tai testin heikkeneminen; jakso 11.3).

Yksinkertaisen satunnaisotannan toteuttamiseksi tarvitaan *otoskehikko* (*sampling frame*) eli listaus populaation alkioista. Paljon käytetty menettely on, että alkiot numeroidaan 1:stä N:ään, arvotaan satunnaislukuja vastaavasta diskreetistä tasaisesta jakaumasta ($\mathsf{P}(i) = 1/N$, $i = 1, \ldots, N$) ja otokseen poimitaan arvonnan tuloksen mukaiset alkiot.

Esimerkki. Kymmenen alkion yksinkertainen satunnaisotanta. Otoskehikko koostuu 10 000 alkiosta. Kuhunkin liitetään numero $1, \ldots, 10\,000$. Arvotaan kymmenen lukua vastaavasta diskreetistä tasaisesta jakaumasta.

```
set.seed(17032018)      # asetetaan siemenluku (ei välttämätön)
x <- sample(1:10000,10) # poimitaan 10 lukua 10 000:sta (ei takaisinpanoa)
x                       # katsotaan poimittuja lukuja
y <- sort(x)            # järjestetään poimitut luvut
y                       # katsotaan niitä
## [1] 991 1815 2184 2758 4352 6049 6533 8605 8890 9287
```

Koodi järjestää ja tulostaa arvotut luvut. Otokseen poimitaan otoskehikosta 991., 1815., ... ja 9287. alkio. Käsky `sample(1:10000,10, replace=T)` toteuttaisi otannan takaisinpanolla. □

Monissa sovelluksissa populaatio on teoreettinen eikä otoskehikkoa ole. Tällöin usein oletetaan, että — vaikkapa viranomaisen tilastoima — aineisto on satunnaisotos teoreettisesta populaatiosta, ja tilastollista päättelyä sovelletaan tavalliseen tapaan. Oletuksen paikkansapitävyyttä on usein vaikea arvioida luotettavasti.

Esimerkki. Rikokset. Rikoksista ei ole otoskehikkoa. Tilastokeskus julkaisee tilastoa poliisin tietoon tulleista rikoksista. Jos tilastoitu rikollisuus on satunnaisotos rikoksista, tilastollinen päättely tavalliseen tapaan on mahdollista rikoksista. □

Jos otantamenetelmä ei ole tutkijan kontrollissa, otoksen satunnaisuutta on vaikea varmistaa. Jos populaation jotkin ominaisuudet tunnetaan, otoksen koostumusta voi verrata populaatioon. Yhteneväisyys on suotavaa muttei takaa, että otos olisi tutkittavan muuttujan suhteen satunnaisesti saatu.

Satunnaisotos on keskeinen teoreettinen oletus lukemattomien tilastotieteellisten menetelmien taustalla. Silti satunnaisotosta on kuvattu ideaaliksi, jollainen saadaan käytännössä harvoin (Hoerl ja Snee 2016). Kuvaus korostaa tarvetta huolellisuuteen otoksen poimimisessa. Jatkossa otos oletetaan kerätyksi yksinkertaisella satunnaisotannalla — paitsi ei välttämättä seuraavassa jaksossa.

8.4 Epäaito otanta, näyte, valikoitumisharha ja muita pulmia

Satunnaisotanta ei ole aina mahdollista tai on vaikeata tai kallista. Otos populaatiosta voi silti olla saatavissa. Otanta on tällöin *epäaitoa* (*non-probability sampling*). Joidenkin populaation alkioiden todennäköisyys tulla otokseen on nolla tai alkioiden sisältymistodennäköisyyksiä ei tiedetä. Näin hankitusta otoksesta on vaikea tehdä tilastotieteellisiä johtopäätöksiä populaatiosta tai arvioida niiden luotettavuutta. Epäaidolla otannalla saatua otosta kutsutaan *näytteeksi*. Joillain tieteenaloilla ei ole harvinaista, että tilastollista päättelyä yritetään tehdä tavanomaiseen tapaan, vaikka aineistona on näyte. Menettely on epäluotettava ja helposti harhaanjohtava.

Otoksessa voi olla *valikoitumisharhaa* (*selection bias*). Se saattaa syntyä monella tapaa:

- Mukavuusotos (*convenience sample*, *opportunity sample*), jossa otokseen tulee mukaan vain helposti saatavilla olevia populaation alkioita.
- Otos valikoituu piilevän tai eksplisiittisen käytännön, ominaisuuden, taipumuksen, mieltymyksen tai muun vastaavan takia. Harkintaotannassa

[2]Kuva: Sketchplanations (`https://sketchplanations.com/sampling-bias`). Kiitän Jono Heytä kuvan painoluvasta ja englanninkielisen tekstin korvaamisesta suomenkielisellä.

Kuva 8.5: Vastauskatoharha.[2]

(*purposive sampling*, *judgement sampling*) aineiston kerääjä valitsee otokseen mielestään edustavat havainnot. Se ei takaa edustavuutta. Kerääjä ei välttämättä hahmota otantaa vääristävää tekijää. Lumipallo-otannassa (*snowball sampling*) otokseen haalitaan ihmisiä tuttujen ja heidän tuttujensa kautta. Hyvin verkostoituneet ihmiset tulevat otokseen muita todennäköisemmin, ja tutut tapaavat jakaa samankaltaisia piirteitä, jotka korostuvat otoksessa.

- Tutkija tai hänen apulaisensa vääristee otosta tarkoitushakuisesti hyödyntämällä edellä lueteltuja seikkoja, poistamalla tai lisäämällä havaintoja tai raportoimalla tulokset virheellisesti. Tutkija voi myös analysoida osa-aineistoa ymmärtämättä, että se vääristää tuloksia.[83]

- Julkaisuharha (*publication bias*) (Sterling 1959, Sterling ym. 1995). Tutkijat lähettävät tieteellisiin lehtiin mielellään yllättäviä tai muuten huomion-

arvoisia yhteyksiä ja tuloksia, ja tieteelliset lehdet julkaisevat sellaisia mieluusti. Odotetut tai muuten tylsemmänoloiset tulokset voivat jäädä makoilemaan tutkijan pöydälle tai karsiutua lehtien julkaistavaksi valittavien tutkimusten seulassa. Lopputulos on, että erikoisuudet ovat yliedustettuina lehdissä ja että osa julkaistuista tutkimuksista perustuu sattumalta poikkeavaan aineistoon, eikä väitetty erikoinen löytö päde todellisuudessa. Ongelma tuli esille "Useimmat tutkimustulokset vääriä" -esimerkissä (s. 68).

Kyselyihin liittyy erityisiä ongelmia edellisten lisäksi:

- Vastauskatoharha (*nonresponse bias*) voi syntyä, jos kaikki otokseen valitut eivät vastaa kyselyyn. Tulokset vääristyvät, jos tietynlaiset ihmiset jättävät vastaamatta tai kysely ei tavoita heitä. Mitä suurempi vastaamattomien osuus, sitä suurempi on vastauskatoharhan riski.
- Ihmiset hakeutuvat nettilinkin, lehti-ilmoituksen tai muun avoimen kutsun perusteella otokseen. Tällaisen itsevalikoituneen (*self-selective sampling*, *volunteer sampling*) otoksen on syytä olettaa olevan poikkeuksellinen. Tutkimukseen tai kyselyyn hakeutuva ihmistyyppi löytää kutsun muita todennäköisemmin tai pitää tutkimukseen osallistumista erityisen tärkeänä.
- Vastausharhaa (*response bias*), kehysvaikutusta (*framing effect*) tai kontekstivaikutusta (*context effect*) syntyy, jos haastattelijan käyttäytyminen, haastattelun tekotapa, kysymysten muotoilu tai järjestys tai muut psykologiset tekijät vaikuttavat vastaukseen. Tällaiset vaikutukset voivat olla suuria.[84] Haastattelijaharhaa (*interviewer bias*) on monenlaista. Haastateltava saattaa pyrkiä vastaamaan tavalla, jonka kokee yleisesti hyväksytyksi (*courtesy bias*). Jos kysymykset ovat arkaluonteisia, vastaukset saattavat olla erilaisia, jos haastattelu tehdään puhelimitse tai tavalla, jossa vastaukset eivät paljastu haastattelijalle. Jos kyselyllä testataan haastateltujen tietämystä, haastattelija saattaa johdattaa haastateltavaa oikeaan vastaukseen.[85] Itsekeskeisyysharhan (*egocentric bias*) johdosta haastateltava voi hahmottaa oman toimintansa liian positiivisessa valossa tai osapuolet vastata ristiriitaisesti samasta asiasta.
- Haastateltava voi vastata väärin tietämättömyyttään tai tietoisesti. Motiivi tietoisesti väärin vastaamiselle voi olla pilailu tai halu vaikuttaa tutkimuksen tuloksiin.

Esimerkki.[86] Merituulet. Francis Galtonin huomio 1866 on varhainen esimerkki mukavuusotannasta. Purjelaivoilta kerättiin tietoja merituulista. Laivat olivat

vähemmän aikaa alueilla, joilla oli suotuisia tuulia ja enemmän aikaa alueilla, joilla oli tyyntä tai vastatuulta. Laivoilta saadut tiedot tuulista painottuivat siksi jälkimmäisiin. □

Esimerkki.[87] Juutin laatu paalissa. Juuttipaaleja laivattiin Bombaystä Eurooppaan 1930-luvulla. Juutin laatua ja arvoa tarkkailtiin survaisemalla paalin kyljestä pyöreäreunainen terävä holkki. Sen kärjellä saatiin näyte paalissa olevasta juutista. Juutti pyrki tiivistymään paalin keskellä, ja holkki pysähtyi tyypillisesti paalin vahingoittumiselle aralle ulkokehälle. Tulos oli säännöllisesti todellista laatua huonompi arvio paalissa olevan juutin laadusta ja arvosta. □

Esimerkki.[88] Lintujen laulu. Charles Darwin esitti 1859 *Lajien synty* -teoksessaan, että naaraslinnut pyrkivät valitsemaan puolisokseen näyttävimmän ja kauniimmin laulavan koiraslinnun. Koiraslintujen laulusta tuli oppikirjaesimerkki sukupuolivalinnasta ja sen voimasta. Käsitys oli, että lintulaji, jossa myös naaras laulaa, on poikkeus. Kului noin puolitoista vuosisataa kunnes lintutieteilijät havahtuivat huomaamaan, että useimpien lintulajien naaraat laulavat. Esimerkiksi Odomin ym:iden (2014) aineistossa naaras laulaa 71 %:ssa lintulajeista.

Selitys pitkään vallalla olleelle väärälle käsitykselle on valikoitunut otos — mahdollisesti mukavuusotos. Darwinin tutkimilla Galapagos-saarten lintulajeilla yksin koiraat lauloivat. Lintutieteilijät perehtyivät sittemmin lintujen lauluun Pohjois-Amerikassa ja Euroopassa — joissa tyypillisesti vain koiraslinnut laulavat. Vasta kun alettiin tutkia lintujen laulua muualla ilmeni, että useimpien lintulajien naaraat laulavat. Rajoitetun alueen tutkimuksista oltiin päätelty virheellisesti, että kaikkialla pääasiallisesti vain koiraslinnut laulavat.

Väärää käsitystä voi tulkita seuraukseksi valikoituneesta otoksesta toisellakin tavalla. Hainesin ym:iden (2020) mukaan erityisesti naistutkijat huomasivat naaraslintujen laulavan. Lintutieteilijöiden tutkimuskohteet olivat mahdollisesti olleet tutkijoiden sukupuolen mukaan valikoituneita. □

Esimerkki. Rikokset (jatkoa). Poliisille ilmoitetaan tai poliisi saa muuten tietoa eri rikoslajeista mahdollisesti eri todennäköisyyksillä. Poliisin tilastoima rikollisuus on tällöin valikoitunut otos. Myös todennäköisyys raportoida samasta rikosnimikkeestä voi muuttua ajassa:[89]

> Poliisi saa yhä enemmän ilmoituksia seksuaalirikoksista, mutta nuoriin ja aikuisiin kohdistuva seksuaalinen väkivalta ei ole viime vuosina yleistynyt. – – Kansallisen rikosuhritutkimuksen tulokset on julkaistu loppuvuodesta Helsingin yliopiston Kriminologian ja oikeuspolitiikan instituutin (Krimo) katsauksessa. – – Vaikka koetun seksuaalisen väkivallan määrä on pysynyt vakaana, ilmoitusten määrä on lisääntynyt voimakkaasti. – – Ilmoitusten lisääntymistä selittää tutkimusten mukaan ilmoitusalttius ja poliisitutkinnan tehostuminen. Raiskauksia ei

> siis välttämättä tapahdu enemmän, vaan niistä vain ilmoitetaan aiempaa useammin. Poliisihallituksen tilastot perustuvat poliisin rikosilmoitusjärjestelmään.

□

Esimerkki. Rikokset (jatkoa). Poliisilaitokset rekisteröivät tietoja rikoksista ja niiden tekopaikoista. Aineistoista voidaan estimoida todennäköisyyksiä rikoksille tietyillä alueilla ja ohjata partiointia niille rikollisten kiinnisaamiseksi tai rikosten ennaltaehkäisemiseksi. Ongelma on, että rekisteriaineisto voi olla valikoitunutta. Poliisi on saattanut alun perin toimia erityisesti tietyillä alueilla vaikkapa niihin liittyvien ennakkoluulojen takia, ja alueille on sen vuoksi paikannettu paljon rikoksia. Poliisin moderninkuuloinen toimintatapa voikin vahvistaa aiempia ennakkoluuloja tai vinoutuneita käytäntöjä. Toisaalta poliisi voi olla välttänyt tietyille alueille menoa, jos kaupunki on riittävän segregoitunut ja jengiytynyt. Tällöin estimoidut todennäköisyydet eivät nimenomaisesti ohjaisi partiointia alueille, joilla sitä mahdollisesti eniten tarvittaisiin. □

Esimerkki. Ihmepelastumiset.[90] Zachris Topelius matkasi kesällä 1836 yliopistosta kotiinsa Pohjanmaalle. Matkakumppanillansa (“A”) oli ollut mukanaan tulentekovälineet ja voimakasta englantilaista ruutia metsästystä varten.

> – – ajoimme – – kesäkuun 30 p:nä 1836, hiljaista ravia Kangasalan harjua, piippujamme poltellen. – – huomasin – – hienon savun tupruavan A—n vasemmalta sivulta. – – Nyt jo rohkenimme ottaa kukkaron käteemme ja huomasimme sen läpikyteneeksi, mutta sammuneeksi, siltä puolelta, missä sytytysvehkeitä säilytettiin. – – ryhdyimme ruutikääröä lähemmin tarkastamaan. Sen ympärillä oli kolme kerrosta tavallista karkeata paperia. Päällimmäinen kerros vasemmalta puolelta oli *mustaksi palanut* ja puoleksi hiiltynyt. Toinen kerros oli *ruskeaksi palanut*, ja *vaaleankeltainen* palopilkku sisimmässä paperissa osotti, että tuli luultavasti olisi tarvinnut vain muutamia sekunteja ruutiin yhtyäkseen. Miten kalliita sekunteja kolmelle ihmishengelle! Ellei niitä olisi ollut, niin olisi tuo voimakas, tiukasti väliimme ahdettu räjähdysaine lennättänyt rattaat, hevoset, matkustajat ja takalaudalla istuvan kyytipojan tuhansina sirpaleina pitkin Kangasalan harjun rinteitä. – – jos tämä oli sattuma, mitä on silloin kaitselmus?

Tarinoita viime hetken pelastumisista on monia. Topeliuksen kaitselmus-tulkinta on mahdollinen muttei välttämätön. Valikoitunut otos on sekulaari selitys. Vaaran huomaa ennen tai jälkeen kriittisen hetken. Ketkä eivät huomaa ennen, eivät ole tarinaansa kertomassa. □

Esimerkki.[91] Yhdysvaltojen presidentinvaali 1936. Klassinen esimerkki valikoitumisharhasta on *Literary Digest* -lehden kysely presidenttiehdokkaiden kannatuksesta 1936 Yhdysvalloissa. Lehti kokosi ilmeisesti suurimman otoksen koskaan 2.4 miljoonaa yhdysvaltalaista. He kertoivat, äänestävätkö demokraatti

Franklin Rooseveltia vai republikaani Alfred Landonia. Kyselyn mukaan Landonista tulisi presidentti 57 %:n kannatuksella. Ennuste epäonistui täydellisesti. Roosevelt voitti ylivoimaisesti 67 % ääniosuudella. Ero lehden ennustamaan kannatukseen oli $67 - 43 = 24$ %-yksikköä, mikä on tiettävästi suurin ennustevirhe koskaan suureen otokseen perustuneessa vaaliennusteessa.

Selitys väärälle ennusteelle oli, että lehti oli lähettänyt 10 miljoonaa tiedustelua muun muassa puhelinluetteloista, lehden omasta tilaajarekisteristä, auton omistajien rekistereistä ja klubien jäsenlistoista kerättyihin osoitteisiin. Vasta 11 miljoonassa kotitaloudessa oli tuolloin puhelin, ja myös auton omistaminen oli nykyistä harvinaisempaa. Lehden tavoittamat äänestäjät lienevät olleet keskimääräistä vauraampia. Heidän poliittiset näkemyksensä erosivat keskivertoäänestäjän ajatuksista. Vastaamatta jättäneiden suuri osuus (n. 75 %) saattoi myös selittää suurta otantavirhettä.

Vaali 1936 teki kuuluisaksi George Gallupin ja hänen 1935 perustamansa yhtiön. Gallup ennusti *Literary Digest* -lehden ennusteen jo ennen sen julkaisua sekä vaalien voittajan Rooseveltin oikein. Lehden ennusteen hän selvitti kysymällä 3 000 *Literary Digest* -lehden käyttämään äänestäjäluetteloon kuuluvalta, kuinka he aikovat äänestää. Vaalien voittajan Gallup selvitti toisella 50 000 äänestäjän erilailla kerätyllä otoksella.[92]

Vaalituloksen ennusti kyselytutkimuksilla oikein myös suomalaistaustainen Emil Hurja. Hän oli ennustanut oikein myös 1932 presidentinvaalin ja 1934 kongressivaalin tulokset. Hurja vakuutti Gallupille, että luotettavia ennusteita saisi pienemmilläkin otoksilla kuin Gallup keräsi. Hurjan otokset eivät olleet satunnaisotoksia, mutta hän osasi painottaa niitä otantaharhan pienentämiseksi. (Holli 2002, 65, 75, 119–121.) □

Esimerkki.[93] Lanarkshiren maitokoe. Kuuluisa harkintaotantaesimerkki on Lanarkshiren maitokoe Skotlannissa 1930. Tarkoitus oli tutkia, miten maito vaikuttaa lasten pituuteen ja painoon. Maitoa saavat (10 000) ja saamattomat (10 000) lapset arvottiin. Kouluja ohjeistettiin, että jos jompaankumpaan ryhmään tuli erityisen paljon hyvin- tai aliravittuja lapsia, niin heidät tulisi korvata, jotta otos olisi tasapainoisempi. Seuraus oli, että opettajat — ilmeisesti hyväsydämisyyttään tietoisesti tai tiedostamattaan — sijoittivat aliravittuja lapsia maitoa saavaan ryhmään. Maitoa juoneet lapset olivat tutkimuksessa lähes kaikissa ikäryhmissä lyhyempiä ja laihempia kuin maitoa saamattomat. Nuorimmat 5.5-vuotiaat maidolla ravitut tytöt olivat erityisen rimpuloita. Maidon pituuden ja painon kasvua edesauttavan vaikutuksen suuruuden estimointi vaikeutui suuresti. □

Esimerkki.[94] Viikonloppujen sairaalakuolemat. Morag Tolvi tutki väitöskirjas-

saan, kuinka paljon todennäköisemmin potilas kuolee sairaalassa, jos hän tulee hoitoon viikonloppuna eikä arkipäivänä. Mikäli todennäköisyyserosta johtuvat kuolemat katsottaisiin alimiehityksestä tai viikonloppusijaisista johtuviksi ja hoitotyön paremmalla järjestelyllä tai rahalla poistettaviksi, 3 701 ihmisen kuolema olisi vältetty HUSin pääsairaaloissa 2000–2013 (Tolvi 2020, 44).

Iso-Britannian terveysministeri Jeremy Hunt arvioi 16.7.2015, että 6 000 ihmistä kuolee vuosittain Iso-Britanniassa, koska sairaalat toimivat viikonloppuisin arkea pienemmällä henkilökunnalla. Ministeri Huntin mukaan sairaalaan sunnuntaina joutuva kuolee 15 %:a todennäköisemmin kuin keskiviikkona joutuva.

Ehkei näin moni kuitenkaan kuole sairaaloiden viikonloppumiehityksen takia. Sairaaloihin hakeutunee ja sairaalat ottanevat viikonloppuisin hoitoon vain välitöntä apua tarvitsevia. Heidän joukossaan kuolleisuus on sisääntulopäivästä riippumatta suurempaa kuin sairaalassa vähemmän kiireellistä apua saavien keskuudessa. Ajatusta tukee se, että HUSin potilaista 16.7 % on otettu sairaalaan lauantaina tai sunnuntaina. Se on vähemmän kuin näiden päivien osuus viikonpäivistä 28.6 %. (Tolvi 2020, 44, 53.) □

Esimerkki. Asesalva (*weapon salve*). Julkaisuharha ilmaantui jo 1600-luvulla. Jan Baptist van Helmont kertoi latinankielisessä kirjassaan 1621 haavoja parantavasta asesalvasta. Walter Charleton englansi kirjan 1649. Parantavasta asesalvasta kerrottiin myös 1658 ranskaksi ja englanniksi julkaistussa teoksessa. Yksi asesalvan resepteistä koostui karhun rasvasta, villisian ihrasta, muumiojauheesta (*powdered mummy*) ja pääkalloon kasvaneesta sammaleesta. Van Helmontin mielestä jesuiitan pääkallon sammal sopi salvaan erityisen hyvin. Salvalla voideltiin asetta, joka oli aiheuttanut haavan. Salvan kerrottiin toimivan magneetin tavoin etäältä haavaan koskettamatta. Salvan toimivuutta oli koeteltu kokeellisesti, ja arvostetut henkilöt olivat raportoineet sen parantavasta voimasta. Charleton ryhtyi sittemmin epäilemään asesalvaa ja arveli, että sen toimivuudesta oli raportoitu onnistumisia muttei epäonnistumisia. (Wootton 2015, 287, 291–293.) □

Esimerkki.[95] Kinsey ja Hite. Alfred Kinseyn ja Shere Hiten kyselytutkimukset ovat vaikuttaneet monien käsityksiin seksuaalisuudesta ja parisuhteesta ja ovat kuuluisimpia kyselytutkimuksia. Hite kuvaa kotisivullaan tutkimustensa olleen uutispommi ja sensaatio.

Kinseyn 1940-luvulla suurta huomiota herättäneitä tuloksia olivat, että 70 % miehistä on käyttänyt prostituoitujen palveluita ja että maatiloilla asuvista miehistä 17 %:lla on ollut seksuaalinen suhde eläimeen. Hite raportoi 1978-tutkimuksessaan, että 70 % yli viisi vuotta naimisista olleista naisista harrastaa

avioliiton ulkopuolisia suhteita. Osuus oli lähes sama kaikissa kuudessa Hiten tutkimassa etnisessä ryhmässä.

Kinseyn tulokset perustuivat 18 216 yksityiskohtaiseen haastatteluun, joista 11 246 muodosti "perusotoksen". Kinsey oli karsinut aineistosta kolmasosan, joka oli peräisin vankiloista ja homoseksuaalisesta yhteisöstä. Karsitussa otoksessa 84 % oli koulutettuja (*college-educated*), koska työväestön edustajat olivat alkuperäisessä otoksessa olleet järjestään vankilassa. Hite oli lähettänyt 100 000 kyselyä, joista 4 500 vastattiin (vastausprosentti 4.5).

ASA julkaisi raportin sekä artikkelin, joissa kritisoitiin Kinseyn tutkimusten uskottavuutta muun muassa siitä, että hänen aineistonsa ei ollut satunnaisotos. Kuuluisa tilastotieteilijä John Tukey opasti Kinseytä, että hän vaihtaisi hetimmiten Kinseyn keräämät 18 000 tapaushistoriaa 400 havainnon satunnaisotokseen, jos se olisi mahdollista.

Spiegelhalter (2015, 10) pitää mahdottomana, että avioliiton ulkopuolisia suhteita harrastavien vaimojen osuudet otoksessa olisivat luonnostaan niin samoja kuin Hite raportoi sen olevan etnisissä ryhmissä. Spiegelhalter arvioi Hiten aineiston luotettavuuden toiseksi alhaisimpaan luokkaan viisiportaisessa asteikossa. □

Esimerkki.[96] Terry Georgen ohjaama Christian Balen tähdittämä armenialaisten kansanmurhaa 1915 kuvaava *The Promise* esitettiin Toronton elokuvajuhlilla 11.9.2016. Elokuvan näki alle 3 000 ihmistä. *Internet Movie Database*'issä (IMDb) oli pian yli 50 000 arvostelua. Elokuva esitettiin seuraavaksi Virginian elokuvajuhlilla 5.11.2016. IMDb-arvosteluita oli 7.4.2017 jo yli 96 000. Arvosteluista noin 60 000 oli alinta arvosanaa 1, noin 35 000 oli korkeinta arvosanaa 10 ja vain noin 1 000 oli väliltä 2–9. Elokuva tuli kaupalliseen levitykseen 21.4.2017.

Moni on arvioinut elokuvan äärimmäisen huonoksi tai hyväksi näkemättä sitä. "1"-äänet juontavat sosiaalisen median kampanjaan Turkissa, jonka hallituksen mukaan kansanmurhaa ei ole tapahtunut. "10"-äänet ovat armenialaisilta, jotka haluavat saada elokuvalle katsojia. Useimmat äänestäjät ovat poliittiselta katsannoltaan valikoituneita, ja valtaosa äänistä on annettu poliittisesti tarkoitushakuisessa mielessä. Kuvat 8.6 ja 8.7 ovat yhdestä kansanmurhaa muistavista monista muistomerkeistä. □

Esimerkki.[97] Valikoitumisharha suomalaisessa tutkimuksessa I. Kesällä 2005 toteutettiin Suomessa rikosuhritutkimus soittamalla kiinteään puhelinverkkoon kuuluville kotitalouksille. Otokseen ei tullut henkilöitä, joilla oli vain matkapuhelin tai jotka olivat kesäilmoilla ulkona. Aromaan ja Heiskasen (2006) mu-

[3]Kuvat: Pekka Pere (2022).

Kuva 8.6: Albert Mkhitarianin (2002) muistomerkki armenialaisten kansanmurhalle 1915 ja armenialaisten osallistumiselle Ranskan sotiin 1914–1918 ja 1939–1945. Cannes, Ranska.[3]

Kuva 8.7: Muistomerkin yksityiskohta.[3]

kaan seuraus oli, että kerätty otos oli väestörakenteeltaan valikoitunut ja tulokset epäluotettavia. Otoksessa oli huomattavasti vähemmän uhreja kuin vuoden 2000 kyselyssä. Tulos saattaa johtua siitä, että nuoret ovat muita useammin väkivallan uhreja mutta heitä oli otoksessa tavanomaista vähemmän. □

Esimerkki.[98] Valikoitumisharha suomalaisessa tutkimuksessa II. Vuonna 2018 kysyttiin suomalaisilta, kuinka perhevapaiden käyttäminen vaikuttaa työuraan. Vastaajia oli toistakymmentätuhatta. Valtaosa äideistä ei kokenut viimeisen perhevapaan vaikuttaneen työuraan. Useampi äiti näki perhevapaan parantaneen asemaa työpaikalla kuin heikentäneen. Tutkija piti tuloksiaan yllättävinä. Helsingin Sanomien haastattelussa 5.12.2019 hän arveli, että raskauden myötä työttömäksi jääneet eivät ole tulleet huomioiduksi kyselyssä. □

Esimerkki. Kysymysten muotoilu I.[99] Mitä yhdysvaltalaiset ajattelevat ihmis-

kunnan synnystä? *Pew*-tutkimuskeskuksen mukaan kysymyksen muotoilulla on tässä yhteydessä hyvin suuri merkitys. Tutkimuskeskus on tavannut kysyä yhdysvaltalaisilta, onko ihmiskunta kehittynyt evoluutioteorian tapaan vai ollut aina olemassa Raamatun luomiskertomuksen tapaan. Tutkimuskeskuksen uudessa kysymyksessä on kolme vastausvaihtoehtoa: 1. Ihmiskunta on kehittynyt evoluution kaltaisten prosessien tuloksena, joissa Jumalalla ei ole roolia. 2. Ihmiskunta on kehittynyt prosessien tuloksena, jotka ovat Jumalan hyväksymiä tai ohjaamia. 3. Ihmiskunta on ollut olemassa ajan alusta alkaen. Tutkimuskeskus arpoi vastaajat vastaamaan joko vanhaan tai uuteen kysymykseen. Vanhan kysymyksen saaneista 31 % valitsi vastauksen, ihmiskunta on ollut aina olemassa. Uuden kysymyksen saaneista 18 % valitsi sen eli 3. vastauksen. Ero on suuri. Tutkimuslaitos arvelee, että vanhat vastausvaihtoehdot saattavat olla liian jyrkät. Uudet vaihtoehdot mahdollistavat vastauksen, joka sallii sekä evoluutioteorialle että uskonnolle merkityksen ihmiskunnan synnyssä. Tutkimuslaitos siirtyi käyttämään uutta kysymysmuotoilua. □

Esimerkki. Kysymysten muotoilu II.[100] Euroopan unionin tilastotoimisto Eurostat julkaisi 4.2.2019 tiedotteen odotettavissa olevista terveistä elinvuosista Euroopan unionin kansalaisilta. Asiaa oli selvitetty kyselyillä kansalaisille. Tulos oli, että ruotsalaiset voivat odottaa pisintä tervettä elämää (miehet 73.0 ja naiset 73.2 vuotta). Suomi oli yksi kolmesta maasta, jossa kyselyn mukaan miehillä on edessä enemmän terveitä elinvuosia kuin naisilla (miehillä 59.1; naisilla 57.0). Tiedotteessa todetaan, että tuloksiin saattaa vaikuttaa, kuinka terveitä elinvuosia on maissa mitattu.

Tutkimusprofessori Seppo Koskinen arvioi tiedotetta Helsingin Sanomissa 21.2.2019:

> Ei ole uskottavaa eikä muiden väestön terveyttä koskevien tietojen valossa mahdollista, että naisten elinajanodote on yhtä pitkä kuin Ruotsissa, mutta terveitä elinvuosia olisi 16 vähemmän – – .

Tilastokeskuksen yliaktuaari Kaisa-Mari Okkonen selittää:

> Suomi pärjää vertailussa heikosti: miehillä on terveitä elinvuosia (59.1) kahdeksanneksi vähiten ja naisilla (57) peräti kolmanneksi vähiten vertailumaista. – – Ruotsin lomakkeella aina vuoteen 2013 asti kysymys esitettiin ainakin pintapuolisesti samaan tapaan kuin Suomessa, yhdellä kysymyksellä. Vuonna 2014 kysymys kuitenkin pilkottiin Ruotsin tutkimuksessa neljään erilliseen kysymykseen. – – Uusi kysymistapa pitää sisällään periaatteessa kaikki samat elementit kuin alkuperäinenkin kysymys. Verrattuna aikaisempaan muotoiluun uusi kysymystapa kuitenkin kiinnittää vastaajan huomion jokaiseen toimintarajoitteen määritelmän osaan erikseen, vaatii harkitsemaan niitä ja sitä myötä nostaa kynnystä päätyä toimintarajoitteisten joukkoon. Tämä näkyy Ruotsin aikasarjatiedoissa selvästi: vuodesta 2014 eteenpäin vakavasti toimintarajoitteisten osuus väestöstä

> karkeasti ottaen puolittui, ja ero Suomen lukuihin kasvoi merkittävästi – – . – – Ruotsin tekemän kysymysmuutoksen myötä ei Suomen ja Ruotsin tietoja voi verrata keskenään – – . – – Ihmisten omat arviot terveydentilastaan ovat luotettavaa tietoa sinänsä, mutta miten kysymys- ja vastausvaihtoehdot muotoillaan, vaikuttavat vastaamiseen merkittävästi.

□

Esimerkki.[101] Kehysvaikutus. Vuonna 2016 hyväksytyn väitöskirjan mukaan 85 % suomalaisista hyväksyy eutanasian. Suhonen (2016) epäilee tulosta: Väittelijä oli kerännyt otoksensa sosiaalisen median kautta, jolloin otos lienee itsevalikoitunut. Eutanasian hyväksymiskysymystä oli edeltänyt pitkä "syväluotaava" kysymyssarja, joka loi eutanasiavastaukseen mahdollisesti vaikuttavan kehyksen. □

Esimerkki.[102] Tarkoittamattomat väärät vastaukset. Oppimistuloksia selvittävissä PISA-tutkimuksissa (*Programme for International Student Assessment*) koululaisilta kysytään heidän vanhempiensa koulutusta. Joidenkin tutkimusten mukaan perhetaustan merkitys koululaisten oppimistuloksissa olisi korostunut 2000-luvulla. Lehti ja Laaninen (2021) argumentoivat, että merkitys ei ole korostunut ja että vaikutelma korostumisesta johtuu suuresta mittausvirheestä: koululaiset eivät ole osanneet kertoa oikein vanhempiensa koulutusta. □

Esimerkki.[103] Vastausosuudet kyselytutkimuksissa. Yhteiskunta- ja käyttäytymistieteellisissä tutkimuksissa vastausosuudet ovat pienentyneet trendimäisesti. Suomalaisten seksuaalielämää kartoittavassa Finsex-tutkimushankkeessa vastausosuus on romahtanut 91.4 %:sta 36.0 %:iin runsaassa kolmessakymmenessä vuodessa:

vuosi	1971	1992	1999	2007	2015
vastaus-%	91.4	75.9	45.8	43.4	36.0

Yhdysvaltalaisen Gallup-yhtiön puhelinkyselyissä vastausosuus on kutistunut 28 %:sta 7 %:iin kahdessakymmenessä vuodessa:

vuosi	1997	2002	2007	2012	2017
vastaus-%	28	15	15	9	7

Nature Portfolio on maailman arvostetuimpia tieteellisiä kustantamoja, ja sen *Scientific Reports* on maailman viitatuimpia tieteellisiä lehtiä. Lehdessä julkaistu Derrickin ym:iden (2022) tutkimus perustui kyselyyn, jossa vastausosuus oli alle 1 %.

Pieni vastausosuus heikentää tulosten luotettavuutta. Vastaajat eivät ole tällöin välttämättä satunnainen otos. Suureen vastausosuuteen voidaan edelleen yltää. Halon ym:iden (2011) tutkimuksessa vastausprosentti oli 97. □

Katoanalyysilla voidaan yrittää selvittää, mikä vaikutus vastaamattomuudella on ollut otantatutkimuksen tuloksiin.

Esimerkki.[104] Katoanalyysi. Ylioppilaiden terveydenhoitosäätiö lähetti sähköisen terveyskyselyn opintonsa aloittaneille yliopisto-opiskelijoille. Opiskelijoista 8 750 eli 52 % vastasi. Vastaamatta jättäneistä poimittiin satunnaisotos. Sen opiskelijoille soitettiin (794 puhelua), kunnes saatiin tehtyä etukäteen päätetty määrä puhelinhaastatteluja 498 (vastausprosentti 63). Vastaamattomien terveydentila osoittautui paremmaksi. Näin ollen voisi ajatella, että ylioppilaiden terveysongelmia ei ole jäänyt piiloon alkuperäisessä tutkimuksessa. Silti on mahdollista, että kumpaankaan otokseen ei saatu vakavimmista terveysongelmista kärsiviä ylioppilaita. □

Esimerkki.[105] Vastaamattomuus- ja vastausharha suomalaisessa tutkimuksessa. Alkoholin suurkuluttajat vastaavat kulutuskyselyyn keskimääräistä epätodennäköisemmin, jolloin suurkuluttajien osuus pyrkii jäämään liian pieneksi. Vastaamattomuusharhaa voidaan korjata. Yhden Kopran (2018) harhakorjatun estimaatin mukaan suurkuluttajien osuus on 7.1 % harhakorjaamattoman estimaatin ollessa 4.8 %. Suurkuluttajia vaikuttaa olevan puolitoistakertainen määrä harhakorjaamattomalla menetelmällä arvioituun verrattuna. Kopran väitöskirjan (mt.) mukaan myös päivittäin tupakoivien osuus estimoituu liian pieneksi, jos vastaamattomuusharhaa ei huomioida.

Kaikki estimaatit edellä saattavat olla liian pieniä. Väitöskirjassa ei huomioitu vastausharhaa eli tässä yhteydessä sitä, että ihmiset usein ilmoittavat alkoholin- ja tupakankulutuksensa todellista pienemmiksi. □

Kyselyiden hupenevat vastausprosentit ovat tärkeä metodologinen ongelma. Se koskee myös monien maiden — Suomenkin — virallisia tilastoja, kun niitä tuotetaan kyselytutkimuksilla.

Valikoitumismekanismi tunnettaessa tilastollinen päättely saattaa olla mahdollista. Jos tietyn ryhmän tiedetään vastaavan kyselyihin huonosti — kuten vähän koulutetut ja miehet monissa yhteyksissä tapaavat tehdä — voidaan käyttää ositettua otantaa tai ryhmän vastauksia voidaan korostaa niitä painottamalla. Jos niukasti vastaavan ryhmän vastaajat ovat satunnaisotos ryhmästään, painotettu otos kuvaa populaatiota harhattomasti. Monimutkaisempiin tilanteisiin on olemassa perusopintotasoa vaativampia menetelmiä, joilla voidaan korjata otok-

sen valikoitumisen vaikutukset sopivissa tilanteissa. Joskus kekseliäällä otantamenetelmällä voi päästä samaan lopputulokseen.

Esimerkki.[106] Suomalaisten koulutustasoa arvioidaan Tilastokeskuksen työvoimatutkimuksella, joka on kyselytutkimus. Vastauskato on tutkimuksessa kasvanut 9 prosentista 37 prosenttiin vuosina 1995–2019. Vastaamattomat ovat muita useammin matalasti koulutettuja. Työvoimatutkimus uudistettiin 2021 korjaamaan vastauskatoharhaa. Aiemman työvoimatutkimuksen mukaan korkea-asteen tutkinnon oli suorittanut 47.9 prosenttia 25–64-vuotiaista suomalaisista vuonna 2020. Uuden työvoimatutkimuksen mukaan osuus oli 42.3 prosenttia vuonna 2021. Ero on 5.6 prosenttiyksikköä eli noin 150 000 tutkintoa. Aiemman tutkimuksen mukaan osuus oli Suomessa kymmenenneksi korkein OECD:ssä; uuden mukaan osuus on OECD-maiden keskitasoa. Käsitys suomalaisten koulutustasosta muuttui merkittävästi. □

Esimerkki.[107] Kekseliäs otanta. Missä määrin turvaistuin, kolmipisteturvavyö tai lantiovyö suojaavat lasta kuolemalta autokolarissa? Jones ja Ziebhart (2016) tutkivat sitä yhdysvaltalaisella aineistolla. Aineisto koostui alun perin autokolareista, joissa kuoli kuljettaja tai matkustaja. Aineisto oli valikoitunutta, sillä se kattoi kolarit, joissa kuoli lapsi muttei kolareita, joissa turvavälineet pelastivat lapsen. Tutkijat rajoittivat siksi — Levittin (2008) oivalluksen mukaisesti — aineiston kolareihin, joissa kuoli ihminen autossa B ja autossa A oli lapsi ja mallittivat todennäköisyyttä lapsen kuolemiselle autossa A. Näin rajoitettu aineisto ei liene valikoitunutta, sillä autossa A matkustaneen lapsen turvavälineet eivät vaikuttane siihen, kuoleeko joku vai ei kolarin toisena osapuolena olevassa autossa. Osa-aineistolla voidaan tutkia, kuinka hyvin eri turvavälineet suojaavat lasta kolarissa. □

Kyselyiden yhteydessä ilmoitettavat virhemarginaalit ovat yleensä oleellisesti luottamusvälejä (luku 10), jotka huomioivat satunnaisuuden vain yksinkertaisen satunnaisotannan kehikossa. Todellisuudessa kyselyihin liittyy muitakin epävarmuutta lisääviä tekijöitä, joiden takia luottamusvälit on monesti syytä olettaa huomattavasti raportoituja leveämmiksi.
Joskus on ilmeistä, että otos ei ole satunnainen (moni esimerkki edellä). Jos otos tiedetään valikoituneeksi mutta tavalla, jota ei tunneta, voi olla järkevintä pidättäytyä tilastollisesta päättelystä. Tavanomaiseen teoriaan tukeutuva päättely-yritys voi johtaa harhaan, kun oletus havaintojen valikoitumattomuudesta ei päde.

Esimerkki. Ei tilastollista päättelyä (Meeker ym. 2017, 393). Seurakunnan 730 jäseneltä kysyttiin kirjeitse, tekeekö heidän pappinsa työnsä hyvin (kyllä tai

ei). Vastauksia tuli 105, joista 58 oli kyllä-vastauksia. Tilastotieteen oppikirjan kirjoittajaa (mt.) pyydettiin arvioimaan kyllä-vastausten osuuteen liittyvää tilastollista epävarmuutta. Otos vaikutti vahvasti valikoituneelta: Äänestäneet ilmeisesti kokivat asian muita merkityksellisempänä, saattoivat olla kannustaneet samanmielisiä vastaamaan kyselyyn ja olivat mahdollisesti seurakunnassa keskimääräistä aktiivisempia. Kirjoittajat (mt.) arvioivat, että luottamusvälin (luku 10) lasku tällaisesta otoksesta ei ole järkevää. He suosittelevat, että äänestystulos julkistettaisiin sellaisenaan kera selityksen mahdollisesta valikoitumisesta ja kannustuksen jatkokyselyyn kyselyyn vastaamattomille. □

Lukuisat tutkimustulokset mediassa perustuvat otantaan. Kyky lukea niitä kriittisesti on tilastotieteellistä perusosaamista.

Lohr (2022) selittää oppikirjassaan, kuinka toimia vastauskadon tai epäaidon otannan tilanteissa.

8.5 Oudokit

Aineistossa on joskus muista havainnoista selvästi poikkeavia *oudokkeja* (*outliers*). Sellaisen voi tuottaa tilastollisen analyysin oletusten mukainen jakauma sattumalta, jolloin oudokki on luonteva osa otosta. Toisaalta oudokki voi olla päätynyt otokseen tarkoittamattomalla tavalla ollen vaikkapa virhekirjaus. Usein ei ole selvää, onko oudokki asianmukainen havainto vai ei. Oudokit voivat vaikuttaa suuresti — varsinkin pienissä otoksissa — tilastolliseen analyysiin. Jos oudokki on virheellinen havainto, se vääristää tuloksia ja pitää poistaa aineistosta.

Oudokkeja syntyy paksuhäntäisistä jakaumista. Sellaisessa tilanteessa saattaa olla järkevää käyttää toista tilastollista menetelmää kuin esimerkiksi normaalijakauman pätiessä (jakso 9.1). Monesti ei ole selvää, miten toimia oudokin kanssa. Päätös voi olla ratkaiseva tilastotieteellisen analyysin tulokselle. Yksi mahdollisuus on tehdä analyysit oudokin kanssa ja ilman ja raportoida molemmat tulokset. Oudokit muistuttavat aineiston keskeisyydestä tilastotieteessä.

Esimerkki. Opiskelijoiden ansiot (jatkoa jaksosta 8.2). Opiskelijoiden ansiot ovat 0, 0, 0, 0, 5 000, 6 000, 7 000, 8 000, 9 000, 10 000 ja 50 000 euroa. Viimeksi mainittu on poikkeuksellisen suuri ja oudokki. Ansio 50 000 euroa on suurempi kuin palkansaajan mediaaniansio 39 948 euroa vuonna 2018.[108] Luku 50 000 euroa saattaisi olla virhekirjaus ansion ollessa todellisuudessa 5 000 euroa. Jos 50 000

euroa on oikea ansio, herää kysymys, millainen opiskelija ansaitsee 50 000 euroa vuodessa. Ehkä kyseessä on työn ohessa sivutoimisesti opiskeleva korkeakoulutettu? Tällöin voisi olla syytä pohtia, onko otoksen kohdepopulaatio tarkoituksenmukainen. Pitäisikö kohdepopulaatioksi määritellä opintotukea saavat ensimmäistä tutkintoa täysipäiväisesti opiskelevat? Jos mielekäs kohderyhmä on kaikki opiskelijat — mukaan lukien työelämässä olevat — niin oudokin sisällyttäminen aineistoon on perusteltua. □

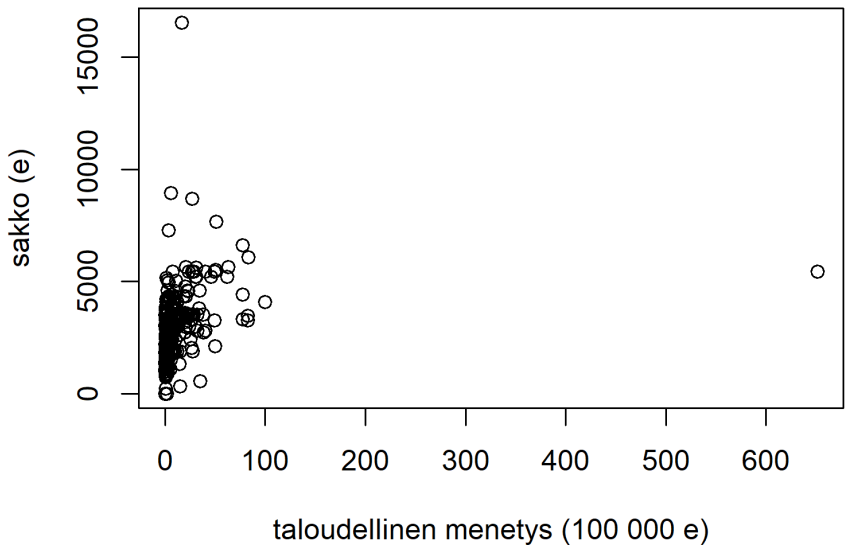
Esimerkki.[109] Kuvassa 8.5 on piirretty vastakkain työtuomioistuimen 2000–2009 määräämiä sakkoja laittomista työtaistelutoimenpiteistä ja niistä aiheutuneista taloudellisista menetyksistä ($n = 272$). Yksi työtaistelu on aiheuttanut yli 65 000 000 euron menetykset eli tavattomasti enemmän kuin muut työtaistelut. Tutkijat päätyivät poistamaan sen aineistosta ennen tilastollisia analyyseja. Sirontakuviossa erottuu myös työtaistelu, josta määrättiin yli 16 000 euron sakko. Tutkijat sisällyttivät sen tilastollisiin analyyseihinsa. □

Toinen empiirinen esimerkki on kuvassa 8.3 jaksossa 8.2. Useampi havainto poikkeaa muista täysin, mutta lukuarvot ovat mahdollisia. Ne tulisi ilmeisesti sisällyttää analyysiin. Kolmas empiirinen esimerkki oudokista on kuvassa 13.13 jaksossa 13.8. Oudokki ei ole virhekirjaus, sillä havaintovuosi voidaan todentaa poikkeukselliseksi.

8.6 Pintaremontti

Moneen tutkimukseen aineisto joudutaan vartavasten keräämään. Tällaista tutkimusta voi verrata pintaremonttiin tai maalaamiseen: Suurin osa urakasta on pohjatöiden tekemistä (otoksen keräämistä). Itse maalaaminen (analyysi) on pienempi ja monesti helpompi osa urakasta. Tämä kannattaa muistaa, jos kerää otoksen itse.

Pintaremonttivertaus toimii toisin, jos aineisto on valmiiksi kerätty: Edellytys tutkimuksen tai pintaremontin onnistumiselle on, että sen pohjatyöt ja ainekset ovat kunnossa.


sakko (e)
0
5000
10000
15000
0
100
200
300
400
500
600
taloudellinen menetys (100 000 e)

Luku 9

Piste-estimointi

> Laskennan päämäärä on näkemys, ei numerot.[110]
> Richard W. Hamming (1915–1998)

9.1 Hyvän estimaattorin ominaisuuksia

Olkoon parametri θ (jakso 6.2) ja sen estimaattori $\hat{\theta}$ (jakso 8.1). Estimaattorin *harha* (*bias*) on erotus

$$b(\theta) = \mathsf{E}(\hat{\theta}) - \theta.$$

Estimaattorin toivotaan monesti olevan *harhaton* (*unbiased*), jolloin

$$\mathsf{E}(\hat{\theta}) = \theta.$$

Tällöin riippumattomista otoksista laskettu $\hat{\theta}$ saa keskimäärin oikean arvon θ. Harhattomuus on intuitiivinen ja yleensä toivottava hyvän estimaattorin ominaisuus — muttei olennainen! On olemassa hyviä estimaattoreita, jotka eivät ole harhattomia eli ovat *harhaisia* (*biased*). On myös tilanteita, joissa harhattomuus ei ole hyvälle estimaattorille välttämätön tai ehkä edes suotava ominaisuus.

Esimerkki. Otoskorrelaatio $\hat{\rho}$ (kaava (8.2)) on tavattomasti käytetty hyvä mutta ylipäänsä harhainen estimaattori. Sen harhaa selvitetään jaksossa 9.8. □

Esimerkki. Parametrin θ tiedetään olevan välillä $[0, 1]$. Olkoon $\theta = 1$ ja θ:n estimaattorilla positiivinen varianssi kaikilla θ:n arvoilla. Jos θ:n estimaattori olisi harhaton, se saisi mahdollisesti usein 1:stä suurempia arvoja, jotka ovat mahdottomia. □

Oleellinen hyvän estimaattorin ominaisuus on, että se on *tarkentuva* (*consistent*): Havaintojen lukumäärän kasvaessa kohti ääretöntä estimaattorin arvo poikkeaa parametrin todellisesta arvosta θ todennäköisyydellä 0. Jos niin ei käy, estimaattori on *tarkentumaton* (*inconsistent*).

Esimerkki. Suurten lukujen laki. Olkoot satunnaismuuttujat X_i riippumattomia ja olkoon niillä sama odotusarvo $\mathsf{E}(X_i) = \mu$ ja varianssi $\mathsf{V}(X_i) = \sigma^2$, $i = 1, \ldots, n$. Havaintojen lukumäärän kasvaessa kohti ääretöntä todennäköisyys, että keskiarvo

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

poikkeaa μ:stä enemmän kuin mielivaltaisen pienen nollaa suuremman vakion verran, on nolla. Keskiarvo $\bar{X}$ on tällöin odotusarvon μ tarkentuva estimaattori.

Huom1! Suurten lukujen laki on todella vanha ja mullistava! Jakob I Bernoulli esitti 1692 lauseen ensimmäisen version. Se julkaistiin postuumisti 1713. Huom2! Laki on todennäköisyyden frekvenssitulkinnan taustalla. Huom3! Laista on erilaisia versioita. Esimerkiksi oleellisesti sama laki pätee, jos satunnaismuuttujien X_i varianssi ei ole vakio. Huom4! Laki takaa tarkentuvuuden muttei normaalisuutta, koska laki edellyttää vähemmän kuin keskeinen raja-arvolause (jakso 7.3). □

Parametrille on olemassa usein monia estimaattoreita. Monesti parhaana pidetään sitä, jonka *keskineliövirhe* (*mean-squared error*)

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{E}\,(\hat{\theta} - \theta)^2$$

on pienin. Määritelmä muistuttaa varianssin määritelmää. Mikäli estimaattori on harhaton, keskineliövirhe typistyy estimaattorin varianssiksi. Muulloin keskineliövirhe on estimaattorin varianssin ja harhan neliön summa:

$$\begin{aligned}
\mathsf{E}\,(\hat{\theta} - \theta)^2 &= \mathsf{E}\{\hat{\theta} - \mathsf{E}(\hat{\theta}) + [\mathsf{E}(\hat{\theta}) - \theta]\}^2 \\
&= \mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})]^2 + [\mathsf{E}(\hat{\theta}) - \theta]^2 + 2\mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})][\mathsf{E}(\hat{\theta}) - \theta] \\
&= \mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})]^2 + [\mathsf{E}(\hat{\theta}) - \theta]^2 \\
&= \mathsf{V}(\hat{\theta}) + [b(\theta)]^2.
\end{aligned}$$

Kolmas termi hävisi, koska $\mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})][\mathsf{E}(\hat{\theta}) - \theta] = [\mathsf{E}(\hat{\theta}) - \theta] \times \mathsf{E}\,[\hat{\theta} - \mathsf{E}(\hat{\theta})] = [\mathsf{E}(\hat{\theta}) - \theta] \times [\mathsf{E}(\hat{\theta}) - \mathsf{E}(\hat{\theta})] = 0$. (Siirrettiin vakio odotusarvon eteen, ja vakion odotusarvo on vakio. Jakso 6.3.)

Estimaattori $\hat{\theta}$ on keskineliövirheellä mitattuna tarkempi kuin estimaattori $\tilde{\theta}$, jos

$$\mathsf{E}\,(\hat{\theta}-\theta)^2 < \mathsf{E}\,(\tilde{\theta}-\theta)^2.$$

Jos estimaattorit ovat harhattomia, ehto pelkistyy $\hat{\theta}$:n varianssin pienemmyydeksi:

$$\mathsf{V}(\hat{\theta}) < \mathsf{V}(\tilde{\theta}).$$

Ehto voidaan yhtä hyvin ilmaista keskihajontojen

$$\mathsf{SD}(\hat{\theta}) < \mathsf{SD}(\tilde{\theta})$$

avulla. Niitä kutsutaan tässä yhteydessä *keskivirheiksi*.

Esimerkki. Normaalijakauman varianssin estimointi. Estimoidaan $\mathsf{N}(\mu,\sigma^2)$-jakautuneen satunnaismuuttujan X_i varianssia:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{9.1}$$

ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Voidaan osoittaa, että

$$\mathsf{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 < \sigma^2,$$

$$\mathsf{E}(s^2) = \sigma^2$$

ja että

$$\mathsf{MSE}(\hat{\sigma}^2) = \mathsf{E}\,(\hat{\sigma}^2-\sigma^2)^2 < \mathsf{E}\,(s^2-\sigma^2)^2 = \mathsf{MSE}(s^2)$$

(esim. Lindgren 1976, 216, 256). Estimaattori $\hat{\sigma}^2$ on harhainen mutta keskineliövirheen mielessä tarkempi kuin s^2. Jälkimmäinen on harhaton. Kumpaa tahansa voi perustellusti käyttää. □

Esimerkki. Odotusarvon estimointi I. Symmetrisen jakauman odotusarvoa voidaan estimoida keskiarvolla tai mediaanilla. Oletetaan, että otoksen havainnot ovat riippumattomia ja noudattavat normaalijakaumaa $\mathsf{N}(\mu,\sigma^2)$. Jaksossa 7.3 osoitettiin, että keskiarvo on odotusarvon harhaton estimaattori, jonka varianssi on σ^2/n. Esimerkin tilanteessa mediaani on myös harhaton estimaattori. Sen

varianssi on suurilla havaintomäärillä noin $1.57 \times \sigma^2/n$. Keskiarvo on keskineliövirheen mielessä tarkempi odotusarvon estimaattori kuin mediaani, jos havainnot ovat normaalijakautuneita. □

Esimerkki. Odotusarvon estimointi II. Olkoon tutkittava jakauma symmetrinen ja paksuhäntäinen. Tällöin suuresti odotusarvosta poikkeavat havainnot, oudokit, ovat todennäköisempiä kuin normaalijakauman tilanteessa. Voidaan osoittaa, että mediaani voi olla tällöin keskiarvoa tarkempi estimaattori odotusarvolle keskineliövirheellä mitattuna. Intuitio on, että paksuhäntäisen jakauman tilanteessa aineistoon voi tulla muista havainnoista poikkeavia havaintoja. Poikkeava havainto voi vaikuttaa suuresti keskiarvoon (pienillä otoskoilla) ja kasvattaa sen varianssia muttei muuta mediaania. □

Jos estimaattorin keskineliövirhe suppenee nollaan havaintojen lukumäärän mennessä äärettömään, estimaattori on tarkentuva. Näin ollen harhaton estimaattori on tarkentuva, jos sen varianssi suppenee nollaan havaintojen lukumäärän kasvaessa.

Keskiarvo ja mediaani eivät välttämättä estimoi samaa parametria (odotusarvoa). Mediaani on usein erityisen hyödyllinen tilanteissa, joissa odotusarvo ja teoreettinen mediaani eroavat. Mediaani on myös useammassa mielessä *vakaa* (*robust*) estimaattori. Estimaattori on vakaa tietyn poikkeavuuden suhteen, jos estimaattori ei muutu "paljon" kyseisessä poikkeavassa tilanteessa. Estimaattori voi olla vakaa esimerkiksi satunnaismuuttujan jakauman tai oudokkien suhteen.

Esimerkki. Opiskelijoiden ansiot (jatkoa jaksosta 8.5). Opiskelijoiden ansioiden (0, 0, 0, 0, 5 000, 6 000, 7 000, 8 000, 9 000, 10 000 ja 50 000 euroa) tyyppiarvo, mediaani ja keskiarvo ovat 0, 6 000 ja 8 636.36 euroa. Aineistoa huolella tutkittaessa huomataan, että oudokki 50 000 euroa on virhekirjaus. Korjataan se oikeaksi arvoksi 5 000 euroksi. Tyyppiarvo, mediaani ja keskiarvo ovat nyt 0, 5 000 ja 4 545.45 euroa. Tyyppiarvo ei muuttunut. Mediaani pieneni 1 000 eurolla eli 17 %:lla. Keskiarvo lähes puolittui. Korjatussa aineistossa se on jopa pienempi kuin mediaani. Tyyppiarvo ja mediaani ovat oudokkien suhteen vakaita estimaattoreita; keskiarvo ei ole. □

9.2 Estimointimenetelmistä

Estimaattoreita voidaan johtaa eri tavoilla. Monien estimaattorien lähtökohta on *tilastollinen malli* (*statistical model*). Tilastollinen malli on teoreettinen kuvaus — käytännössä yksinkertaistus — tutkittavasta satunnaisilmiöstä. Mallissa

on systemaattinen ja satunnainen osa. Malli määrittelee satunnaismuuttujista tehtävät oletukset kuten todennäköisyysjakauman ja niiden riippuvuussuhteet ja havaintojen poimintamekanismin. Mallin parametrien numeeriset arvot pyritään selvittämään estimoinnilla.

Esimerkki. Normaalijakautunut otos I. Satunnaismuuttuja X noudattaa normaalijakaumaa $\mathsf{N}(\mu, \sigma^2)$:

$$X = \mu + \varepsilon.$$

Odotusarvo μ on mallin systemaattinen ja satunnaistermi ε satunnainen osa. Molempiin liittyy estimoitava parametri (μ ja σ^2). Otoksen havainnot oletetaan poimituksi riippumattomasti toisistaan.

Riippumattomasti poimitun otoksen $X_1, \ldots, X_n$ *yhteistiheysfunktio* (*joint density function*) on

$$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} = L(x_1, \ldots, x_n; \mu, \sigma^2). \tag{9.2}$$

Intuitio kaavalle on riippumattomien tapahtumien tulosääntö (4.9). Kaavan yllä avulla voidaan laskea havaintoihin $x_1, \ldots, x_n$, liittyviä todennäköisyyksiä (tällaiset laskut sivuutetaan). Kaava (9.2) summeeraa esimerkin tilastollisen mallin. Merkintä $L(x_1, \ldots, x_n; \mu, \sigma^2)$ tulee tulkita tässä niin, että parametrit μ ja σ^2 ovat kiinteitä, ja todennäköisyydet vaihtelevat havaintojen $x_1, \ldots, x_n$ mukaan. □

Suurimman uskottavuuden (SU, *maximum likelihood*) menetelmä on tärkeimpiä estimointimenetelmiä. Menetelmän idea: Valitaan jakauman parametrien estimaateiksi lukuarvot, joilla todennäköisyys havaitulle aineistolle on suurin mahdollinen. Teknisempi kuvaus: Jakauman ja oletusten (kuten havaintojen riippumattomuus) perusteella muodostetaan uskottavuusfunktio, jonka arvo maksimoidaan valitsemalla sopivat parametriarvot.

SU-estimaatit ovat usein, mutteivät aina, helposti laskettavissa. SU-estimaattoreiden voidaan sopivin oletuksin osoittaa olevan suurilla havaintomäärillä keskineliövirheen mielessä tarkimpia mahdollisia ja normaalijakautuneita. SU-estimaattorit eivät ole välttämättä harhattomia.

Esimerkki. Binomijakautunut otos. On tehty n riippumatonta havaintoa x_i Bernoulli-jakautuneesta satunnaismuuttujasta X. Tapahtuman $y = \sum_{i=1}^{n} x_i$ to-

dennäköisyys määräytyy binomitodennäköisyydestä (7.3). Tulkitsemalla siinä y kiinteäksi ja π funktion argumentiksi saadaan uskottavuusfunktio

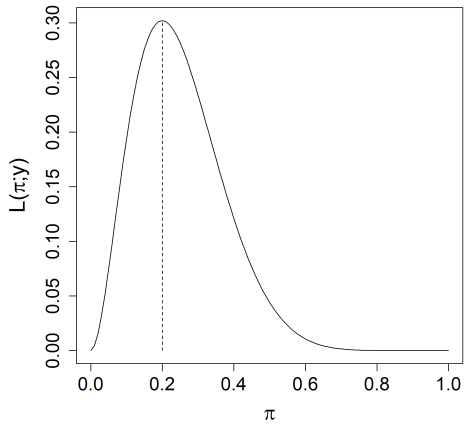
$$L(\pi; y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}.$$

Luvussa 7 oletettiin parametri π tunnetuksi ja laskettiin todennäköisyyksiä tapahtumien lukumäärälle y sijoittamalla sille eri arvoja kaavassa yllä. Nyt toimitaan toisinpäin: Kiinnitetään lukumäärä y havaituksi ja lasketaan $L(y; \pi)$:n arvoja eri π:n arvoilla. Se π:n arvo, jolla $L(y; \pi)$ saa suurimman arvonsa, on π:n SU-estimaatti $\hat{\pi}$. Voidaan osoittaa, että $\hat{\pi} = y/n$ (harjoitustehtävä).

Kuvassa 9.1 on uskottavuusfunktio $L(y; \pi)$ tilanteessa $y = 2$ ja $n = 10$:

$$L(\pi; y) = \binom{10}{2} \pi^2 (1-\pi)^{10-2} = 45 \times \pi^2 (1-\pi)^8.$$

Uskottavuusfunktio saa suurimman arvonsa π:n arvolla $0.2 = 2/10$. Se on π:n SU-estimaatti. □



Kuva 9.1: Binomijakauman parametrin π uskottavuusfunktio, jos $y = 2$ ja $n = 10$.

Esimerkki. Normaalijakautunut otos II. Satunnaismuuttuja X noudattaa normaalijakaumaa $\mathsf{N}(\mu, \sigma^2)$. Jakaumasta poimitaan n riippumatonta havaintoa. Us-

kottavuusfunktio on $L(x_1, \ldots, x_n; \mu, \sigma^2)$ (kaava (9.2)) tulkiten havainnot x_i kiinteiksi ja parametrit μ ja σ muuttujiksi, joiden suhteen kaava maksimoidaan. Voidaan osoittaa, että SU-estimaatit normaalijakauman parametreille ovat $\hat{\mu} = \bar{x}$ (havaintojen keskiarvo) ja $\hat{\sigma}^2$ (kaava (9.1)). □

Uskottavuusfunktiota on kuvattu tilastotieteen keskeisimmäksi käsitteeksi (Barndorff-Nielsen 1991) ja SU-menetelmää sovelletun matematiikan vaikutusvaltaisimmaksi saavutukseksi 1900-luvulla (Efron ja Hastie 2016, 91). (Allekirjoittaneesta aineisto on silti tilastotieteen käsitteistä keskeisin.)

Toinen paljon käytetty estimointimenetelmä on *pienimmän neliösumman* (*least squares*, PNS) menetelmä. Tavallisissa normaalijakauman tilanteissa se ja SU-menetelmä tuottavat samat estimaattorit. PNS-menetelmä kuvataan regressioanalyysin yhteydessä (luku 13).

Alla ei yleensä todeta, millä menetelmällä estimaattori on johdettu. Estimaattorin intuitiivisuus on riittävä peruste tutkia ja soveltaa sitä luentomateriaalissa. Havaintojen (n kappaletta) oletetaan noudattavan otsikoissa nimettyjä jakaumia ja olevan riippumattomia, jollei toisin todeta.[111] “Hatulla” (“$\wedge$”) merkityt estimaattorit ovat luvussa SU-estimaattoreita.

9.3 Binomijakauman parametrin estimointi

Binomijakauman parametrille π luonteva estimaattori on tapahtumien osuus otoksessa eli tapahtumien lukumäärä (Y) jaettuna otoskoolla (n):

$$\hat{\pi} = \frac{Y}{n}.$$

Se on harhaton estimaattori:

$$\mathsf{E}(\hat{\pi}) = \mathsf{E}\left(\frac{Y}{n}\right) = \frac{1}{n}\mathsf{E}(Y) = \frac{n\pi}{n} = \pi$$

(kaava (7.4)). Estimaattorin varianssi on

$$\mathsf{V}(\hat{\pi}) = \mathsf{V}\left(\frac{Y}{n}\right) = \frac{1}{n^2}\mathsf{V}(Y) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

(kaava (7.5)). Estimaattorin varianssi menee nollaan otoksen koon kasvaessa kohti ääretöntä, ja estimaattori on tarkentuva. Normeerattu estimaattori $n\hat{\pi}$ on binomijakautunut

$$n\hat{\pi} = Y \sim \mathsf{Bin}(n, \pi)$$

ja suurilla havaintomäärillä likimain normaalijakautunut

$$\hat{\pi} \sim \mathsf{N}(\pi, \pi(1-\pi)/n)$$

keskeisen raja-arvolauseen mukaan (jakso 7.4.3).

9.4 Multinomijakauman parametrien estimointi

Multinomijakauman luokkatodennäköisyyksille luonteva estimaattori on luokkafrekvenssien (N_i) osuudet otoksessa:

$$\hat{\pi}_i = \frac{N_i}{n},$$

$i = 1, \ldots, c$. Normeeratut estimaattorit $n\hat{\pi}_i = N_i$ noudattavat multinomijakaumaa $\mathsf{Mul}(n, \pi_1, \ldots, \pi_c)$. Yksittäinen luokkalukumäärä N_i ja siten $n\hat{\pi}_i$ on binomijakautunut jaksoissa 7.1.4 ja 9.3 kuvatulla tavalla, ja $\hat{\pi}_i$ on harhaton ja tarkentuva estimaattori. Keskeisen raja-arvolauseen (jakso 7.3) perusteella kukin $\hat{\pi}_i$ on normaalijakautunut suurilla havaintomäärillä:

$$\mathsf{E}(\hat{\pi}_i) = \pi_i,$$

$$\mathsf{V}(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n}$$

ja

$$\hat{\pi}_i \sim \mathsf{N}(\pi_i, \pi_i(1-\pi_i)/n).$$

9.5 Poisson-jakauman parametrin estimointi

Kun $Y_i \sim \mathsf{Poi}(\mu)$, niin keskiarvo on luonteva estimaattori myös Poisson-jakauman parametrille μ:

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

Estimaattori $\hat{\mu}$ on harhaton ja sen varianssi suppenee nollaan otoskoon kasvaessa:

$$\mathsf{E}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{E}(Y_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \mu$$

ja

$$\mathsf{V}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{V}(Y_i)}{n^2} = \frac{\sum_{i=1}^{n} \mu}{n^2} = \frac{\mu}{n}.$$

Varianssin lasku perustuu havaintojen riippumattomuuteen. Estimaattori on tarkentuva.

Kaavan (7.12) ja jakson 7.4.5 perusteella $\sum_{i=1}^{n} Y_i \sim \mathsf{Poi}(n\mu)$. Näin ollen normeerattu estimaattori $n\hat{\mu} = \sum_{i=1}^{n} Y_i$ on Poisson-jakautunut:

$$n\hat{\mu} \sim \mathsf{Poi}(n\mu).$$

Keskeisestä raja-arvolauseesta (7.14) seuraa, että suurilla havaintomäärillä pätee likimäärin

$$\hat{\mu} \sim \mathsf{N}(\mu, \mu/n).$$

9.6 Normaalijakauman parametrien estimointi

Normaalijakauman $\mathsf{N}(\mu, \sigma^2)$ odotusarvolle ilmeinen estimaattori on otoskeskiarvo

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Se on harhaton, ja sen varianssi menee nollaan havaintojen lukumäärän n kasvaessa:

$$\mathsf{E}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{E}(X_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \mu$$

ja

$$\mathsf{V}(\hat{\mu}) = \frac{\sum_{i=1}^{n} \mathsf{V}(X_i)}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Voidaan osoittaa, että

$$\frac{\hat{\mu} - \mu}{s/\sqrt{n-1}} \sim t(n-1)$$

(esim. Lindgren 1976, 344). Tilastollinen päättely odotusarvosta μ kannattaa perustaa ylipäänsä tunnuslukuun yllä. Keskeisen raja-arvolauseen (7.14) perusteella $\hat{\mu}$ noudattaa suurilla havaintomäärillä likimäärin normaalijakaumaa

$$\mathsf{N}(\mu, \sigma^2/n).$$

Varianssin estimointia pohdittiin jo esimerkissä jaksossa 9.1. Voidaan osoittaa, että sekä $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$ että $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ ovat σ^2:n tarkentuvia estimaattoreita ja että

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{9.3}$$

(esim. Lindgren 1976, 334).

9.7 Odotusarvon estimointi ilman jakaumaoletusta

Ilmiön taustalla olevaa jakaumaa ei voida aina määritellä. Tällöin voidaan turvautua keskeiseen raja-arvolauseeseen (7.14) odotusarvoa estimoitaessa, jos lauseen oletukset ovat voimassa. Odotusarvon otosvastine ja ilmeinen estimaattori on keskiarvo. Keskeisen raja-arvolauseen mukaan suurilla otoskoilla keskiarvo on likimain normaalijakautunut:

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1) \quad \text{eli} \quad \bar{Y} \sim \mathsf{N}(\mu, \sigma^2/n).$$

Jaksojen 7.3 ja 9.1 johdot keskiarvoestimaattorin harhattomuudelle ja varianssille pätevät tässäkin tilanteessa. Keskiarvo on harhaton ja tarkentuva estimaattori odotusarvolle vaikkei taustalla olevaa jakaumaa rajattaisi tarkasti.

9.8 Korrelaation estimointi ja ekologinen korrelaatio

Oletetaan, että tutkittavat satunnaismuuttujat ovat binormaalijakautuneita (jakso 7.5) ja että havaintoparit $(X_1, Y_1), \ldots, (X_n, Y_n)$ ovat riippumattomia. Otoskorrelaation $\hat{\rho}$ (kaava 8.2) jakauma riippuu vain korrelaatiosta ρ ja otoskoosta n. Jakauma vinoutuu ρ:n lähestyessä ± 1:htä ja voi olla hyvin vino (Cramer 1946, 399–400). $\hat{\rho}$ on lievästi harhainen kohti 0:aa, jos $\rho \neq 0$ ($|\mathsf{E}(\hat{\rho})| < |\rho|$). Harha on suurimmillaan, kun $|\rho|$ on noin 0.5–0.7 (Demidenko 2020, 392, Shieh 2010).

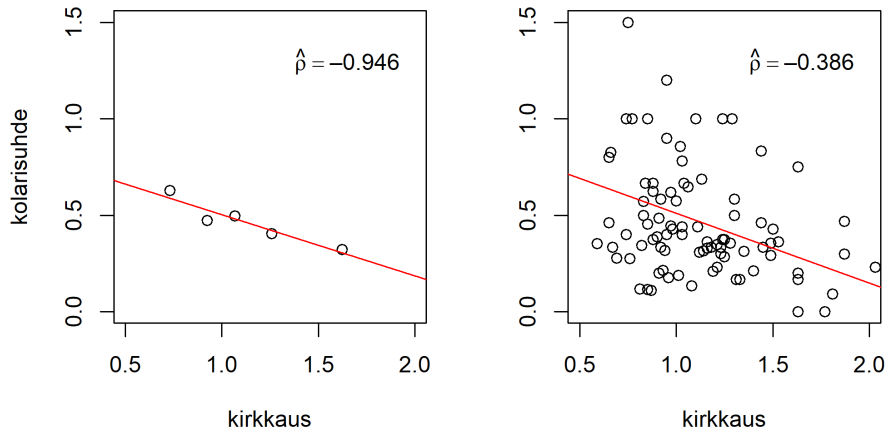
Esimerkki. Olkoon $n = 20$ ja $|\rho|$ noin $0.5 - 0.7$. Otoskorrelaation harha kohti 0:aa on tällöin noin 0.01. Harha on suurimmillaankin pieni. □

Otoskorrelaation jakaumaa tarkastellaan lähemmin korrelaation testaamista kuvaavassa jaksossa 12.7.

Kuvan 9.2 sirontakuviot havainnollistavat, kuinka keskiarvoista laskettu korrelaatio voi olla paljon suurempi kuin alkuperäisestä aineistosta laskettu. Sirontakuvioiden mukaan pimeän ajan kolarien lukumäärän suhde valoisan ajan kolarien lukumäärään pienenee katuvalaistuksen kirkkauden kasvaessa. Keskiarvojen otoskorrelaatio on -0.946; havaintojen -0.386. Yhteys vaikuttaa kovasti tiukemmalta keskiarvoista laskettuna.

Otoskorrelaatioita lasketaan monesti ryhmäkeskiarvoista. Tällaisia otoskorrelaatioita kutsutaan *ekologisiksi korrelaatioiksi* (Robinson 1950, te Grotenhuis

ym. 2011). Ekologinen korrelaatio voi olla hyödyllinen suure, jos ollaan kiinnostuneita ryhmätason korrelaatiosta. Joskus yksilötason korrelaatio kiinnostaa mutta otoskorrelaatio lasketaan ryhmätasolla, koska vain ryhmätason aineisto on saatavilla. Ekologiset korrelaatiot ovat usein itseisarvoltaan suurempia kuin vastaavat yksilötason korrelaatiot. Ekologiset korrelaatiot voivat siksi antaa liioitellun vaikutelman tilastollisen yhteyden vahvuudesta. *Ekologinen virhepäätelmä* (*ecological fallacy*) tehdään, jos ryhmätason korrelaatio samaistetaan yksilötason korrelaatioon.



Kuva 9.2: Pimeän ajan kolarien lukumäärän suhde valoisan ajan kolarien lukumäärään ja katuvalaistuksen kirkkaus (cd/m^2). Vasen: Keskiarvot. Oikea: Alkuperäiset havainnot.[2]

[2]Kiitän Paul Marchantia aineiston luovuttamisesta 23.3.2020. Olen tehnyt vasemmanpuoleisen kuvan jakamalla aineiston kirkkausmuuttujan suuruuden mukaan 20, 17, 18, 17 ja 17 havainnon ryhmiin ja korvaamalla kunkin ryhmän kirkkaus- ja kolarisuhdehavainnot niiden ryhmäkeskiarvoilla Marchantin (2019) esimerkin mukaisesti. Olen piirtänyt molempiin kuviin regressiosuorat (luku 13). Aineisto on Hargrovesin ja Scottin (1979) tutkimuksesta kolareista Iso-Britanniassa.

Luku 10

Väliestimointi

> Mikä esitetyistä näkemyksistä on totta, sen päättäköön joku jumala. Mikä niistä on todennäköisin, se on suuri kysymys.[112]
>
> Marcus Tullius Cicero (106–43 eKr.)

Luvussa kuvataan teoreettisia tuloksia ja havainnollistetaan niiden käyttöä empiirisillä esimerkeillä. Luvussa ei ole mahdollista tutustua yksityiskohtaisesti esimerkkiaineistoihin. Todellisessa tutkimuksessa aineistoon tulee perehtyä huolellisesti ennen väliestimointia tai muuta tilastollista päättelyä. Muun muassa kannattaa piirtää kuvioita havainnoista ja niiden jakaumista. Aineistojen oletetaan olevan satunnaisotoksia luvussa.

10.1 Idea

Tilastotieteen ytimessä on päätelmiin liittyvän epävarmuuden kuvaaminen ja mittaaminen. Ajatus oli rivien välissä edellisessä luvussa, jossa estimaattorit noudattivat erilaisia jakaumia tilanteesta riippuen. *Väliestimointi* (*interval estimation*) on monien suosittelema tapa kuvata estimaatin tai muun jakaumaan liittyvän suureen tarkkuutta ja tehdä tilastollista päättelyä. Väliestimointi eroaa piste-estimoinnista kuin verkon heittäminen keihään heittämisestä kalaa kohti. (Ismay ja Kim 2020, 254–255).

Jos moni sekaantuu todennäköisyyslaskennassa, niin tilastollisen päättelyn käsitteissä on sekä opiskelijoilla että soveltajilla syvällemeneviä väärinkäsityksiä (esim. Gigerenzer ym. 2004, Sotos ym. 2007). Niitä saattaa kuulla väitöstilaisuuksissa tai lukea julkaistuista empiirisistä artikkeleista. Päättelyn teoriaan

kannattaa syventyä — yksinkertaisesti jotta ymmärtää, mitä tekee tilastotieteellisiä menetelmiä soveltaessaan. Käsitteet eivät ole niin vaikeita, etteivät peruskurssien opiskelijat niitä voisi omaksua. Väärät tulkinnat johtunevat pikemminkin jonkinlaiseen (tarkoitukseen sopimattoman) maalaisjärkeen perustuvan ajattelun soveltamisesta oikotienä kuviteltuun ymmärrykseen. Ilmeinen virhelähde ovat soveltajien laatimat oppikirjat, joissa käsitteet on tulkittu heppoisesti (mt:t).

Koostukoon otos satunnaismuuttujista $X_1, \ldots, X_n$, ja olkoon estimoitava parametri θ. Luottamusväli θ:lle toteuttaa (tyypillisissä tilanteissa) epäyhtälön

$$\mathsf{P}[L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n)] = \mathsf{P}(L \leq \theta \leq U) \geq 1 - \alpha. \qquad (10.1)$$

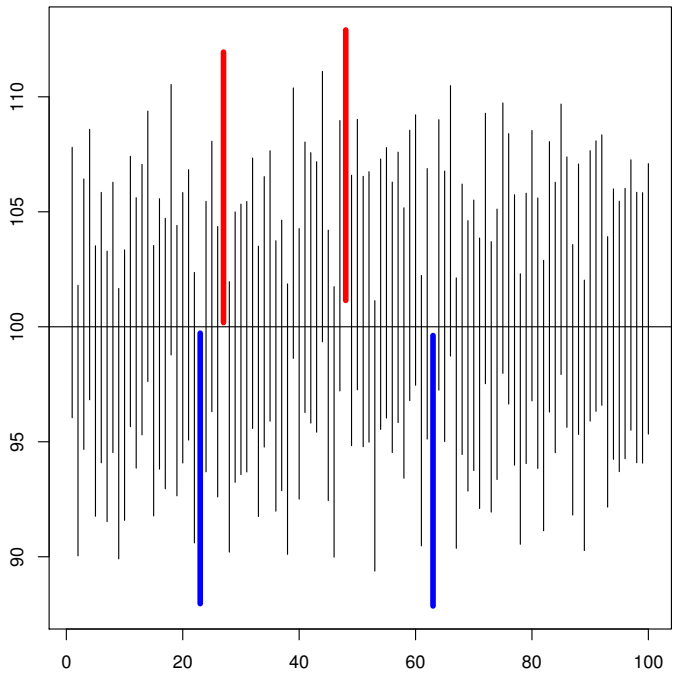
Siinä $L = L(X_1, \ldots, X_n)$, $U = U(X_1, \ldots, X_n)$, $L \leq U$ ja $\alpha \in (0, 1)$. Luottamusvälin ala- ja ylärajat ovat L ja U. Luottamusväli $[L, U]$ on θ:n *väliestimaattori*, ja havaitusta otoksesta laskettu väli $[l, u] = [L(x_1, \ldots, x_n), U(x_1, \ldots, x_n)]$ on θ:n *väliestimaatti* (*interval estimate*). Ala- ja ylärajat ovat otoksesta laskettavia tunnuslukuja, joiden määritelmä riippuu jakaumasta, johon θ liittyy.

Epäyhtälön (10.1) mukaan todennäköisyys, että väli $[L, U]$ peittää θ:n, on vähintään $1-\alpha$. Tässä yhteydessä todennäköisyyttä $1-\alpha$ kutsutaan *luottamustasoksi* (*confidence level*). Väliä kutsutaan $100 \times (1 - \alpha)$ %:n *luottamusväliksi* (*confidence interval*). Tyypillisesti α on 0.05 tai 0.01, jolloin lasketaan 95 %:n tai 99 %:n luottamusväli.

Väliestimoinnissa puhutaan luottamuksesta todennäköisyyden sijaan, jotta parametriin θ ei sekoitettaisi todennäköisyyden käsitettä (Neyman 1934, 590). Parametri on kiinteä populaatiota kuvaava suure eikä ole satunnaismuuttuja. Satunnaismuuttujia ovat luottamusvälin ala- ja ylärajat. Ne vaihtelevat otoksesta toiseen, ja niiden rajaama väli peittää θ:n (vähintään) $100 \times (1-\alpha)$ %:ssa hypoteettisista riippumattomista toistokokeista. Luottamuskin saattaa olla liian latautunut käsite. Luottamusvälille on ehdotettu neutraalimpaa nimitystä *yhteensopivuusväli* (*compatibility interval*). Sen peittämät parametriarvot ovat yhteensopivia aineiston kanssa.

Esimerkki. Sata luottamusväliä.[113] Kuvassa 10.1 on sata luottamusväliä odotusarvolle (100). Neljä väritettyä luottamusväliä ei peitä odotusarvoa. Tässä sadan otoksen toistokokeessa 95 %:n luottamusväleistä 96 % peittää estimoitavan parametrin. Jos toistojen lukumäärää kasvatettaisiin kohti ääretöntä, odotusarvon peittävä osuus väleistä menisi 95 %:iin. Kuvan luottamusvälien laskutapa selitetään jaksossa 10.4.1. □

Väliestimaatin etu piste-estimaattiin verrattuna on, että väliestimaatista saa käsityksen estimaatin tarkkuudesta. Mikäli luottamusväli on leveä, piste-esti-



Kuva 10.1: Tilastotieteellinen Sibelius-monumentti. Sadasta 36 havainnon otoksesta lasketut odotusarvon 95 %:n luottamusvälit (kaava (10.6)). Havaintojen jakauma: $\mathsf{N}(100, 18^2)$.

maatti $\hat{\theta}$ ei ole osunut välttämättä lähelle θ:aa. Jos väli on kapea, θ vaikuttaa tulleen estimoitua tarkasti.

Mikäli $\hat{\theta}$:n jakauma on jatkuva ($\hat{\theta}$ voi saada äärettömän määrän arvoja), on piste-estimaattorin todennäköisyys peittää θ nolla. Todennäköisyydellä mitattuna väliestimoinnilla saavutetaan tavaton parannus piste-estimointiin verrattuna!

Useimmiten lasketaan symmetrisiä kaksisuuntaisia luottamusvälejä, joissa $\hat{\theta}$ on luottamusvälin keskipiste ja $\hat{\theta} - L = U - \hat{\theta}$. Epäyhtälön (10.1) määrittä-

mä luottamusväli ei ole välttämättä symmetrinen. Joskus on perusteltua laskea yksisuuntainen luottamusväli kuten $(-\infty, U]$, $[L, \infty)$, $[0, U]$ tai $[L, 1]$.

Luottamusvälejä voidaan laskea monella tavalla. Yleensä luottamusväli yritetään muodostaa niin, että se olisi mahdollisimman kapea mutta silti toteuttaisi epäyhtälön (10.1) ja että todennäköisyys siinä olisi tasan $1 - \alpha$. Jos jakauma on diskreetti ja havaintoja on vähän, se ei ole aina mahdollista. Luottamusvälin peittävyystodennäköisyys on siksi määritelty epäyhtälönä kaavassa (10.1).

Luottamusvälin tulkinnassa kannattaa pitäytyä parametrin todellisen arvon peittävyystulkinnassa riippumattomissa toistokokeissa. Venytetyt väärät tulkinnat ovat yleisiä.

Esimerkki. Luottamusvälin tulkinta. Väliestimoidaan parametria θ. Sen (95 %:n) luottamusvälin vääriä tulkintoja:

- "θ sijoittuu 95 %:n luottamusvälille 95 %:n todennäköisyydellä." Ei!: θ on kiinteä eikä liiku minnekään. Luottamusväli on satunnainen ja sijoittuu eri kohtiin riippumattomissa toistokokeissa.
- "95 %:n luottamusvälin peittämät θ:n arvot ovat yhtätodennäköisiä." Ei!: θ:n arvot eivät ole satunnaismuuttujia, joihin liittyisi todennäköisyys.
- "Riippumattomissa toistokokeissa parametrin θ piste-estimaateista 95 % sijoittuu (yhdestä otoksesta lasketulle) 95 %:n luottamusvälille θ:lle." Ei!: Otos on voinut olla poikkeava, ja piste-estimaatit voivat tavata osua toisaalle. □

Yleinen on myös väärä käytäntö katsoa, menevätkö kahden odotusarvon luottamusvälit päällekkäin ja päätellä siitä, eroavatko parametrit. Oikea menettely on laskea luottamusväli odotusarvojen erotukselle, mikä selitetään eri jakaumien tilanteessa jaksoissa alla. Väärän käytännön ongelmia setvitään harjoitustehtävissä.

Esimerkki. Luottamusvälien vertailu. Niewenhuis ym. (2011) kävivät läpi 513 viidessä huippulehdessä (mm. *Science* ja *Nature*) julkaistua tutkimusta.[114] Tutkijat vertasivat niissä yhtä usein luottamusvälejä väärin kuin käyttivät niitä oikein. Helsingin, Tampereen ja Turun yliopistojen psykologian pääsykoekirjassa 2017 opetettiin virheellinen menettely kahden luottamusvälin vertaamiseksi. □

Fagerlandin ym:iden (2017), Newcomben (2013) sekä erityisesti Meekerin ym:iden (2017) kirjoissa opastetaan luottamusvälien laskua yksityiskohtaisemmin ja kattavammin.

10.2 Luottamusvälejä osuuksille

10.2.1 Osuuden luottamusväli

Binomijakauman todennäköisyyden π:n eli osuuden luottamusväli on käytetyimpiä luottamusvälejä. Tilastotieteen perusoppikirjoissa järjestään opetetaan menettely alla sen laskemiseksi. Johto lähtee osuuden estimaattorin $\hat{\pi} = Y/n$ normaalisuudesta suurilla havaintomäärillä:

$$\hat{\pi} \sim \mathsf{N}(\pi, \pi(1-\pi)/n)$$

(jakso 9.3). Tällöin pätee likimäärin

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} < \mathsf{z}_{1-\alpha/2}\right)$$
$$\mathsf{P}\left(-\mathsf{z}_{1-\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} < \mathsf{z}_{1-\alpha/2}\right) = 1 - \alpha$$

(vrt. kaava (7.14)). Tässä $\mathsf{z}_{\alpha/2} = -\mathsf{z}_{1-\alpha/2}$ on standardinormaalijakauman $\alpha/2$. kvantiili (esim. $\mathsf{z}_{0.025} = -\mathsf{z}_{0.975} = -1.960$). Estimoidaan estimaattorin $\hat{\pi}$ nimittäjässä varianssi $\hat{\pi}(1-\hat{\pi})/n$:llä. Näin saadaan likimääräiset yhtälöt

$$\mathsf{P}\left(-\mathsf{z}_{1-\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} < \mathsf{z}_{1-\alpha/2}\right) \approx 1 - \alpha \Leftrightarrow$$
$$\mathsf{P}\left(\hat{\pi} - \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} < \pi < \hat{\pi} + \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) \approx 1 - \alpha.$$

Osuuden $100 \times (1-\alpha)$ %:n likimääräisen luottamusvälin ylä- ja alaraja ovat

$$\hat{\pi} \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{10.2}$$

Jos havaintoja on paljon eikä π ole lähellä nollaa tai yhtä, luottamusvälin peittävyys on noin $100 \times (1-\alpha)$ %. Väliä (10.2) kutsutaan tässä Waldin luottamusväliksi tilastotieteilijä Abraham Waldin mukaan.

Luottamusvälin leveys riippuu luottamustasosta $1-\alpha$, π:n suuruudesta ja otoskoosta n. Luottamustason kasvattaminen $(1-\alpha)$:sta $(1-\alpha^*)$:iin leventää luottamusväliä ($\alpha^* < \alpha$). Tällöin luottamusvälin leveyden määräävä termi suurenee: $\mathsf{z}_{1-\alpha^*/2}[\hat{\pi}(1-\hat{\pi})/n]^{1/2} > \mathsf{z}_{1-\alpha/2}[\hat{\pi}(1-\hat{\pi})/n]^{1/2}$, koska $\mathsf{z}_{1-\alpha^*/2} > \mathsf{z}_{1-\alpha/2}$.

Esimerkki. Luottamusvälin leveys ja luottamustaso. Olkoon $\hat{\pi} = 0.5$ ja $n = 100$. Jos luottamustaso on 0.95, niin $\alpha = 0.05$ ja $\mathtt{z}_{1-\alpha/2} = 1.960$. Jos luottamustaso on 0.99, niin $\alpha = 0.01$ ja $\mathtt{z}_{1-\alpha/2} = 2.576$. Standardinormaalijakauman 0.975. ja 0.995. kvantiilit 1.960 ja 2.576 on laskettu R-komennoilla `qnorm(0.975)` ja `qnorm(0.995)`. Kaksisuuntaisen 95 %:n luottamusvälin rajat ovat

$$0.5 \pm 1.960\sqrt{\frac{0.5 \times 0.5}{100}} = 0.5 \pm 1.960 \times 0.05 = 0.5 \pm 0.098,$$

eli luottamusväli on

$$[0.402, 0.598].$$

Kaksisuuntaisen 99 %:n luottamusvälin rajat ovat

$$0.5 \pm 2.576\sqrt{\frac{0.5 \times 0.5}{100}} = 0.5 \pm 2.576 \times 0.05 = 0.5 \pm 0.1288,$$

jolloin luottamusväli on

$$[0.371, 0.629].$$

Luottamustason suurentaminen leventää luottamusvälin pituuden $0.598-0.402 = 0.196$:sta $0.630 - 0.371 = 0.258$:aan.

Edelliset luottamusvälit selviävät kätevästi mosaic-paketin komennoilla:

```
install.packages("mosaic")
library(mosaic)
binom.test(x=50,n=100,conf.level=0.95,ci.method="Wald")
binom.test(x=50,n=100,conf.level=0.99,ci.method="Wald")
```

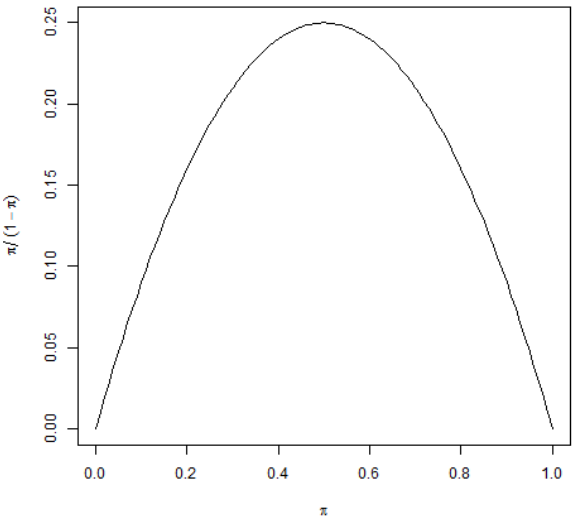
□

Luottamusväli tapaa olla sitä leveämpi, mitä lähempänä π ja siten $\hat{\pi}$ on 0.5:ttä. Ääriarvoja 0 tai 1 lähellä olevat osuudet estimoituvat tarkemmin kuin 0.5:n tienoilla olevat. Kuva 10.2 havainnollistaa tuloa $\pi(1-\pi)$, kun $\pi \in [0, 1]$.

Otoskoon n kasvattaminen kuristaa luottamusväliä, koska n:n suuretessa luottamusvälin (10.2) neliöjuuritermi pienenee. Luottamusväli kapenee kuitenkin hitaammin kuin n suurenee.

Esimerkki. Estimaatin $\hat{\pi}$ pysyessä samana luottamusvälin (10.2) leveys puolittuu, jos n nelinkertaistuu:

$$1.960 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{4n}} = \frac{1}{2} \times 1.960 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$



Kuva 10.2: Todennäköisyys π ja tulo $\pi(1-\pi)$.

Vasemmalla on kaksisuuntaisen luottamusvälin puolikas, kun otoskoko on $4n$. Oikealla on 1/2:lla kerrottuna otoskoolla n pätevän kaksisuuntaisen luottamusvälin (10.2) puolikas. □

On esitetty peukalosääntöjä, joiden pätiessä approksimaation (10.2) pitäisi toimia.[115] Luottamusvälin (10.2) peittävyys voi olla paljon pienempi kuin $100 \times (1-\alpha)$ %, vaikka peukalosäännön mukaan approksimaatio olisi pätevä.

Esimerkki. Osuuden luottamusvälin (10.2) peittävyys.[116] Jos havaintoja on 25 ja π on noin 0.05, niin 95 %:n luottamusvälin (10.2) peittävyys on noin 70 %. □

Esimerkki. Luottamusväli, jos tapahtumia ei ole. Jos $\hat{\pi} = y/n = 0/n = 0$, luottamusväli (10.2) typistyy pisteeksi $[0, 0]$ kaikilla luottamustasoilla. Se on epätyydyttävää: Jos voitaisiin olla varmoja, että tapahtuman todennäköisyys on 0, luottamusväliä ei olisi laskettu. □

Newcombista (1998a, 868) luottamusväliä (10.2) ei tulisi käyttää tieteellisessä

tutkimuksessa ja sen käyttö tulisi rajata otoskoon suunnitteluun ja opetustarkoituksiin. Andersson (2023), Fagerland ym. (2017, 65), Meeker ym. (2017, 105, 108) sekä Schilling ja Doi (2014) arvioivat luottamusväliä (10.2) samaan tapaan.

Paljon paremmin toimiva kaksisuuntainen 95 %:n luottamusväli on *plus neljä -luottamusväli* (Agresti ja Coull 1998). Havaintoihin lisätään neljä havaintoa taulukon alla mukaisesti (alun perin n havaintoa ja y tapahtumaa):

tapahtuma		
kyllä	ei	Σ
$y+2$	$n-y+2$	$n+4$

Σ:lla on merkitty summaa rivin lukumääristä. Luottamusväli lasketaan näin muokatusta aineistosta kaavalla (10.2). Plus neljä -luottamusväli on hieman leveämpi ja peittää todellisen osuuden todennäköisyydellä, joka tapaa olla selvästi lähempänä 95 %:a kuin alkuperäisestä aineistosta kaavalla (10.2) laskettu luottamusväli (mt., Newcombe 2013, 106, 109, Andrés ja Hernández 2014). Jos π on hyvin lähellä nollaa tai yhtä, plus neljä -luottamusväli on liian leveä. Komennoilla

```
library(mosaic)
binom.test(x=3,n=20,ci.method="Plus4")
```

mosaic-paketti laskee plus neljä -välin nanosekunnissa (käskyssä esimerkkinä 3 tapahtumaa 20 toistossa).

Plus neljä -luottamusväli on suunniteltu 95 %:n luottamusvälin laskemiseen, mutta se on Waldin luottamusväliä (10.2) parempi muillakin luottamustasoilla (Andrés ja Hernández 2014). Myös yksisuuntainen plus neljä -luottamusväli kannattaa laskea ennemmin kuin yksisuuntainen Waldin luottamusväli (Pradhan ym. 2016). Näissä tilanteissa on silti suositeltavaa laskea luottamusväli muilla menetelmillä (jakson lopun lisämateriaali).

Joskus binomikokeessa tapahtumia ei tule lainkaan. Tällöin kätevä kaava π:n 95 %:n luottamusväliksi on *kolmen sääntö* (*rule of three*)

$$\left[0,\ \frac{3}{n}\right] \approx \left[0,\ \frac{3}{n+1}\right].$$

Jälkimmäinen likiarvoistus on parempi (Jovanovic 2005). Likiarvoistukset ovat toimivia, jos $n > 30$. Nämä ovat yksisuuntaisia luottamusvälejä.

Esimerkki.[117] Helsingin Sanomat 19.8.2022:

> Kun Mariette Hägglund oli tekemässä – – laskuvarjohyppyä neljästä kilometristä, hän huomasi, että laskuvarjo ei avautunutkaan – – . Tilanne oli Hägglundin mukaan verrattain harvinainen. Hän kertoo, että moni laskuvarjohyppääjäkaverikin on turvautunut varavarjon käyttöön. Hägglund kuitenkin arvioi, että näin tapahtuu vain noin kerran 500 hyppyä kohden.

Tehdas valmistaa 300 laskuvarjoa uudella menetelmällä. Jokaisen uuden laskuvarjon aukeamista käytössä kokeillaan, ja kaikki aukeavat. Laske 95 %:n luottamusväli uusille laskuvarjoille, jotka eivät aukea käytössä. Kolmen säännön mukaan yksisuuntainen 95 %:n luottamusväli on

$$\left[0,\, \frac{3}{301}\right] \approx [0, 0.01].$$

Laskuvarjon avautumattomuuden todennäköisyys vaikuttaa pieneltä. □

Luottamusvälin laskemiseksi on parempia menetelmiä kuin plus neljä -luottamusväli. Luottamusvälit voidaan jakaa kahteen ryhmään: Konservatiivisiin, joiden peittävyys on aina vähintään nimellinen luottamustaso (jakso 11.3), ja muihin.

Clopperin–Pearsonin luottamusväli sekä Blakerin luottamusväli ovat konservatiivisia. Fagerland ym. (2017, 64) sekä Park ja Leemis (2019) pitävät Blakerin luottamusväliä parempana. Sen peittävyys on lähempänä nimellistä luottamustasoa, ja se tapaa olla kapeampi kuin Clopperin–Pearsonin luottamusväli.

Muiden luottamusvälien peittävyys voi olla vielä lähempänä nimellistä luottamustasoa. Fagerland ym. (2017, 64) suosittelevat Wilsonin luottamusväliä. Newcombista (1998a) se on ainoa helppolaskuinen (ratkaistaan kvadraattinen yhtälö) toimiva menetelmä. Wilsonin luottamusväli toimi hyvin myös Parkin ja Leemisin (2019) tutkimuksessa. Fagerland ym. (2017, 64) pitävät myös keski-p-korjattua (*mid-p*) Clopperin–Pearsonin ja Agrestin–Coullin luottamusvälejä hyvinä. Edellinen on laskennallisesti työläämpi. Jälkimmäinen on laskennallisesti helppo, plus neljä -luottamusvälin tapainen, luottamusväli.

Mikään luottamusväli ei ole yksikäsitteisesti paras, sillä eri asioita voidaan perustellusti painottaa vertailussa. Fagerland ym. (2017, 47) toteavat puutteita Jeffreysin luottamusvälissä; Meeker ym. (2017, 108) suosittelevat sitä.

Edellä nimetyt menetelmät selostetaan Agrestin (2007, 2013, 2019), Bilderin ja Loughinin (2015), Fagerlandin ym:iden (2017), Fleissin ym:iden (2003, jakso 2.2), Newcomben (2013) sekä Meekerin ym:iden (2017) oppikirjoissa. Missään tutkimuksessa ei ole verrattu kaikkia menetelmiä.

Kolmen sääntö on yksisuuntaisen 95 %:n Clopperin–Pearsonin luottamusvälin likiarvoistus. Yksisuuntaisen luottamusvälin laskemiseksi Hernández ym. (2016) suosittelevat Wilsonin luottausväliä jatkuvuuskorjauksella tai toissijaisesti laskennallisesti yksinkertaisempaa Borkowf-luottamusväliä (Borkowf 2006). Luottamustasolla $1 - \alpha$ Borkowf-luottamusvälin ala- ja yläraja ovat

$$\hat{\pi}_L^B - \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_L^B(1-\hat{\pi}_L^B)}{n}}$$

ja

$$\hat{\pi}_U^B + \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_U^B(1-\hat{\pi}_U^B)}{n}}.$$

Yllä $\hat{\pi}_L^B = y/(n+1)$ ja $\hat{\pi}_U^B = (y+1)/(n+1)$. Yksisuuntaista luottamusväliä laskettaessa $z_{1-\alpha/2}$ korvataan $z_{1-\alpha}$:lla. Jos tapahtumia ei ole lainkaan ($y = 0$), Borkowfin menetelmä tuottaa lähes saman 95 %:n luottamusvälin ylärajan kuin Jovanovicin parannettu versio kolmen säännöstä:

$$\frac{0+1}{n+1} + 1.960 \times \sqrt{\frac{\frac{0+1}{n+1}\left(1-\frac{0+1}{n+1}\right)}{n}} = \frac{2.96}{n+1}.$$

Meeker ym. (2017, 108) pitävät (neljästä arvioimastaan menetelmästään) parhaana Jeffreysin menetelmää yksisuuntaisen luottamusvälin laskemisessa.

Paketilla mosaic voi laskea plus neljä -luottamusvälin, Agrestin–Coullin ja Clopperin–Pearsonin luottamusvälit sekä Wilsonin luottamusvälin. Conf-paketti kattaa mm. Blakerin, Jeffreysin ja Wilsonin luottamusvälit sekä Clopperin–Pearsonin luottamusvälin. Paketti PropCIs sisältää mm. plus neljä -luottamusvälin, Agrestin–Coullin ja Clopperin–Pearsonin luottamusvälit, keski-p-korjatun Clopperin–Pearsonin luottamusvälin sekä Blakerin ja Wilsonin luottamusvälit. Keski-p-korjatun Clopperin–Pearsonin luottamusvälin saa myös Anna Gottardin R-koodilla.[118]

10.2.2 Osuuksien erotuksen luottamusväli, jos osuudet ovat riippumattomia

Ollaan kiinnostuneita, kuinka paljon tapahtumien todennäköisyydet π_1 ja π_2 eroavat kahdessa binomijakautuneessa ilmiössä. Oletetaan, että käytettävissä on n_1:n ja n_2:n kokoiset riippumattomat otokset, joiden avulla voidaan estimoida tapahtumien havaitut osuudet $\hat{\pi}_1 = y_1/n_1$ ja $\hat{\pi}_2 = y_2/n_2$ ja niiden erotus $\hat{\pi}_1 - \hat{\pi}_2$. Näissä y_1 ja y_2 ovat tapahtumien lukumäärät otoksissa.

Molemmat osuuden estimaattorit ovat suurilla havaintomäärillä normaalijakautuneita:

$$\hat{\pi}_1 \sim \mathsf{N}(\pi_1, \pi_1(1-\pi_1)/n_1) \quad \text{ja} \quad \hat{\pi}_2 \sim \mathsf{N}(\pi_2, \pi_2(1-\pi_2)/n_2)$$

(jakso 9.3). Koska otokset ovat riippumattomia, erotuksen $\hat{\pi}_1 - \hat{\pi}_2$ varianssi on

$$\mathsf{V}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

(jakso 6.3). Erotuksen keskihajonnan luonteva estimaatti on

$$\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

Erotuksen $100 \times (1-\alpha)$ %:n likimääräisen luottamusvälin ala- ja ylärajat ovat

$$\hat{\pi}_1 - \hat{\pi}_2 \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}, \tag{10.3}$$

koska

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} < \mathsf{z}_{1-\alpha/2}\right) \approx 1-\alpha \Leftrightarrow$$

$$\mathsf{P}\left(\hat{\pi}_1 - \hat{\pi}_2 - \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} < \pi_1 - \pi_2 < \right.$$

$$\left. \hat{\pi}_1 - \hat{\pi}_2 + \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}\right) \approx 1-\alpha.$$

Tämäkin luottamusväli perustuu jakson 9.3 normaalisuuslikiarvoistukseen. Tässä se toimii paremmin kuin yhden osuuden suuruutta arvioitaessa, mutta sen peittävyys tapaa silti olla tarkoitettua pienempi (Agresti ja Caffo 2000, Bilder ja Loughlin 2015, 33–34). Erilaisia peukalosääntöjä likiarvoistuksen kelvollisuudelle on annettu.[119] Ne eivät takaa, että luottamusväli peittäisi parametrien erotuksen (noin) todennäköisyydellä $1-\alpha$, jos π_1 tai π_2 on lähellä nollaa tai yhtä tai havaintoja on vähän.

Huomattavasti paremmin toimivia tapoja muodostaa luottamusväli osuuksien erotukselle on olemassa. Oletetaan, että on estimoitu osuudet $\hat{\pi}_1 = n_{11}/n_1$ ja $\hat{\pi}_2 = n_{21}/n_2$ kahdesta riippumattomasta otoksesta:

	tapahtuma		
	kyllä	ei	Σ
ryhmä 1	n_{11}	n_{12}	n_1
ryhmä 2	n_{21}	n_{22}	n_2

Riuska parannuskeino on lisätä yksi havainto kuhunkin lukumäärään:

	tapahtuma		
	kyllä	ei	Σ
ryhmä 1	$n_{11}+1$	$n_{12}+1$	n_1+2
ryhmä 2	$n_{21}+1$	$n_{22}+1$	n_2+2

Näin muokatusta aineistosta kaavalla (10.3) laskettua luottamusväliä kutsutaan *Agrestin–Caffon luottamusväliksi*. Sen peittävyyden on havaittu olevan lähellä tarkoitettua pienilläkin havaintomäärillä ($n_1 = n_2 = 20$ tai jopa $n_1 = n_2 = 10$) paitsi jos π_i:t ovat molemmat lähellä nollaa tai yhtä (Agresti ja Caffo 2000, Bilder ja Loughlin 2015, 33–34). Samaa muokkausta voidaan käyttää eri luottamustasoilla.

Luottamusväli (10.3) voidaan laskea R-komennolla `prop.test` ja sen `conf.level`-määreellä. Komentoa havainnollistetaan jaksossa 12.1.2 ja harjoitustehtävissä. PropCIs-paketin `wald2ci(n11,n1,n21,n2,conf.level=0.99,adjust="AC")` ja DescTools-paketin `BinomDiffCI(n11,n1,n21,n2,conf.level=0.99,method=c("ac"))` tapaiset komennot palauttavat Agrestin–Caffon luottamusvälin. Harjoitustehtävissä kokeillaan komentoja.

Bilder ja Loughin (2015, 29) suosittelevat Agrestin–Caffon luottamusväliä. Sitä parempi on silti esimerkiksi neliöi ja summaa -Wilson-luottamusväli (*square–and–add*, *hybrid score*; Newcombe 1998b, 2013, Agresti ja Caffo 2000, Fagerland ym. 2015, taulukko 8). Ne ovat molemmat helppolaskuisia. Fagerland ym. (2015, taulukko 8) pitävät parhaana Agrestin–Minin ehdollistamatonta eksaktia luottamusväliä. Mainittujen lisäksi Fagerland ym. (2017, 176) pitävät suositeltavana myös Miettisen–Nurmisen luottamusväliä. Päätekstissä esiteltiin Agrestin–Caffon luottamusväli, koska se on erityisen helppo ja silti pätevä.

10.2.3 Osuuksien erotuksen luottamusväli, jos osuudet eivät ole riippumattomia

Ollaan edelleen kiinnostuneita kahden Bernoulli-kokeen todennäköisyyden vertaamisesta, mutta aineisto koostuu nyt *kaltaistetuista pareista* (*matched pairs*). Kaltaistetuissa pareissa kukin havainto liittyy kahteen kokeeseen, jolloin kokeet eivät ole riippumattomia eikä aineisto koostu kahdesta riippumattomasta otoksesta jakson 10.2.2 tapaan. Pari voi muodostua luonnollisella tavalla saman kohteen kahdesta mittauksesta tai kyselytutkimuksessa, jossa kysytään haastateltavalta kaksi kysymystä ja vastausvaihtoehdot ovat samanlaiset. Pari voidaan myös rakentaa hakemalla tutkittavalle mahdollisimman samanlainen verrokki ja kirjaamalla molemmista saman kokeen tulokset. Jaksossa huomion kohteena ovat kaltaistetut parit kaksiarvoisista satunnaismuuttujista (jatkuva-arvoiset jaksossa 12.4.7).

Esimerkki. Nuoren uskonnollisuus ennen ja jälkeen rippileirin. Kahden autokorjaamoketjun hinnoittelu. Tehdään autoon vika, ja viedään se toisen ketjun korjaamoon. Tehdään sama vika uudestaan samaan autoon, ja viedään se toisen

ketjun korjaamoon. Verrataan tupakoijan keuhkojen hapenottokykyä iältään ja muilta elintavoiltaan mahdollisimman samanlaisen ihmisen keuhkojen hapenottokykyyn. Tällaisista kokeista saadaan havaintoina kaltaistettuja pareja. Mittaukset voivat olla jatkuva- tai kaksiarvoisia (uskonnollinen tai ei-uskonnollinen, kallis tai halpa, suuri tai pieni). □

Tutkitaan kaksiarvoisia satunnaismuuttujia X ja Y:

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

Muuttujien yhtäaikaisiin arvoihin liittyvät todennäköisyydet on järjestetty 2×2 -taulukoksi. Siinä π_{ij} on *solutodennäköisyys* eli todennäköisyys, että X on saanut arvon x_i ja Y on saanut arvon y_j eli että molempien satunnaismuuttujien arvo osuu (i, j)-soluun $(i, j = 1, 2)$. Solut voidaan hahmottaa luokiksi, joiden todennäköisyydet määräävät multinomijakauman $\mathsf{Mul}(1, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$.

Rivien ja sarakkeitten todennäköisyydet on summattu *reunatodennäköisyyksiksi*. Ne kertovat todennäköisyyden, että X saa arvon x_1 tai x_2 (π_{1+} tai π_{2+}) tai Y saa arvon y_1 tai y_2 (π_{+1} tai π_{+2}). Jos x_i:n tai y_i:n alaindeksi "1" viittaa tapahtumaan (ja "2" "vastatapahtumaan"), tapahtumisen todennäköisyydet ovat π_{1+} ja π_{+1} satunnaismuuttujille X ja Y.

Taulukkoa kutsutaan *paritaulukoksi*, jos se liittyy kaltaistettuihin pareihin. Kaltaistettujen parien tilanteessa reunatodennäköisyydet ovat erityisen mielenkiintoisia.

Johdetaan luottamusväli erotukselle $\pi_{1+} - \pi_{+1}$, kun käytettävissä on n:n havainnon satunnaisotos:

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

Tässä n_{ij} on havaintojen lukumäärä (i, j)-solussa. Lukumäärät on summattu riveittäin ja sarakkeittain *reunalukumääriksi* (n_{1+}, n_{2+}, n_{+1} ja n_{+2}).

Ilmeiset estimaatit solu- ja reunatodennäköisyyksille ovat

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n},\ \hat{\pi}_{i+} = \frac{n_{i+}}{n} \quad \text{ja } \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

Erotuksen $\pi_{1+} - \pi_{+1}$ estimaatti on

$$\hat{\pi}_{1+} - \hat{\pi}_{+1} = \frac{n_{1+}}{n} - \frac{n_{+1}}{n} = \frac{n_{11} + n_{12} - (n_{11} + n_{21})}{n} = \frac{n_{12} - n_{21}}{n}.$$

Mitä suurempi erotus $n_{12} - n_{21}$, sitä suurempi erotus $\hat{\pi}_{1+} - \hat{\pi}_{+1}$.

Suurilla havaintomäärillä $\hat{\pi}_{1+} - \hat{\pi}_{+1}$ on normaalijakautunut. Erotuksen varianssia ei voida laskea termien varianssien summana kuten jaksossa 10.2.2: Estimaattorit $\hat{\pi}_{1+}$ ja $\hat{\pi}_{+1}$ koostuvat osin lukumäärästä n_{11}, joten ne eivät ole riippumattomia eivätkä jakson 6.3 laskusäännöt päde. Voidaan osoittaa, että erotuksen estimoitu varianssi on suurilla havaintomäärillä

$$\frac{n_{12} + n_{21} - (n_{12} - n_{21})^2/n}{n^2}$$

(esim. Agresti 2007, 246). Erotuksen $\pi_{1+} - \pi_{+1}$ approksimatiivisen $100 \times (1-\alpha)$ %:n luottamusvälin rajat ovat

$$\hat{\pi}_{1+} - \hat{\pi}_{+1} \pm \mathsf{z}_{1-\alpha/2} \frac{\sqrt{n_{12} + n_{21} - (n_{12} - n_{21})^2/n}}{n}.$$

Esimerkki.[120] Parisuhdeväkivalta. Taulukoissa on lukumäärät ja osuudet kuudes- ja yhdeksäsluokkalaisten lasten havainnoista vanhempiensa toisiinsa kohdistamasta väkivallasta. Taulukon voi ajatella syntyneen kyselytutkimuksesta, jossa kultakin lapselta on kysytty ensin, onko hän nähnyt tai kuullut isän kohdistavan äitiin väkivaltaa, sen jälkeen, onko hän nähnyt tai kuullut äidin kohdistavan isään väkivaltaa ja lopuksi lapsen vastausyhdistelmä on kirjattu yhteen taulukon soluista.

Nähnyt tai kuullut parisuhdeväkivaltaa, joka kohdistuu							
		isään (lkm)			isään (%)		
		kyllä	ei	Σ	kyllä	ei	Σ
äitiin	kyllä	516	674	1190	3.8	5.0	8.8
	ei	231	12038	12269	1.7	89.4	91.2
	Σ	747	12712	13459	5.6	94.4	100

Lapsista 8.8 % ($\hat{\pi}_{1+}$) on aistinut äitiin ja 5.6 % ($\hat{\pi}_{+1}$) isään kohdistunutta väkivaltaa. Osuuksien erotus on 3.3 %-yksikköä. Erotuksen $\pi_{1+} - \pi_{+1}$ 99 %:n luottamusvälin ylä- ja alarajat ovat

$$\begin{aligned}&\frac{674-231}{13459} \pm 2.575829 \frac{\sqrt{674+231-(674-231)^2/13459}}{13459} \\ &= 0.03291478 \pm 0.005710859.\end{aligned}$$

Luottamusväli on noin $[0.027, 0.039]$. Se on laskettu kahdella vaihtoehtoisella tavalla alla. Paketin epibasix komento `mcNemar` käyttää ns. jatkuvuuskorjausta ($\pm 1/n$). Se on eliminoitu alla lisäämällä ja vähentämällä $1/n$ luottamusvälin ala- ja ylärajasta. Näin molemmat laskutavat antavat täsmälleen saman tuloksen.

```
n12 <- 674
n21 <- 231
n <- 13459
z <- qnorm(0.995)
(n12-n21)/n-z*sqrt(n12+n21-(n12-n21)^2/n)/n
(n12-n21)/n+z*sqrt(n12+n21-(n12-n21)^2/n)/n

# Vaihtoehtoinen tapa:
nahnyt <- matrix(c(516,231,674,12038),nrow=2)
install.packages("epibasix")
library(epibasix)
mcNemar(nahnyt,alpha=0.01)$rd.CIL+1/n
mcNemar(nahnyt,alpha=0.01)$rd.CIU-1/n
```

Luottamusväli ei peitä nollaa. Luottamustasolla 0.99 ero on 2.7 − 3.9 %-yksikköä. Lapset ovat havainneet enemmän äitiinsä kuin isäänsä kohdistettua väkivaltaa. Luottamusvälin leveys on 1.2 %-yksikköä. Ero on saatu estimoitua melko tarkasti. □

Pienillä havaintomäärillä edellä kuvatun luottamusvälin peittävyys tapaa olla liian pieni. Agrestin–Minin luottamusväli (Agresti ja Min 2005) lasketaan kuten edellä mutta lisäämällä 0.5 solulukumääriin n_{ij} $(i, j = 1, 2)$ ja 2 otoskokoon n. Näin peittävyys on huomattavasti parempi mutta voi edelleen poiketa tarkoitetusta erityisesti alaspäin. PropCi-paketin komento `diffpropci.mp` palauttaa Agrestin–Minin luottamusvälin.

Fagerland ym. (2017, 384–385) pitävät hieman konservatiivista (jakso 11.3) Bonettin–Pricen luottamusväliä (Bonett ja Price 2012) parhaana. Solutodennäköisyyksien estimaateiksi asetetaan

$$\tilde{\pi}_{12} = \frac{n_{12}+1}{n+2} \quad \text{ja } \tilde{\pi}_{21} = \frac{n_{21}+1}{n+2}$$

ja luottamusvälin rajoiksi

$$\tilde{\pi}_{12} - \tilde{\pi}_{21} \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\tilde{\pi}_{12} + \tilde{\pi}_{21} - (\tilde{\pi}_{12} - \tilde{\pi}_{21})^2}{n+2}}.$$

Fagerland ym. (mt.) suosittelevat sitä laskettaviksi aina otoskoosta riippumatta, koska se on sekä paras että yksinkertainen laskea. Esimerkin edellä tilanteessa se voidaan laskea R-koodilla alla:

```
n12 <- 674
n21 <- 231
n <- 13459
p12 <- (n12+1)/(n+2)
p21 <- (n21+1)/(n+2)
z <- qnorm(0.975)
(p12-p21)-z*sqrt((p12+p21-(p12-p21)^2)/(n+2))
(p12-p21)+z*sqrt((p12+p21-(p12-p21)^2)/(n+2))
```

10.3 Luottamusvälejä havaintojen ollessa Poisson-jakautuneita

10.3.1 Poisson-jakauman odotusarvon luottamusväli

$\mathsf{Poi}(\mu)$-jakautuneesta satunnaismuuttujasta on n riippumatonta havaintoa Y_i. Keskeisen raja-arvolauseen (7.14) perusteella suurilla havaintomäärillä

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sqrt{\mu/n}} < \mathsf{z}_{1-\alpha/2}\right) \approx 1 - \alpha,$$

jossa odotusarvon estimaattori on $\hat{\mu} = \bar{Y} = \sum_{i=1}^{n} Y_i/n$ ja $\mathsf{z}_{\alpha/2}$ ja $\mathsf{z}_{1-\alpha/2}$ on standardinormaalijakauman $\alpha/2$. ja $(1-\alpha/2)$. kvantiili. Sijoitetaan estimaattorin varianssin paikalle sen estimaattori $\hat{\mu}/n$:

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sqrt{\hat{\mu}/n}} < \mathsf{z}_{1-\alpha/2}\right) \approx 1 - \alpha \quad \Leftrightarrow$$

$$\mathsf{P}\left(\hat{\mu} - \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}} < \mu < \hat{\mu} + \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}\right) \approx 1 - \alpha.$$

Poisson-jakauman odotusarvon $100 \times (1-\alpha)$ %:n likimääräisen luottamusvälin ala- ja yläraja ovat

$$\hat{\mu} \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}. \tag{10.4}$$

Myös yhtäpitävää muotoa

$$\hat{\mu}^* \pm \mathsf{z}_{1-\alpha/2}\sqrt{\hat{\mu}^*} = Y \pm \mathsf{z}_{1-\alpha/2}\sqrt{Y}$$

käytetään. Siinä $\mu^* = n\mu$ ja $\hat{\mu}^* = n\hat{\mu} = Y = \sum_{i=1}^{n} Y_i$. Muoto seuraa kertomalla epäyhtälöt yllä n:llä. Tällöin aineisto hahmotetaan yhtenä yhdistettynä otoksena $\mathsf{Poi}(\mu^*)$-jakaumasta.

Tällainen luottamusväli kärsii samantapaisista ongelmista kuin osuuden luottamusväli (10.2). Välin peittävyys on pienillä μ:n ja n:n arvoilla yleensä tarkoitettua pienempi (Bilder ja Loughin 2015, 198–202, Byrne ja Kabaila 2005). Jos tapahtumia ei ole ($\hat{\mu} = \sum_{i=1}^{n} Y_i/n = 0/n = 0$), luottamusväli surkastuu pisteeksi $[0, 0]$. Approksimaation toimivuus paranee μ:n ja n:n kasvaessa. Molempien tulisi olla melko suuria tai n:n erityisen suuri. Peukalosääntö toimivuudelle on, että $n\mu > 100$ (Armitage ym. 2002, 154).

Esimerkki.[121] Rintasyöpä. Kohorttitutkimuksissa seurataan tyypillisesti kahta ihmisryhmää, joista toinen altistuu riskille ja toinen ei. Monesti altistumista mitataan henkilövuosilla eli otosten koolla kerrottuna seuranta-ajalla.

Tuberkuloosia sairastaneiden naisten keuhkoja on tutkittu röntgenillä fluoresoivan varjoaineen avulla (*fluoroscopy*). Seuraavien 28 010 henkilövuoden aikana 41 naista sairastui rintasyöpään. Vertailuryhmässä, jonka keuhkoja ei oltu tutkittu, naisille kehittyi 15 rintasyöpää 19 017 henkilövuoden aikana. Estimoidaan syövän ilmaantuvuudet tuhatta henkilövuotta kohden: $41/(28010/1000) = 41/28.01 = 1.463763$ ja $15/(19017/1000) = 15/19.017 = 0.7887679$. Mallitetaan sairastuneiden lukumääriä $\mathsf{Poi}(\mu_i)$-jakaumalla ($i = 1, 2$). Estimoitu ilmaantuvuus tuhatta henkilövuotta kohden vastaa $\hat{\mu}$:a ja tuhannet henkilövuodet otoskokoa n kaavassa (10.4). Otoskoko ei ole kokonaisluku aineiston muodostamistavasta johtuen.

Ensimmäisessä otoksessa 95 %:n luottamusvälin μ_1:lle rajat ovat

$$1.463763 \pm 1.959964 \times \sqrt{1.463763/28.01} = 1.463763 \pm 0.4480505.$$

Luottamusväli on noin $[1.02, 1.91]$. Vastaavan luottamusvälin μ_2:lle rajat ovat

$$0.7887679 \pm 1.959964 \times \sqrt{0.7887679/19.017} = 0.7887679 \pm 0.3991643.$$

Luottamusväli on noin $[0.39, 1.19]$.

Peukalosääntö odotusarvojen luottamusvälien käyttökelpoisuudelle ei toteudu: $28.01 \times 1.463763 = 41 < 100$ ja $19.017 \times 0.7887679 = 15 < 100$. Vaikka havaintoja on paljon, on syöpäriski pieni. Niiden tulo ei ole tarpeeksi suuri takaamaan normaalisuuslikiarvoistuksen toimivuutta. □

Meeker ym. (2017, 131) näkevät luottamusvälille (10.4) käyttöä vain tilanteisiin, joissa tarvitaan yksinkertaista karkeaa laskua. Poisson-jakauman odotusarvon luottamusväli voidaan laskea monilla paremmilla tavoilla (Barker 2002, Byrne ja Kabaila 2005, Patil ja Kulkarni 2012). Yhtä yksinkertaista ja toimivaa parannusta kuin plus neljä -luottamusväli (osuutta väliestimoitaessa) ei ole kehitetty.

Monissa yhteyksissä käytetty jatkuvuuskorjaus on helppo ja tuottaa paremman peittävyyden (Byrne ja Kabaila 2005). Laskennallisesti helpoimpia on skooriin (*score*) perustuva luottamusväli (esim. Agresti ja Coull 1998, Andersson 2015, Bilder ja Loughin 2015, 198–202, Byrne ja Kabaila 2005, Davison 2003, Meeker ym. 2017, 132, Newcombe 2013, luku 6, Swift 2009, Stuart ja Ord 1991, 755): Jos $Y \sim \mathsf{Poi}(\mu^*)$, niin suurilla havaintomäärillä

$$\frac{Y - \mu^*}{\sqrt{\mu^*}}$$

on standardinormaalijakautunut. Tällöin $1 \times (1 - 2\alpha)$ %:n luottamusvälin ala- ja yläraja saadaan ratkaisuna yhtälöstä

$$(Y - \mu^*)^2 = \mathsf{z}_{1-\alpha/2}\mu^*.$$

Myös voidaan laskea eksakti luottamusväli, jonka peittävyys tiedetään liian suureksi (Agresti ja Coull 1998, Barker 2002, Casella ja Berger 2002, 434–435, Fleiss, Levin ja Paik 2013, 342, Meeker ym. 2017, 129–130, Pawitan 2013, 134–135) tai keski-p-korjattu eksakti luottamusväli (Byrne ja Kabaila 2005, Cohen ja Young 1994, Newcombe 2013, luku 6). Vaihtoehtoja on muitakin. Bilder ja Loughin (2015, 198–202) suosittelevat yllä esitettyä skooriluottamusväliä: Se on parempi kuin luottamusväli (10.4) eikä muiden menetelmien edut siihen verrattuna ole suuria. Newcombe (2013, 120) on kriittisempi. Newcombe (mts. 122) laskee keski-p-korjatut eksaktit luottamusvälit rintasyöpäesimerkin aineistolla. Välit poikkeavat jonkin verran edellä lasketusta ($[1.03, 1.97]$ ja $[0.46, 1.27]$). Meeker ym. (2017, 133) suosittelevat Jeffreysin luottamusväliä.

10.3.2 Riippumattomien Poisson-jakautuneiden satunnaismuuttujien odotusarvojen erotuksen luottamusväli

Käytettävissä on kaksi riippumatonta otosta $\mathsf{Poi}(\mu_i)$-jakautuneista satunnaismuuttujista, $i = 1, 2$. Estimaattorit $\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ ovat normaalijakautuneita

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\mu_i/n_i}} \sim \mathsf{N}(0, 1)$$

suurilla havaintomäärillä (n_i) edellisen jakson tapaan. Edellä Y_{ij} on j. havainto satunnaismuuttujasta i. otoksessa. Erotuksen $\hat{\mu}_1 - \hat{\mu}_2$ varianssi on riippumattomuuden perusteella varianssien summa $\mu_1/n_1 + \mu_2/n_2$ (jakso 6.3). Erotuksen

$100 \times (1-\alpha)$ %:n luottamusväli voidaan jälleen perustaa normaalisuuteen:

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}}} < \mathsf{z}_{1-\alpha/2} \right) \approx 1 - \alpha \quad \Leftrightarrow$$

$$\mathsf{P}\left(\hat{\mu}_1 - \hat{\mu}_2 - \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}} < \mu_1 - \mu_2 < \right.$$

$$\left. \hat{\mu}_1 - \hat{\mu}_2 + \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}} \right) \approx 1 - \alpha.$$

Luottamusvälin rajat ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\hat{\mu}_1}{n_1} + \frac{\hat{\mu}_2}{n_2}}. \tag{10.5}$$

Vaihtoehtoinen yhtäpitävä muotoilu on taas mahdollinen. Erotuksen $\hat{\mu}_1 - \hat{\mu}_2$ komponentit ovat riippumattomia ja likimain pätee $\hat{\mu}_1 \sim \mathsf{N}(\mu_1, \mu_1/n_1)$, $\hat{\mu}_2 \sim \mathsf{N}(\mu_2, \mu/n_2)$, $Y_1 \equiv \sum_{j=1}^{n_1} Y_{1j} = n_1\hat{\mu}_1 \sim \mathsf{N}(n_1\mu_1, n_1\mu_1)$, $Y_2 \equiv \sum_{j=1}^{n_2} Y_{2j} = n_2\hat{\mu}_2 \sim \mathsf{N}(n_2\mu_2, n_2\mu_2)$ ja $\mathsf{V}(Y_1 - Y_2) = n_1\mu_1 + n_2\mu_2$. Merkitään $\mu_i^* = n_i\mu_i$. Yllä olevaan tapaan erotuksen $\mu_1^* - \mu_2^*$ luottamusvälin luottamustasolla $1-\alpha$ rajoiksi saadaan

$$\hat{\mu}_1^* - \hat{\mu}_2^* \pm \mathsf{z}_{1-\alpha/2}\sqrt{\hat{\mu}_1^* + \hat{\mu}_2^*} = Y_1 - Y_2 \pm \mathsf{z}_{1-\alpha/2}\sqrt{Y_1 + Y_2}.$$

(Y_1:n ja Y_2:n merkitys on tässä eri kuin jaksossa 10.3.1.)

Rajojen (10.5) määrittelemä luottamusväli toimii hyvin, jos $\mu_1^* = n_1\mu_1 > 2$, $\mu_2^* = n_2\mu_2 > 2$ ja $n_1 = n_2$. Muulloin välin peittävyys voi olla paljon pienempi kuin sen nimellinen peittävyys $1 - \alpha$. (Krishnamoorthy ja Lee 2013.) Waldin periaatteella johdettu luottamusväli Poisson-odotusarvojen erotukselle voi siten olla luotettava tilanteessa, jossa vastaavat luottamusvälit (10.4) yksittäisille Poisson-odotusarvoille eivät ole luotettavia.

Esimerkki. Rintasyöpä (jatkoa). Erotuksen $\mu_1 - \mu_2$ 95 %:n luottamusvälin rajat ovat

$$1.463763 - 0.7887679 \pm 1.959964 \times \sqrt{1.463763/28.01 + 0.7887679/19.017}$$
$$= 0.6749951 \pm 0.6000678.$$

Luottamusväli on noin $[0.07, 1.28]$. Se ei peitä nollaa. Luottamusvälin mukaan röntgenillä fluoresoivan varjoaineen avulla tutkitut naiset sairastuvat useammin

rintasyöpään kuin tutkimattomat naiset. Luottamusväli on (tutkittavan asian kannalta) leveä, joten eron suuruutta ei ole saatu selvitettyä tarkasti.

Luottamusvälin käyttökelpoisuuden ehdoista kaksi ensimmäistä toteutuvat kirkkaasti: $28.01 \times 1.463763 = 41 > 2$ ja $19.017 \times 0.7887679 = 15 > 2$. Kolmas ehto samansuuruisista otoksista ei päde.

Erotuksen $\mu_1 - \mu_2$ luottamusväli ei kata nollaa, vaikka μ_1:n ja μ_2:n luottamusvälit $[1.02, 1.91]$ ja $[0.39, 1.19]$ menevät toistensa päälle (jakso 10.3.1). Luottamusvälien lomittumisesta ei pidä päätellä, että odotusarvot eivät eroaisi. Asiaan palataan jaksossa 11.3. □

Jos ehdot $n_1\mu_1 > 2$, $n_2\mu_2 > 2$ ja $n_1 = n_2$ eivät täyty, voi olla parempi käyttää muita tekniikoita (Li ym. 2011, Krishnamoorthy ja Lee 2013, Krishnamoorthy 2016, 106–108). Khrishnamoorthy (2016, 107) pitää puntaroimistaan menetelmistä parhaimpana skooriluottamusväliä, jos $n_1\mu_1 > 2$ ja $n_2\mu_2 > 2$.

Li ym. (2011) laskevat luottamusvälejä eri tekniikoilla Poison-odotusarvojen erotukselle esimerkin rintasyöpäaineistolle. Kaikkien luottamusvälien mukaan odotusarvot eroavat. Ng ym. (2007) tutkivat vaihtoehtoisia piste-estimaattoreita ja arvioivat syöpäriskin eron suuruutta esimerkin aineiston avulla.

10.4 Luottamusvälejä havaintojen ollessa normaalijakautuneita

Edellä vertailut perustettiin keskeiseen raja-arvolauseeseen, joka takaa keskiarvon normaalisuuden suurilla otoskoilla. Seuraus oli, että pienillä havaintomäärillä luottamusvälien todellinen peittävyys ei välttämättä ollut tarkoitetunlainen. Tässä jaksossa oletetaan, että havainnot ovat normaalijakautuneita. Tällöin luottamusvälit voidaan laskea niin, että niiden peittävyys on täsmälleen oikea kaikilla havaintomäärillä. Jaksossa selitetään luottamusvälin lasku alkaen yksinkertaisimmasta empiirisesti epärelevanteimmasta tilanteesta edeten kohti monimutkaisinta empiirisesti relevanteinta tilannetta. Polku on pedagogisesti hyödyllinen, mutta kiireisimmät tutustuvat vain jaksoihin 10.4.2 ja 10.4.6.

10.4.1 Normaalijakauman odotusarvon luottamusväli, jos varianssi tunnetaan

Jos $X_i \sim \mathsf{N}(\mu, \sigma^2)$, niin

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0, 1).$$

Yllä $\bar{X} = \sum_{i=1}^{n} X_i/n$. Tällöin pätee eksaktisti

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \mathsf{z}_{1-\alpha/2}\right) = 1 - \alpha.$$

Siinä $\mathsf{z}_{\alpha/2}$ ja $\mathsf{z}_{1-\alpha/2}$ ovat standardinormaalijakauman kvantiileja (esim. $\mathsf{z}_{0.005} = -\mathsf{z}_{0.995} = -2.576$). Jos σ^2 tunnetaan, voidaan johtaa ja laskea jakson 10.2.1 tapaan μ:lle $100 \times (1 - \alpha)$ %:n luottamusväli

$$\hat{\mu} \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}. \tag{10.6}$$

Siinä $\hat{\mu} = \bar{X}$. Kuvan 10.1 luottamusvälit on laskettu tällä kaavalla. Kukin luottamusväli kuvassa on yhtäpitkä, koska varianssi on tunnettu ja siten termi $\mathsf{z}_{1-\alpha/2}\sqrt{\sigma^2/n}$ on vakio.

10.4.2 Normaalijakauman odotusarvon luottamusväli, jos varianssia ei tunneta

Useimmiten varianssia σ^2 ei tunneta. Estimoidaan se kaavalla $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ (jaksot 9.1 ja 9.6). Voidaan osoittaa, että standardoitu tunnusluku

$$\frac{\hat{\mu} - \mu}{s/\sqrt{n}} \sim \mathsf{t}(n-1)$$

noudattaa t-jakaumaa $n - 1$ vapausasteella. Jälleen saadaan yhtäsuuruus

$$\mathsf{P}\left(\mathsf{t}_{\alpha/2}(n-1) < \frac{\hat{\mu} - \mu}{s/\sqrt{n}} < \mathsf{t}_{1-\alpha/2}(n-1)\right) = 1 - \alpha.$$

Ero edelliseen todennäköisyyslaskuun on, että nyt standardoitu satunnaismuuttuja noudattaa $\mathsf{t}(n-1)$-jakaumaa ja epäyhtälöissä on siksi $\mathsf{t}(n-1)$-jakauman $\alpha/2$. ja $(1 - \alpha/2)$. kvantiilit $\mathsf{t}_{\alpha/2}(n-1)$ ja $\mathsf{t}_{1-\alpha/2}(n-1)$. Luottamusväliksi μ:lle luottamustasolla $1 - \alpha$ saadaan

$$\hat{\mu} \pm \mathsf{t}_{1-\alpha/2}(n-1) \times \sqrt{\frac{s^2}{n}}. \tag{10.7}$$

Koska varianssi estimoidaan, termin $\mathsf{t}_{1-\alpha/2}(n-1) \times \sqrt{s^2/n}$ suuruus ja siten luottamusvälin pituus vaihtelevat otoksissa.

Esimerkki. Varianssin estimoinnin hinta. Jos varianssi estimoimaan, luottamusväli levenee. Olkoot $X_i \sim \mathsf{N}(\mu, 1)$ ja $n = 20$. Jos varianssi tunnetaan, 95 %:n luottamusvälin odotusarvolle rajat ovat

$$\hat{\mu} \pm 1.960\sqrt{\frac{1}{20}} = \hat{\mu} \pm 0.4382613.$$

(kaava (10.6)). Luottamusvälin leveys on $2 \times 0.4382613 = 0.8765225$.

Estimoidaan varianssi, ja saadaan ihmeen kautta estimaattorin odotusarvo oikea arvo $s^2 = 1$ (jakso 9.1). 95 %:n luottamusvälin odotusarvolle rajat ovat nyt

$$\hat{\mu} \pm 2.093\sqrt{\frac{1}{20}} = \hat{\mu} \pm 0.4680144$$

(kaava (10.7)). 0.975. kvantiili `t(19)`-jakaumasta 2.093 on laskettu R:n käskyllä `qt(0.975,19)`. Luottamusvälin leveys on $2 \times 0.4680144 = 0.9360288$.

Jälkimmäinen luottamusväli on $0.9360288 - 0.8765225 \approx 0.060$ verran edellistä leveämpi. Se on hinta tietämättömyydestä varianssin suuruudesta eli sen estimoinnista, kun väliestimoidaan odotusarvoa. □

Esimerkki. Miesten keskipituus.[122] Suomeen luotiin 2010 uudet kasvukäyrät, jotka kuvaavat lasten ja nuorten kasvua syntymästä aikuisuuteen. Kasvukäyräaineistossa miehen pituuden otoskeskiarvo ja -hajonta ovat 181.042 cm ja 6.0609 cm. Klassinen empiirinen esimerkki normaalijakautuneesta satunnaismuuttujasta on miesten pituus, joten oletetaan se normaalijakautuneeksi. Oletetaan, että aineisto koostuu 4 000 miehestä ja että havainnot ovat riippumattomia. Lasketaan 95 %:n luottamusväli miesten keskipituudelle. R-käsky `qt(0.975,3999)` laskee `t(3999)`-jakauman 0.975. kvantiiliksi 1.960557. Kvantiili on likipitäen standardinormaalijakauman 0.975. kvantiili 1.959964 (`qnorm(0.975)`), koska vapausasteita on paljon (jakso 7.2.3). Luottamusvälin rajat saadaan kaavasta (10.7):

$$181.042 \pm 1.960557 \times \frac{6.0609}{\sqrt{4000}} = 181.042 \pm 0.1878826.$$

Miesten keskipituuden 95 %:n luottamusvälin (180.9 cm, 181.2 cm) leveys on vajaa 4 mm (pyöristämättömistä rajoista). Keskipituus on saatu estimoitua varsin tarkasti. □

Komento `t.test(x, conf.level=0.95)` palauttaa 95 %:n luottamusvälin, jos aineisto `x` on jo R:ssä.

10.4.3 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit yhtäsuuria ja tunnetaan

Kiinnostuksen kohteena on kahden normaalijakautuneen satunnaismuuttujan X_1:n ja X_2:n odotusarvon erotus. Oletetaan idealisoitu tilanne, jossa jakaumien varianssit tunnetaan ja ne ovat samat: $X_{1j} \sim \mathsf{N}(\mu_1, \sigma^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma^2)$. Populaatioista on poimittu n_1:n ja n_2:n suuruiset riippumattomat otokset. Keskiarvojen erotuksen $\hat{\mu}_1 - \hat{\mu}_2 = \bar{X}_1 - \bar{X}_2 = \sum_{j=1}^{n_1} X_{1j}/n_1 - \sum_{j=1}^{n_2} X_{2j}/n_2$ varianssi on $\sigma^2/n_1 + \sigma^2/n_2 = \sigma^2(1/n_1 + 1/n_2)$. Tällöin

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim \mathsf{N}(0, 1),$$

ja

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} < \mathsf{z}_{1-\alpha/2}\right) = 1 - \alpha.$$

$100 \times (1 - \alpha)$ %:n luottamusvälin erotukselle $\mu_1 - \mu_2$ rajat ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm \mathsf{z}_{1-\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{10.8}$$

10.4.4 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit erisuuria ja tunnetaan

Oletetaan edellisen jakson tilanne paitsi, että jakaumien varianssit ovat erisuuria: $X_{1j} \sim \mathsf{N}(\mu_1, \sigma_1^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma_2^2)$. Varianssit tunnetaan. Nyt

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathsf{N}(0, 1).$$

$100 \times (1 - \alpha)$ %:n luottamusvälin erotukselle $\mu_1 - \mu_2$ rajat ovat vastaavasti

$$\hat{\mu}_1 - \hat{\mu}_2 \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{10.9}$$

10.4.5 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit yhtäsuuria ja tuntemattomia

Palataan yhtäsuurien varianssien tilanteeseen $X_{1j} \sim \mathsf{N}(\mu_1, \sigma^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma^2)$. Uusi realistinen piirre on, että varianssia σ^2 ei tunneta. Estimoidaan se

molempien — edelleen riippumattomaksi oletettujen — otosten avulla:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_{j=1}^{n_1}(X_{1j}-\hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(X_{2j}-\hat{\mu}_2)^2}{n_1+n_2-2}.$$

Yllä $s_i^2 = \sum_{j=1}^{n_i}(X_{ij}-\hat{\mu}_i)^2/(n_i-1)$, $i = 1, 2$. Voidaan osoittaa, että tällöin

$$\mathsf{P}\left(\mathsf{t}_{\alpha/2}(n_1+n_2-2) < \frac{\hat{\mu}_1-\hat{\mu}_2-(\mu_1-\mu_2)}{s\sqrt{1/n_1+1/n_2}} < \mathsf{t}_{1-\alpha/2}(n_1+n_2-2)\right) = 1-\alpha.$$

Siinä $\mathsf{t}_{\alpha/2}(n_1+n_2-2)$ ja $\mathsf{t}_{1-\alpha/2}(n_1+n_2-2)$ ovat t-jakauman n_1+n_2-2 vapausasteella $\alpha/2$. ja $(1-\alpha/2)$. kvantiilit. Luottamustasolla $1-\alpha$ luottamusvälin erotukselle $\mu_1-\mu_2$ rajat ovat

$$\hat{\mu}_1-\hat{\mu}_2 \pm \mathsf{t}_{1-\alpha/2}(n_1+n_2-2)s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Koska t-jakauma on paksuhäntäisempi kuin standardinormaalijakauma, tämä luottamusväli on leveämpi kuin kaavan (10.8) rajaama luottamusväli kuvitteellisessa tilanteessa $s = \sigma$. Luottamusväli levenee, kun satunnaismuuttujista tiedetään vähemmän. R-komento `t.test(x1,x2,var.equal=TRUE, conf.level=0.95)` (`x1`:n ja `x2`:n sisältäessä otosten havainnot) palauttaa tässä kuvatun 95 %:n luottamusvälin.

10.4.6 Normaalijakaumien odotusarvojen erotuksen luottamusväli, jos varianssit erisuuria ja tuntemattomia

Olkoot $X_{1j} \sim \mathsf{N}(\mu_1, \sigma_1^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma_2^2)$. Populaatioista on poimittu n_1:n ja n_2:n suuruiset riippumattomat otokset.

Realistisin tilanne on, että variannssit σ_1^2 ja σ_2^2 ovat tuntemattomia ja mahdollisesti erisuuria. Ilmeiset estimaatit niille ovat $s_i^2 = \sum_{j=1}^{n_i}(X_{ij}-\hat{\mu}_i)^2/(n_i-1)$, $i = 1, 2$. Toisin kuin muualla jaksossa 10.4, nyt joudutaan tyytymään likimääräiseen luottamusväliin. Edellisten jaksojen tapaan muotoillun tunnusluvun

$$\frac{\hat{\mu}_1-\hat{\mu}_2-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$$

jakauma riippuu varianssien suhteesta σ_1^2/σ_2^2 sekä otoskoista n_1 ja n_2 (ns. Behrensin–Fisherin jakauma). Voidaan osoittaa, että tunnusluku yllä noudattaa

likimääräisesti $\mathsf{t}(\nu)$-jakaumaa, jossa

$$\nu = \mathsf{int}\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]. \tag{10.10}$$

Yllä $\mathsf{int}[x]$ on argumentin x kokonaislukuosa (esim. $\mathsf{int}[36.51] = 36$). Likimääräisen $100 \times (1-\alpha)$ %:n luottamusvälin rajat erotukselle $\mu_1 - \mu_2$ ovat

$$\hat{\mu}_1 - \hat{\mu}_2 \pm \mathsf{t}_{1-\alpha/2}(\nu)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}. \tag{10.11}$$

Kaavaa yllä kutsutaan tässä *Smithin–Welchin–Satterthwaiten luottamusväliksi*, mutta sen nimitys vaihtelee kirjallisuudessa.

Voidaan osoittaa, että $\nu \leq n_1 + n_2 - 2$ (Dudewicz ja Mishra 1988, 502), joten $\mathsf{t}_{1-\alpha/2}(\nu) \geq \mathsf{t}_{1-\alpha/2}(n_1+n_2-2)$ ja kaavan (10.11) mukainen luottamusväli on yleensä leveämpi kuin verrannollisessa tilanteessa $s_1^2 = \sigma_1^2$ ja $s_2^2 = \sigma_2^2$ kaavasta (10.9) saatava. Jälleen oletuksista luopuminen venyttää luottamusväliä.

Jos molemmissa otoksissa havaintoja on paljon, $\mathsf{t}_{1-\alpha/2}(\nu)$:n voi korvata luottamusvälissä $\mathsf{z}_{1-\alpha/2}$:lla. Tällöin kaavojen (10.9) ja (10.11) mukaiset luottamusvälit yhtyisivät edellä mainitussa kuvitteellisessa verrannollisessa tilanteessa.

Esimerkki.[123] Rikollisuus ja leposyke. Latvala ym. (2015 ja 2016) havaitsivat yli 700 000:n ja 1 000 000:n havainnon aineistoissa asevelvollisuusikäisillä miehillä yhteyden alhaisen leposykkeen ja taipumuksen myöhemmällä iällä väkivaltaiseen rikollisuuteen välillä. Choyn ym:iden (2017) aineistossa jo 11-vuotiaana alhaisen leposykkeen omaavat olivat 23-vuotiaana syyllistyneet rikoksiin muita useammin. Aineistossa 23 vuoden ikään mennessä rikoksesta tuomittujen ja tuomitsemattomien miesten ja naisten leposykkeiden 11-vuotiaana otoskeskiarvot ja -hajonnat ovat alla.

	pojat		tytöt	
leposykkeen	tuomio	ei tuomiota	tuomio	ei tuomiota
otoskeskiarvo	86.66	89.61	97.72	98.17
otoskeskihajonta	12.74	13.60	12.90	15.52
n	271	226	44	353

Tuomittujen lepopulssi on aineistossa verkkaisempi molemmilla sukupuolilla. Lasketaan likimääräinen 99 %:n luottamusväli 23-vuoden ikään mennessä tuomittujen ja tuomitsemattomien 11-vuotiaiden poikien lepopulssien odotusarvojen erotukselle. Oletetaan, että poikien leposyke on normaalijakautunut. Tarvittava vapausasteluku on

$$\nu = \mathsf{int}\left[\frac{\left(\frac{12.74^2}{271}+\frac{13.60^2}{226}\right)^2}{\frac{\left(\frac{12.74^2}{271}\right)^2}{271-1}+\frac{\left(\frac{13.60^2}{226}\right)^2}{226-1}}\right] = \mathsf{int}[466.5826] = 466.$$

Luottamusvälin rajat ovat

$$86.66 - 89.61 \pm 2.586421\sqrt{\frac{12.74^2}{271}+\frac{13.60^2}{226}} = -2.95 \pm 3.079175.$$

Siinä 2.586421 on t(466)-jakauman 0.995. kvantiili (`qt(0.995,466)`). Luottamusväli $(-6.039, 0.129)$ peittää nollan.

Luottamusväli on lähes sama, jos se lasketaan käyttäen standardinormaalijakauman 0.995. kvantiilia 2.575829 (`qnorm(0.995)`): $(-6.016, 0.117)$. Näinkin laskettu väli peittää nollan. Luottamusvälit ovat yhtenevät, koska ν:n ollessa suuri, t- ja standardinormaalijakauman kvantiilit ovat liki samat.

Luottamusvälit eivät anna vahvaa perustetta argumentoida, että rikoksista tuomittujen ja tuomitsemattomilla pojilla ei voisi olla 11-vuotiaana sama leposyke. Yksi mahdollinen selitys on, että Choylla ym:illa on vähemmän havaintoja kuin Latvalalla ym:illa. Se suurentaa otosvariansseja ja luottamusvälejä. Choyn ym:iden aineistoon palataan harjoitustehtävässä. □

Yleensä variansseja σ_1^2 ja σ_2^2 ei tunneta eikä niiden yhtäsuuruudesta ole selkeää tietoa. Luottamusväli odotusarvojen erotukselle tulisi siksi lähtökohtaisesti aina laskea jaksossa kuvatulla menettelyllä, jos havainnot ovat normaalijakautuneita ja riippumattomia. R-komento `t.test(x1,x2, var.equal=FALSE, conf.level=0.95)` palauttaa juuri kuvatun 95 %:n luottamusvälin.

10.5 Luottamusvälejä ilman jakaumaoletusta

10.5.1 Odotusarvon luottamusväli, jos jakauma on tuntematon

Olkoot havainnot X_i riippumattomia ja jatkuva-arvoisia mutteivät välttämättä normaalijakautuneita. Tukeudutaan jakson 7.3 keskeiseen raja-arvolauseeseen: Suurilla havaintomäärillä

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0, 1),$$

vaikka havainnoista ei tiedettäisi juuri muuta kuin, että ne ovat riippumattomia ja että niillä on odotusarvo (μ) ja varianssi (σ^2). Estimoidaan odotusarvo ($\hat{\mu} = \bar{X} = \sum_{i=1}^{n} X_i/n$) ja varianssi ($s^2$ tai $\hat{\sigma}^2$). Luottamustason $1 - \alpha$ luottamusvälin rajat ovat

$$\hat{\mu} \pm \mathsf{z}_{1-\alpha/2}\sqrt{\frac{s^2}{n}}. \tag{10.12}$$

Siinä on käytetty s^2:sta varianssin estimaattorina.

Luottamusvälit (10.12) ja (10.7) yhtyvät suurilla otoskoilla. Ylipäänsä luottamusväli (10.12) on kapeampi kuin (10.7), koska $\mathsf{z}_{1-\alpha/2} < \mathsf{t}_{1-\alpha/2}(n-1)$. Ei ole mielekästä, että tietämättömyydestä jakaumasta seuraisi kapeampi luottamusväli. Luottamusvälin laskeminen kaavalla (10.7) on siksi lähtökohtaisesti suositeltavaa, jos havaintojen jakaumaa ei tiedetä.

Luottamusväli (10.7) toimii ylipäänsä varsin hyvin, jos havainnot eivät tule hyvin vinosta jakaumasta ja $n \geq 15$ (Agresti ja Finlay 2009, 122). Yleensä $25 - 40$ havainnon pitäisi takaa luottamusvälin (10.7) toimivuus (Wilcox 2012, 135).

Wilcox (2012, 139–140) argumentoi, että paksuhäntäiset jakaumat voivat johtaa liian leveisiin luottamusväleihin pienillä havaintomäärillä. Liian kapea tai leveä luottamusväli saattaa seurata vinosta jakaumasta. Paksuhäntäinen vino jakauma voi vaatia erityisen paljon havaintoja, jotta luottamusväli olisi pätevä. Tilanteesta riippuen saatetaan tarvita 200–300 havaintoa, jotta luottamusvälin (10.7) peittävyys olisi oikeanlainen. (Wilcox 2012, 138–147 ja 200.) Erityisesti oudokit vääristävät odotusarvon luottamusvälin laskua (Chihara ja Hesterberg 2019, 197–198).

10.5.2 Odotusarvojen erotuksen luottamusväli, jos jakauma on tuntematon

Kahdesta populaatiosta on poimittu n_1:n ja n_2:n kokoiset riippumattomat otokset jatkuva-arvoisista satunnaismuuttujista X_1 ja X_2. Niillä on mahdollisesti erisuuret odotusarvot ja varianssit. Muuta jakaumista ei tiedetä. Odotusarvojen erotukselle voi tällöinkin laskea luottamusvälin Smithin–Welchin–Satterthwaiten kaavalla (10.11). Sen pitäisi olla likimäärin pätevä keskeisen raja-arvolauseen (jakso 7.3) perusteella, jos havaintoja on riittävästi. Peukalosääntö toimivuudelle on, että $n_1 \geq 30$ ja $n_2 \geq 30$.

Hyvin vinot satunnaismuuttujien X_1 ja X_2 jakaumat tai oudokit voivat tehdä luottamusvälistä (10.11) epäluotettavan. (Agresti ja Finlay 2009, 191.) Vinouden voi odottaa olevan kuitenkin vähemmän ongelmallista odotusarvojen erotuksen kuin odotusarvon luottamusväliä laskettaessa, jos ryhmien otokset ovat samallalailla vinoja ja suhtsamankokoisia (Chihara ja Hesterberg 2019, 201–202).

10.6 Korrelaation luottamusväli

Agresti ja Finlay (2009, 299), Anderson (2003, 135), Fagerland ym. (2017, 317–318), Hu ym. (2020), Krishnamoorthy (2016, jakso 30.3), Lindgren (1976, 478–479), Meeker ym. (2017, 241–242) sekä Stuart ja Ord (1991, 984–985) selittävät, kuinka korrelaatiokertoimelle voi laskea likimääräisen tai eksaktin luottamusvälin, kun satunnaismuuttujat ovat binormaalijakautuneita (jakso 7.5). R-komento `cor.test(x,y)` tukeutuu kaavaan (12.11) ($n \geq 4$) ja palauttaa likimääräisen luottamusvälin.

Luku 11

Testiteoriaa

Kuva 11.1: Henrik Karlssonin piirros Helsingin Sanomissa 28.9.2019.[1]

[1]Kuva: `https://www.hs.fi/karlsson/car-2000006252771.html` (haettu 5.10.2019). Piirroksen taustalla on lyhyen matematiikan ylioppilaskokeen 24.9.2019 joitakin kokelaita hämmentänyt kysymys. Kiitän Henrik Karlssonia luvasta käyttää kuvaa.

> Ei ole olemassa rutiininomaisia tilastollisia ongelmia — on vain ongelmallisia tilastollisia rutiineja.[124]
>
> David Cox (1924–2022)

Usein tutkimuksessa ja varsinkin käytännön elämässä, vaikkapa liike-elämässä, täytyy tehdä selkeitä ratkaisuja ja päätöksiä. Testillä saa selkeän kyllä- tai ei-vastauksen päätöksentekoon. (Laber ja Shedden 2018, Benjamini ym. 2021.) Selkeys ilmeisesti viehättää soveltajia ja selittää testien suosiota. Testit ovat käytetyimpiä tilastollisen päättelyn työkaluja.

Testejä voidaan käyttää hienovaraisemmin, mikä on usein suositeltavaa. Testin tuloksesta ei ole välttämätöntä tehdä yksiselitteistä johtopäätöstä. On järkevämpää käyttää kaikkea tietoa asiasta johtopäätösten tekemiseen. Yksiselitteistä johtopäätöstä ei ole aina myöskään tarpeen tehdä. Ylipäänsä testit selkiyttävät ja antavat ryhtiä empiiriselle analyysille. Yksi suurimmista tilastotieteilijöistä, David Cox, muotoilee testien merkityksen näin: Merkitsevyystestien tavoite ei ole korvata tutkijan henkilökohtaista harkintaa vaan selventää hänelle ja muille, mitä aineisto osoittaa (Cox 1977, 61).

Karl Pearsonin (1900) χ^2-testi (jakso 12.2) ja sen ohessa esittämä p-arvo (jakso 11.2) ohjasivat tilastotiedettä voimallisesti testaamista painottavaan suuntaan. Ronald Fisher sekä työpari Jerzy Neyman ja Egon Pearson kehittivät testiteoriaa 1900-luvun alkupuolella. Teoria alla on synteesi edellisten näkemyksistä. Fisheriläinen perinne painottaa testisuureen p-arvoa tilastollisen päättelyn yhtenä työkaluna; Neymanin–Pearsonin traditio korostaa testaamista mekanistisempana päätöksentekomenetelmänä. Ensin mainittu perinne on suositeltavampi.

Termien merkitys ei ole täysin vakiintunut kirjallisuudessa. Käsitteet selitetään yksinkertaisimmassa yhden parametrin tilanteessa. (Parametri voi olla kahden parametrin erotus tai muu niiden välinen rajoitus.) Myös oletettua tilastollista mallia ylipäänsä voidaan testata. Siitä on esimerkkejä jaksoissa 12.2 ja 12.3.

11.1 Merkitsevyystestaus

Testaus tehdään useimmiten tilastollisen mallin puitteissa (jakso 9.2). *Nollahypoteesi* (H_0; *null hypothesis*) on testattava tilastollista mallia rajaava oletus. Tyypillinen nollahypoteesi on, että parametrin arvo tai parametrien arvojen erotus on nolla. (Nollahypoteesi-nimitys juontanee tästä.) Tällöin yleensä testataan ajatusta, että yhteyttä tai vaikutusta ei ole tai että eroja ei ole. Nollahypoteesi voi olla myös, että parametrin arvo on tietty (nollasta poikkeava) luku tai

parametreja sitova rajoite (muukin kuin yhtäsuuruus). Luku tai rajoite voi seurata sovellusalan teoriasta. Parametrin nollahypoteesin mukaisuutta osoitetaan usein alaindeksillä (esim. θ_0). Nollahypoteesi voi olla myös jakaumaoletus tai muu oletus tilastollisesta mallista. Nollahypoteesista luovutaan, mikäli aineisto on vahvassa ristiriidassa sen kanssa. Johtopäätösten teossa ollaan varovaisia: Yhteys, vaikutus tai ero todetaan olemassa olevaksi tai nimetään toista suuremmaksi vain, jos siihen on vahva peruste.

Vaihtoehtoista maailmantilaa kuvaa *vastahypoteesi* (H_1; *alternative hypothesis*). Sen mukaan nollahypoteesin asettama rajoitus ei pidä paikkaansa: Parametrin tai parametrien arvot eivät ole rajoitusten mukaiset tai jakauma- tai muu oletus tilastollisesta mallista ei pidä paikkaansa.

Nollahypoteesia testataan *testisuureella*. Testisuure on tunnusluku (jakso 6.5), jota käytetään testaamiseen. *Merkitsevyystesti* (*significance test*) on päätössääntö: Mikäli testisuure osuu *hylkäysalueelle* (*critical region*), nollahypoteesi hylätään; mikäli *hyväksymisalueelle* (*acceptance region*), nollahypoteesi hyväksytään (tai jää voimaan).

Testisuureen *kriittiset arvot* (*critical values*) rajaavat hylkäys- ja hyväksymisalueet. Ne lasketaan *nollajakaumasta* (*null distribution*) eli testisuureen jakaumasta nollahypoteesin pätiessä. Todennäköisyyttä, että testisuure osuu hylkäysalueelle nollahypoteesin pätiessä, kutsutaan testin *merkitsevyystasoksi* (*level of significance*) tai *kooksi* (*size*). Merkitsevyystason symboli on yleisesti α. Todennäköisyys määritellään merkitsevyystestauksen yhteydessä frekventistisenä todennäköisyytenä. Tässä se on hylkäysalueelle osuvien arvojen osuus ääretömässä määrässä riippumattomia toistokokeita nollahypoteesin pätiessä. Hylkäystilanteessa nollahypoteesin sanotaan tulevan hylätyksi merkitsevyystasolla α tai $100 \times \alpha$ %:n merkitsevyystasolla. Testisuureen sanotaan tällöin olevan *tilastollisesti merkitsevä* kyseisellä merkitsevyystasolla.

Testin *voima* on todennäköisyys, että testisuure saa arvon hylkäysalueelta, kun vastahypoteesi pätee. Voimaa merkitään monesti $(1 - \beta)$:lla. Voiman toivotaan olevan mahdollisimman suuri. Testi on yleensä sitä voimakkaampi, mitä suurempi on testin merkitsevyystaso ja mitä enemmän on havaintoja. Testi on *tarkentuva*, mikäli testin voima suurenee 1:hteen havaintojen lukumäärän kasvaessa kohti ääretöntä.

Testaamisessa voidaan tehdä *hylkäysvirhe* (*rejection error*, *type I error*, *false positive*) tai *hyväksymisvirhe* (*acceptance error*, *type II error*, *false negative*). Hylkäysvirhe tehdään, jos paikkansapitävä nollahypoteesi hylätään. Hyväksymisvirhe tapahtuu, jos paikkansapitämätön nollahypoteesi hyväksytään. Nollahypoteesin pätiessä hylkäysvirheen todennäköisyys on α. Vastahypoteesin pätiessä hyväksymisvirheen todennäköisyys on β.

Taulukko summeeraa päätösten oikeellisuuden, niiden todennäköisyydet ja vastaavat termit. Testien yhteydessä ei tavata puhua luottamustasosta $(1-\alpha)$, mutta se nivoutuu testin merkitsevyystasoon testien ja luottamusvälien dualiteetin kautta (jakso 11.3).

		H_0 hylätään	
		kyllä	ei
H_0 tosi	kyllä	hylkäysvirhe α merkitsevyystaso	oikea päätös $1-\alpha$ luottamustaso
	ei	oikea päätös $1-\beta$ testin voima	hyväksymisvirhe β

Testisuureen toteuma poikkeaa aina nollahypoteesin mukaisesta, jos testisuureen jakauma on jatkuva. Ainoastaan tilastotieteellisessä mielessä riittävän poikkeavan eli pienen todennäköisyyden toteuman tilanteessa nollahypoteesi hylätään. Hylkäysvirheen todennäköisyys eli testin merkitsevyystaso α asetetaan siksi pieneksi merkitsevyystestauksessa. Ajatus on edellä esitetty, että nollahypoteesista luovutaan vain vakuuttavan todistusaineiston edessä.

Testin merkitsevyystaso on tutkijan päätettävissä. Testin voima ei ole yhtä helposti asetettavissa jos lainkaan. Voima riippuu vastahypoteesista — esimerkiksi sen mukaisista parametriarvoista. Tutkija saattaa joskus pystyä välillisesti vaikuttamaan voimaan kasvattamalla otoskokoa. Tutkijalla saattaa olla myös valinnanvaraa poimia mahdollisimman voimakas testi työkalukseen.

Testin merkitsevyystaso tulisi valita päätösteoreettisten tai karkeampien arvioiden päätösten seurauksista ja niiden todennäköisyyksien perusteella. Tilannetta verrataan usein oikeudenkäyntiin (Feinberg 1971). Oikeudenkäyntivertaus juontaa ainakin Laplaceen (1749–1827) asti (Neyman ja Pearson 1933, 296).

Lähtökohta on, että syytetty on syytön (nollahypoteesi). Syyttömän tuomitsemista tulee välttää viimeiseen asti (α:n tulee olla pieni). Toisaalta halutaan maksimoida todennäköisyys, että syyllinen tuomitaan ($1-\beta$:n tulisi olla suuri). Jos mikään todistusaineisto ei olisi riittävän vakuuttava tuomioon ($\alpha = 0$), ketään ei tuomittaisi ($1-\beta = 0$). Siksi täytyy hyväksyä pieni riski syyttömän tuomitsemiseen ($\alpha > 0$), jotta rikoksiin todella syyllistyneitä saataisiin tuomittua ($1-\beta > 0$). Joskus ellei monesti todistusaineisto ei riitä rikokseen syyllistyneen tuomitsemiseen, ja hänet katsotaan syyttömäksi ($\beta > 0$).

Oikeudessa sovelletaan eri näyttökynnyksiä erilaisissa rikoksissa. Oikeudenkäynnissä murhasta vaaditaan vastaansanomattomampaa näyttöä (α hyvin pieni) kuin sakko määrättäessä (α suurempi). Koska murhasta katsotaan olevan tärkeämpi saattaa tekijä vastuuseen kuin sakonalaisista rikkeistä, murhaoikeudenkäynnissä todistusaineiston keräämiseen käytetään enemmän aikaa ja vaivaa. Se kasvattaa todennäköisyyttä tuomita syyllinen (β pienenee ja $1-\beta$ suurenee).

Testaamisessa voidaan pyrkiä toimimaan vastaavasti. Pitäydytään pienessä merkitsevyystasossa (α pieni), jos hylkäysvirhe voi johtaa suuriin kielteisiin seurauksiin. Mahdollisuuksien mukaan panostetaan havaintojen laatuun ja määrään. Toisaalta jos sairauteen tai ongelmaan ei ole parannuskeinoa tai ratkaisua, lääke tai apu siihen tahdotaan mahdollisesti löytää ja hyväksyä käyttöön tavallista keveämmin perustein. Tällöin kasvatetaan hylkäysvirheen todennäköisyyttä (α suurenee), jotta testin voima ($1-\beta$) kasvaa ja hyväksymisvirheen todennäköisyys (β) pienenee.

Empiirisessä tutkimuksessa testin merkitsevyystasoa puntaroidaan harvoin ainakaan eksplisiittisesti. Useimmiten sovelletaan rutiininomaisesti merkitsevyystasoja 0.05 tai 0.01 ilman perusteluja. Sopivaa merkitsevyystasoa olisi hyvä pohtia tapauskohtaisesti, ja myös tyypillisiä pienempien merkitsevyystasojen käyttö olisi suotavaa (jakso **??**).

Merkitsevyystason lisäksi voi olla tarpeen päättää, onko vastahypoteesi kaksi- vai yksisuuntainen. Kaksisuuntaisessa vastahypoteesissa parametrin arvo poikkeaa nollahypoteesin asettamasta arvosta:

$$\mathsf{H}_1\colon \theta \neq \theta_0.$$

Tässä θ_0 on nollahypoteesin mukainen parametrin arvo. Yksisuuntaisessa vastahypoteesissa huomio kohdistuu arvoihin, jotka ovat vain pienempiä tai vain suurempia kuin nollahypoteesin mukainen parametrin arvo:

$$\mathsf{H}_1\colon \quad \theta < \theta_0 \quad \text{tai} \quad \theta > \theta_0.$$

Kaksisuuntaisen hypoteesin tilanteessa hylkäysalue muodostuu testisuureen arvoista, jotka liittyvät testattavan parametrin sekä erityisen “pieniin” että “suuriin” arvoihin; yksisuuntaisen hypoteesin tilanteessa vain “pieniin” tai “suuriin” arvoihin.

Testiä sanotaan kaksi- tai yksisuuntaiseksi vastahypoteesin mukaisesti. Tavallisesti tehdään kaksisuuntainen testi. Jos kaksisuuntaisessa testauksessa nollahypoteesi hylätään, käytännössä usein jatketaan vielä päätelmään, onko $\theta < \theta_0$ vai $\theta > \theta_0$. Yksisuuntainen testi on sopiva, jos sovellusalan teorian tai aiempien

empiiristen tutkimusten mukaan on ilmeistä, että poikkeama nollahypoteesista voi olla vain toiseen suuntaan.

Esimerkki. Kaksi- ja yksisuuntaisen testin hylkäysalueet (normaalijakauma). Kuvassa 11.2 havainnollistetaan hylkäysaluetta vasemmalla kaksisuuntaisen ja oikealla yksisuuntaisen testin tilanteissa testisuureen noudattaessa nollahypoteesin pätiessä standardinormaalijakaumaa.[125] Standardinormaalijakauman väritetyt hännät rajautuvat vasemmalla 0.025. ja 0.975. kvantiileihin -1.960 ja 1.960 ja oikealla 0.95. kvantiiliin 1.645. Sekä kaksi- että yksisuuntaisessa testauksessa testisuure osuu nollahypoteesin pätiessä hylkäysalueelle todennäköisyydellä 0.05. Kaksisuuntaisessa testauksessa itseisarvoltaan pieni tai suuri testisuureen arvo voi johtaa nollahypoteesin hylkäämiseen. Kuvan esimerkissä ainoastaan suuri testisuureen arvo voi johtaa nollahypoteesin hylkäämiseen yksisuuntaisessa testauksessa.

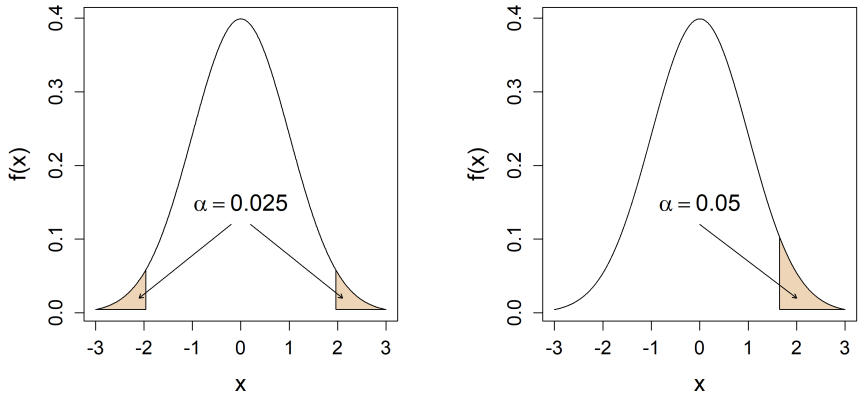
Yksisuuntainen testaus kasvattaa testin voimaa valittuun suuntaan. Kaksisuuntaisessa testauksessa testisuureen arvo 1.7 ei johda nollahypoteesiin hylkäämiseen merkitsevyystasolla 0.05; yksisuuntaisessa testauksessa johtaa. Yksisuuntaisella testillä ei ole voimaa lainkaan toiseen suuntaan: Oli testisuureen arvo kuinka pieni tahansa, nollahypoteesia ei hylätä.

Yksisuuntaisessa testauksessa hylkäysalue voi tietenkin koostua myös pienistä testisuureen arvoista. Tällöin hylkäysaluetta havainnollistaisi kuvan 11.2 oikeanpuolen peilikuva, jossa jakauman vasen häntä olisi väritetty. □

Esimerkki. Testin voima.[126] Kuva 11.3 havainnollistaa testin voiman kasvamista havaintomäärän kasvaessa. Lasketaan keskiarvo $\hat{\mu}$ normaalijakaumaa $\mathsf{N}(40, 36)$ noudattavista satunnaismuuttujista. Testisuure on $Z = (\hat{\mu} - 40)/(\sqrt{36/n})$. Nollahypoteesi $\mu_0 = 40$ hylätään merkitsevyystasolla 0.05, jos testisuureen havaitun arvon itseisarvo $|z|$ ylittää standardinormaalijakauman 0.975. kvantiilin 1.960.

Merkinnällä $n = 9$ — havaintojen lukumäärä — osoitettu käyrä kuvaa, kuinka testin voima kasvaa μ:n etääntyessä 40:stä. Nollahypoteesin $\mu = 40$ ympäristössä voima on jonkin verran suurempi kuin testin merkitsevyystaso 0.05. Testin voima lähenee yhtä, kun μ etääntyy 40:stä. Merkinnällä $n = 36$ osoitettu käyrä toistaa samat kuviot mutta kärjekkäämmin, kun havaintoja on 36. Odotusarvon μ poikkeama 40:stä johtaa nyt kauttaaltaan suurempaan testin voimaan kuin silloin, kun havaintoja oli vain 9. Samansuuruinen erotus $\hat{\mu} - 40$ johtaa useammin nollahypoteesin hylkäämiseen, kun $n = 36$ kuin silloin, kun $n = 9$. Testin merkitsevyystaso on sama 0.05 molemmilla havaintomäärillä. □

Hyväksymisalue-termiin ei pidä suhtautua kirjaimellisesti. Jos testisuureen arvo ei ole tilastollisesti merkitsevä, siitä ei tule päätellä, että nollahypoteesi olisi tosi. Ei-merkitsevä arvo voi johtua vaikkapa pienestä otoskoosta ja siten pienestä



Kuva 11.2: Kaksi- ja yksisuuntaisen testin hylkäysalueet (normaalijakauma).

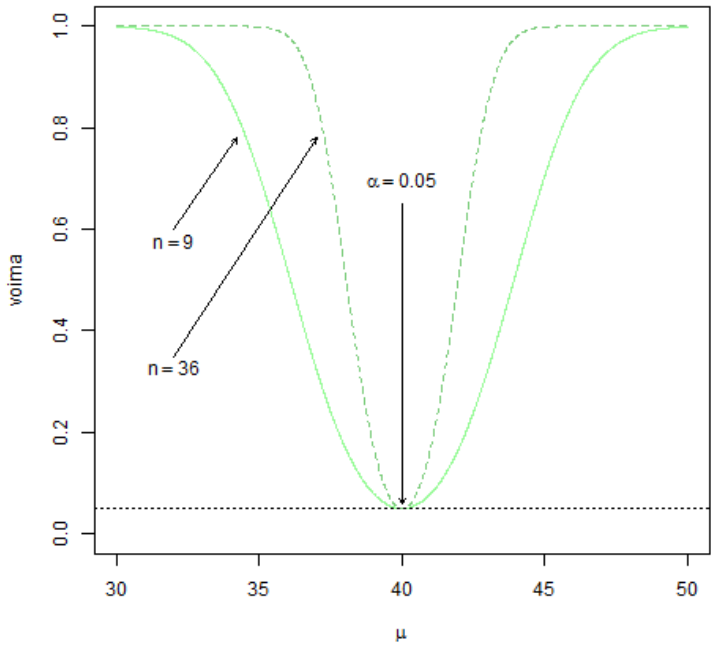
testin voimasta. Sopiva päätelmä on pikemminkin, että nollahypoteesi jää voimaan kuin että se hyväksytään. Jaksossa 15.3 on esimerkkejä tilanteista, joissa on selvää, että nollahypoteesia ei tule hyväksyä, vaikka se jäisi voimaan.

11.2 *p*-arvo

Empiirisessä tutkimuksessa tyypillisesti raportoidaan testisuureen *p-arvo* (*p-value*) eli *havaittu merkitsevyystaso* (*observed level of significance*). On tärkeää ymmärtää, mitä *p*-arvo tarkoittaa — ja mitä ei. Monet soveltajat tulkitsevat *p*-arvon väärin (esim. Gigerenzer 2018). Psykologian oppikirjoista 89 %:n on raportoitu kuvaavan *p*-arvon väärin (Cassidy ym. 2019). Väärintulkintojen on katsottu johtaneen "valtaisaan sekaannukseen ja sanaharkkaan" (Kuffner ja Walker 2019). Väärintulkinnat eivät rajoitu tutkimukseen: oikeudenkäynnissä tuomari voi ymmärtää testisuureeseen liittyvän todennäköisyyden väärin (Fatti ja Greenacre 1991). Alla selitetään asian ydin. Yleispätevämpiä esityksiä on teoreettisemmissa oppikirjoissa.

Olkoon nollahypoteesi $\theta = \theta_0$, kaksisuuntainen vastahypoteesi $\theta \neq \theta_0$ ja testisuureen $t(X_1, \ldots, X_n)$ jakauma symmetrinen nollan ympärillä. Testisuureen toteuman

$$t(x_1, \ldots, x_n)$$



Kuva 11.3: Kaksisuuntaisen testin voima, kun testisuure on $(\hat{\mu} - 40)/(\sqrt{36/n})$, testin merkitsevyystaso on 0.05, havainnot noudattavat jakaumaa $\mathsf{N}(40, 36)$ ja nollahypoteesi on $\mu_0 = 40$.

p-arvo on todennäköisyys

$$\mathsf{P}\left(|t(X_1, \ldots, X_n)| \geq |t(x_1, \ldots, x_n)|\right)$$

nollahypoteesin pätiessä. Edellä $X_1, \ldots, X_n$ on satunnaisotos ja $x_1, \ldots, x_n$ sen toteuma. Testisuureen p-arvo on kaksisuuntaisessa testauksessa todennäköisyys, että testisuure saa arvon, joka on nollahypoteesiin verrattuna yhtä poikkeava tai vielä poikkeavampi kuin havaittu arvo, kun nollahypoteesi pätee. Jos vasta-

hypoteesi on yksisuuntainen $\theta > \theta_0$, p-arvo on todennäköisyys

$$\mathsf{P}\left(t(X_1, \ldots, X_n) \geq t(x_1, \ldots, x_n)\right),$$

jos $t(x_1, \ldots, x_n)$:n suuret arvot määrittelevät hylkäysalueen. Epäyhtälön suunta kääntyy yllä, jos vastahypoteesi on $\theta < \theta_0$ ja hylkäysalue koostuu $t(x_1, \ldots, x_n)$:n pienistä arvoista.

Esimerkki. Miesten keskipituus (jatkoa jaksosta 10.4.2). Olkoon nollahypoteesi, että normaalijakautuneen satunnaismuuttujan $\mathsf{N}(\mu, \sigma^2)$ odotusarvo $\mu = \mu_0$ ja kaksisuuntainen vastahypoteesi, että $\mu \neq \mu_0$. Jaksossa 12.4.2 osoitetaan sopivaksi testisuureeksi

$$t(x_1, \ldots, x_n) = \frac{\hat{\mu} - \mu_0}{s/\sqrt{n}}.$$

Sen p-arvo on todennäköisyys

$$\mathsf{P}\left(|(t(X_1, \ldots, X_n)| \geq \left|\frac{\hat{\mu} - \mu}{s/\sqrt{n}}\right|\right),$$

Testisuure noudattaa $\mathsf{t}(n-1)$-jakaumaa nollahypoteesin pätiessä (jakso 12.4.2).

Tutkijan kasvuteoria ehdottaa miesten keskipituudeksi 181.5 cm:ä vuonna 2010. Onko kasvukäyräaineisto ristiriidassa tutkijan teorian kanssa? Lasketaan testisuure ja sen p-arvo kaksisuuntaisessa testauksessa (H_1: $\mu \neq 181.5$). Testisuure on

$$\frac{\hat{\mu} - \mu_0}{s/\sqrt{n}} = \frac{181.042 - 181.5}{6.0609/\sqrt{4000}} = -4.779235.$$

Todennäköisyys, että $\mathsf{t}(3999)$-jakaumaa noudattava satunnaismuuttuja saisi itseisarvoltaan 4.779235:ttä suuremman arvon, on

$$\begin{aligned}&\mathsf{P}\left(t(X_1, \ldots, X_n) \leq -4.779235\right) + \mathsf{P}\left(t(X_1, \ldots, X_n) \geq 4.779235\right)\\ &= 0.0000009114233 + 0.0000009114233 \approx 0.000002.\end{aligned}$$

Todennäköisyys voidaan laskea R-komennolla `pt(-4.779235,3999)+1-pt(4.779235,3999)` tai `t`-jakauman symmetrisyyden perusteella komennolla `2*pt(-4.779235,3999)`. Mikäli vastahypoteesi olisi yksisuuntainen (H_1: $\mu <$ 181.5), testisuureen p-arvo puolittuisi:

$$\mathsf{P}\left(t(X_1, \ldots, X_n) \leq -4.779235\right) = 0.0000009114233 \approx 0.000001.$$

Näyttö nollahypoteesia vastaan vahvistuisi.

Kumpi vastahypoteesi onkaan, p-arvo on tavattoman pieni. Aineisto on vahvassa ristiriidassa tutkijan teorian kanssa — ainakin tilastotieteellisessä mielessä. Tutkijalla on syytä hylätä teoriansa. □

Yksinkertaisimmillaan p-arvoa käytetään merkitsevyystestin tapaan joko–tai-päätösten tekoon: Jos p-arvo on pienempi kuin etukäteen valittu merkitsevyystaso, nollahypoteesi hylätään; muuten nollahypoteesi jää voimaan.

p-arvoa voidaan hyödyntää merkitsevyystestiä joustavammin mittana todistusaineiston vahvuudesta nollahypoteesia vastaan: Mitä pienempi p-arvo on, sitä jyrkemmin testisuure puhuu nollahypoteesia vastaan. Tutkija voi pohtia aineiston ja hypoteesien yhteensopivuutta tai -sopimattomuutta ja johtopäätöksiä sitoutumatta etukäteen tiettyyn merkitsevyystasoon. Menettelyn huono puoli on, että merkitsevyystasosta ei ole selvyyttä nollahypoteesi hylättäessä. Empiirisessä tutkimuksessa kannattaa kuitenkin muutenkin huomioida kaikki asiaan liittyvät seikat johtopäätöksissä pelkän p-arvon ja merkitsevyystason napittamisen sijaan. Nollahypoteesia ei ole välttämätöntä hyväksyä tai hylätä yksinomaan sen perusteella, onko aineisto ristiriidassa sen kanssa. Aina ei ole myöskään tarpeen tehdä jyrkkää päätöstä merkitsevyystestauksen tapaan.

Esimerkki. Tutkitaan tuotteen kulutuksen ja hinnan välistä yhteyttä. Lasketaan nollahypoteesia “ei yhteyttä” haastava testisuure. Sen p-arvoksi saadaan 0.07. Jos on aivan ilmeistä, että hinta vaikuttaa kulutukseen, ei tyydytä jatkamaan analyyseja nollahypoteesin pohjalta. Suurehko p-arvo heijastelee vaikkapa otoskoon pienuutta. □

Esimerkki. Pienen p-arvon mukaan aineisto on ristiriidassa yksinkertaisen tilastollisen mallin kanssa puoltaen monimutkaista mallia. Tutkija saattaa silti pitäytyä yksinkertaisemmassa mallissa. Se on helpompi selittää soveltajille ja kenties helpompi estimoida ja ylipäänsä käyttää. Jos se esimerkiksi ennustaa kohtuullisesti, sen käyttö voi olla perusteltua. Tutkija ei silti hyväksy nollahypoteesia, että malli olisi oikea. □

Tuosta poikki -tyyppisen merkitsevyystestauksen perinnettä selittää tuonnoinen jakaumien vain valikoitujen kvantiilien taulukointi ja ylipäänsä laskennan työläys. p-arvoa saattoi vain haarukoida. Nykyään p-arvo on useimmiten helposti laskettavissa, ja sitä kannattaa hyödyntää muuhun tietoon verraten. Sitä ei tule ylitulkita.

Esimerkki. Yleisiä p-arvon vääriä tulkintoja:[127]

- p-arvo on todennäköisyys, että nollahypoteesi pätee. Ei!: p-arvo on nimenomaan nollahypoteesin pätiessä laskettu todennäköisyys. Frekventistisessä tilastotieteessä ei arvioida hypoteesien todennäköisyyksiä.

- Todennäköisyys, että vastahypoteesi on tosi, on $1-p$. Ei! Laskettaessa p-arvoa ei oteta kantaa myöskään vastahypoteesin todennäköisyyteen. Pieni p-arvo merkitsee vain, että on saatu hyvin poikkeuksellinen testisuureen arvo nollahypoteesin pätiessä.

- p-arvo on todennäköisyys saada sattumalta havaittu tai poikkeavampi testisuureen arvo. Ei!: Todennäköisyys pitää laskea nimenomaan nollahypoteesin pätiessä, mitä ei edellä todettu.

- Pieni p-arvo tarkoittaa käytännön kannalta tärkeää havaintoa. Ei! Suurilla havaintomäärillä hyvinkin pienet poikkeamat nollahypoteesista voivat johtaa mikroskooppiseen p-arvoon. Asiaa pohditaan tarkemmin jaksossa 11.4. □

Jos jakauma on epäsymmetrinen, p-arvo voidaan määritellä kaksisuuntaisessa testauksessa eri tavoilla (esim. Agresti 2013, 92, Cox 2020). Asiaa sivutaan jaksossa 12.5.1. p-arvon käyttöä empiirisessä tutkimuksessa pohditaan lisää jaksossa **??**.

11.3 Luottamusvälien ja testien yhteys

Tarkastellaan kaksisuuntaisen testin merkitsevyystasolla α ja samasta testisuureesta muodostetun kaksisuuntaisen $100 \times (1-\alpha)$ %:n luottamusvälin yhteyttä. Oletetaan yksinkertaisuuden ja konkreettisuuden vuoksi, että havainnot noudattavat normaalijakaumaa $\mathsf{N}(\mu, \sigma^2)$, testataan nollahypoteesia "odotusarvo on μ", varianssin estimaatti on $s^2 > 0$, $n > 0$ on otoskoko, testisuure on $(\hat{\mu} - \mu)/(s/\sqrt{n})$ ja sen kriittiset arvot ovat $\mathsf{t}_{\alpha/2}(n-1) < 0$, $\mathsf{t}_{1-\alpha/2}(n-1) > 0$ ja $-\mathsf{t}_{\alpha/2}(n-1) = \mathsf{t}_{1-\alpha/2}(n-1)$. Tällöin jakson 12.4.2 teorian ja jakson 11.2 p-arvo-esimerkin (miesten keskipituuden testauksesta) mukaan

$$
\begin{aligned}
1-\alpha &= \mathsf{P}\left(\mathsf{t}_{\alpha/2}(n-1) < \frac{\hat{\mu}-\mu}{s/\sqrt{n}} < \mathsf{t}_{1-\alpha/2}(n-1)\right) \\
&= \mathsf{P}\left(\mathsf{t}_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \hat{\mu}-\mu < \mathsf{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) \\
&= \mathsf{P}\left(-\hat{\mu} + \mathsf{t}_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} < -\mu < -\hat{\mu} + \mathsf{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right) \\
&= \mathsf{P}\left(\hat{\mu} - \mathsf{t}_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} > \mu > \hat{\mu} - \mathsf{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)
\end{aligned}
$$

$$= \mathsf{P}\left(\hat{\mu} - \mathtt{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} < \mu < \hat{\mu} + \mathtt{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right).$$

Ensimmäisen rivin mukaan riippumattomissa toistokokeissa testisuureen $(\hat{\mu} - \mu)/(s/\sqrt{n})$ arvo osuu todennäköisyydellä $1-\alpha$ välille $(\mathtt{t}_{\alpha/2}(n-1), \mathtt{t}_{1-\alpha/2}(n-1))$. Mikäli näin käy, nollahypoteesi jää voimaan. Jos testisuureen arvo sijoittuu välin $(\mathtt{t}_{\alpha/2}(n-1), \mathtt{t}_{1-\alpha/2}(n-1))$ ulkopuolelle, nollahypoteesi hylätään.

Viimeisen rivin muoto on tuttu luottamusvälin lausekkeista edellä: Riippumattomissa toistokokeissa väli

$$\left(\hat{\mu} - \mathtt{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \hat{\mu} + \mathtt{t}_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$$

peittää odotusarvon μ todennäköisyydellä $1-\alpha$ (kaava (10.7) ja jakso 10.4.2).

Jos testisuure sijoittuu kriittisten arvojen väliin, yhtälöketjun ensimmäisellä rivillä olevat epäyhtälöt pätevät, eikä nollahypoteesia hylätä. Tällöin yhtälöketjun viimeisellä rivillä olevat epäyhtälöt pätevät eli luottamusväli peittää μ:n. Jos nollahypoteesia, että odotusarvo on μ ei hylätä, testin merkitsevyystasoa $100 \times \alpha$ %:a vastaava $100 \times (1-\alpha)$ %:n luottamusväli peittää μ:n.

Jos testisuure ei sijoitu kriittisten arvojen väliin, yhtälöketjun yllä ensimmäisellä rivillä olevat epäyhtälöt eivät päde. Nollahypoteesi hylätään. Tällöin myöskään yhtälöketjun viimeisellä rivillä olevat epäyhtälöt eivät ole voimassa eli luottamusväli ei peitä μ:tä. Jos nollahypoteesi "odotusarvo on μ" hylätään, testin merkitsevyystasoa vastaava luottamusväli ei peitä μ:tä.

Esimerkki. Miesten keskipituus (jatkoa jaksoista 10.4.2 ja 11.2). Miesten keskipituuden 95 %:n luottamusväli (180.9 cm, 181.2 cm) ei peitä arvoa 181.5 cm (jakso 10.4.2). Näin ollen nollahypoteesi 181.5 cm:n keskipituudesta pitäisi tulla hylätätyksi merkitsevyystasolla 0.05. Niin käy, sillä vastaavan testisuureen p-arvo on pienempi kuin 0.05 (jakso 11.2). □

Usein testisuureen kriittiset arvot osataan määrittää vain suurille havaintomäärille ja samoja kriittisiä arvoja käytetään pienillekin otoksille. Testin todellista kokoa ei tällöin tunneta. Joskus tiedetään, että todellinen koko on pienempi kuin käytettyihin kriittisiin arvoihin liittyvä nimellinen koko. Testi on tällöin *konservatiivinen*. Vastaavasti luottamusväli on tällöin liian leveä, ja myös se on konservatiivinen. Termi tulee siitä, että ollaan taipuvaisia pitäytymään nollahypoteesissa.

Esimerkki. Jos havainnot eivät ole normaalijakaumasta, testisuure $(\hat{\mu} - \mu)/(s/\sqrt{n})$ ei ylipäänsä sijoitu välille $(\mathtt{t}_{\alpha/2}(n-1), \mathtt{t}_{1-\alpha/2}(n-1))$ todennäköisyydellä

$1-\alpha$. Keskeisen raja-arvolauseen (jakso 7.3) mukaan suurilla havaintomäärillä yhtälöryhmä yllä silti likimain pätee. Jos pienissä otoksissa $(\hat{\mu}-\mu)/(s/\sqrt{n})$ sijoittuu $(1-\alpha)$:aa suuremmalla todennäköisyydellä välille $(\mathtt{t}_{\alpha/2}(n-1), \mathtt{t}_{1-\alpha/2}(n-1))$ ja sitä käytetään hyväksymisalueena, niin testin todellinen koko $\alpha(n)$ on pienempi kuin sen nimellinen koko α. Tällöin luottamusväli $(\hat{\mu}-\mathtt{t}_{1-\alpha/2}(n-1)s/\sqrt{n},\ \hat{\mu}+\mathtt{t}_{1-\alpha/2}(n-1)s/\sqrt{n})$ peittää odotusarvon μ todennäköisyydellä $1-\alpha(n)$, joka on suurempi kuin nimellinen luottamustaso $1-\alpha$. Sekä testi että luottamusväli ovat konservatiivisia. □

Joskus pohditaan, ovatko luottamusvälit vai testit tärkeämpiä päättelyn työkaluja (esim. Cox 2006, 42, 2020). Jotkut pitävät luottamusväliä testiä informatiivisempana, koska se kertoo, millaisia parametrin arvot voisivat olla merkitsevyystestin päätelmän ollessa yksioikoisempi (θ_0 on tai ei). Toisaalta luottamusväli on sidottu luottamustasoon mutta p-arvo mittaa jatkuva-arvoisesti aineiston ja nollahypoteesin yhteensopivuutta. Molemmille on käyttöä. Useimmilla tieteenaloilla testaaminen lienee paljon yleisempää kuin luottamusvälien lasku.

Esimerkki. Luottamusvälit ja p-arvot biolääketieteessä 1990–2015. Chavalarias ym. (2016) laskivat, että biolääketieteellisten artikkelien tiivistelmistä 2 %:ssa raportoitiin luottamusväli ja 16 %:ssa p-arvo.[128] Tiivistelmissä p-arvojen raportointi kaksinkertaistui 1990–2014. □

11.4 Tilastollinen merkitsevyys ja käytännön merkitys

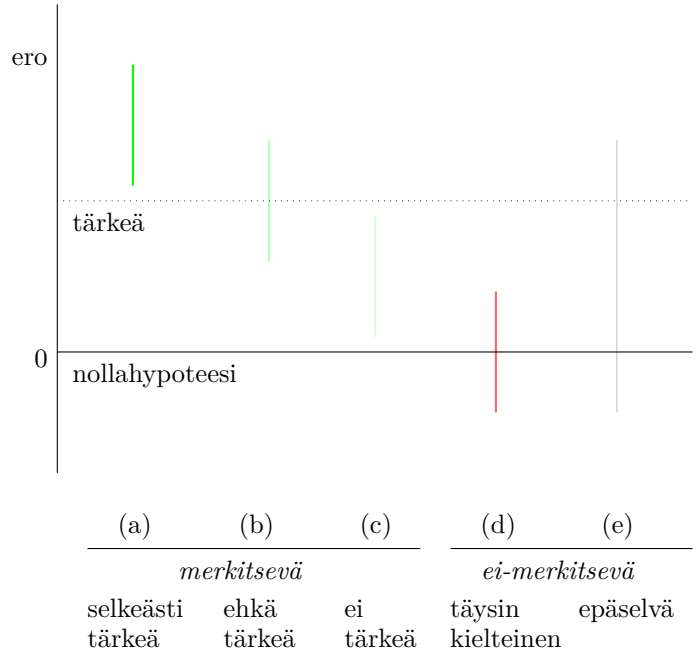
Tilastollinen merkitsevyys ja käytännön merkitys ovat eri asioita. Sovellusten kannalta merkityksetönkin poikkeama nollahypoteesista muodostuu suurilla havaintomäärillä tilastollisesti merkitseväksi, koska testien voima kasvaa kohti yhtä otoskoon kasvaessa. Kuva 11.4 havainnollistaa. Pystyakselilla on estimoidun parametrin tai estimoitujen parametrien erotuksen ero nollahypoteesin mukaisesta arvosta (tyypillisesti 0).

Tilanteissa (a)–(c) luottamusväli ei peitä nollahypoteesin mukaista arvoa ja ero on tilastollisesti merkitsevä. Ainoastaan tilanteessa (a) on tehty selkeästi tärkeä löydös: Ero on tilastollisesti merkitsevä ja käytännön kannalta suuri. Tilanteessa (b) luottamusväli peittää sekä käytännön kannalta tärkeitä että toisarvoisia eroja. Löydös voi olla tärkeä tai toisarvoinen. Vaikka ero on tilastollisesti merkitsevä, kohdan c) löydös on mielenkiinnoton, koska luottamusväli peittää vain eron arvoja, joilla ei ole käytännön merkitystä. Luottamusväli on

mahdollisesti estimoitu suuresta otoksesta, jolloin pienetkin poikkeamat nollahypoteesista tulevat tilastollisesti merkitseviksi.

Tilanteissa (d)–(e) ero ei ole tilastollisesti merkitsevä. Ero on mielenkiinnoton sekä tilastotieteen että sovellusalan näkökulmasta tilanteessa (d). Luottamusväli peittää nollahypoteesin mukaisen arvon eikä yllä käytännön merkitystä omaaviin arvoihin. Löydös on kaikinpuolin negatiivinen. Viimeinen tilanne (e) on moniselitteinen. Luottamusväli on niin leveä, että se peittää eron nollahypoteesiarvon mutta myös käytännön kannalta suuria arvoja. Otoskoko on mahdollisesti ollut pieni, jolloin parametrin tai parametrien estimaatit ovat epätarkkoja.

Tilastollinen merkitsevyys yhdistyy monen soveltajan mielessä merkittävään tulokseen. Kuvan 11.4 mukaan yhteyttä ei välttämättä ole. Tilastollisesti merkitsevä -termin sijalle on ehdotettu termiä *tilastollisesti erottuva* (*statistically discernible*). Se on osuvampi ja neutraalimpi muttei vielä yleisesti käytetty.



Kuva 11.4: Luottamusvälit, tilastollinen merkitsevyys ja käytännön merkitys (Armitage ym. 2002, 92).

Luku 12

Testejä

> Opiskella ja sitten harjoittaa oppimaansa — eikö se toisi tyydytystä?[129]
> Kungfutse (551–479 eKr.)

Testausta havainnollistetaan alla kompakteilla empiirisillä esimerkeillä. Tutkimuksessa aineistoihin tulee perehtyä huolellisesti ennen testaamista. Ugarten ym:iden (2016) sekä Chiharan ja Hesterbergin (2019) kirjoissa on esimerkkejä graafisista tarkasteluista testien yhteydessä. Luvussa aineistojen oletetaan olevan satunnaisotoksia.

12.1 Testejä osuuksille

12.1.1 Osuustesti

Estimoidaan osuutta $\hat{\pi} = y/n$:llä (tapahtumien ja havaintojen lukumäärien suhde). Olkoon nollahypoteesin mukaan osuus $\pi = \pi_0$. Sen pätevyyttä voidaan koetella testisuureella

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}, \tag{12.1}$$

joka on suurilla havaintomäärillä standardinormaalijakautunut (jakso 9.3). Jos $\pi_0 = 0.5$, niin approksimaatio toimii, jos havaintoja on ainakin 20. Muuten pitää olla $n\pi_0 \geq 10$ ja $n(1-\pi_0) \geq 10$. (Agresti ja Finlay 2009, 156, 172. Dickinson Gibbonsin ja Chakrabortin 2020, 181, mukaan normaalisuusapproksimaatio toimii tilanteessa $\pi_0 = 0.5$ jo, kun $n = 12$.)

Esimerkki. Oikeuspoliittinen tutkimuslaitos tutki käräjäoikeuksien päätöksiä lapsen asumisesta ajalla 14.11.2005–13.2.2006 (529 havaintoa).[130] Keskitytään tutkimaan päätöksiä, joissa lapset määrättiin asumaan vain jommankumman vanhemman luona. Näissä päätöksissä lapsi määrättiin asumaan 35:ssä isän ja 83:ssa äidin luona (118 havaintoa).[131] Vastaavat prosenttiosuudet ovat noin 29.7 ja 70.3.

Testataan kaksisuuntaisesti merkitsevyystasolla 0.01 nollahypoteesia, että käräjäoikeudet määräävät lapset asumaan eri sukupuolta olevien vanhempien luona yhtä todennäköisesti ($\pi_0 = 0.5$). Testisuure

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.2966102 - 0.5}{\sqrt{0.5(1-0.5)/118}} = -4.418758$$

laskettiin jo jaksossa 7.4.3. Havaintoja on yli 20, joten normaalisuusapproksimaatioehto tilanteessa $\pi_0 = 0.5$ täyttyy. Testisuureen p-arvo saadaan standardinormaalijakaumasta ja on noin 0.00001 (`2*pnorm(-4.418758)`). Nollahypoteesi hylätään merkitsevyystasolla 0.01. Isät ja äidit voittavat asumisriidan eri todennäköisyydellä. Isien voittotodennäköisyys on pienempi.

Testin voi tehdä yhtä hyvin vertaamalla 0.5:teen äitien voitto-osuutta 88/118. Testisuureen arvo olisi 4.418758. p-arvo, testin tulos ja johtopäätökset olisivat samat.

R:ssä testin laskut voi toteuttaa komennolla `prop.test(x=35,n=118,p=0.5, correct=F)`. (Määre `correct=F` eli `correct=FALSE` merkitsee, että ns. jatkuvuuskorjausta ei tehdä.) Komento palauttaa testisuureen neliön $19.525 = (-4.418758)^2$. Komennon palautteessa sitä verrataan $\chi^2(1)$-jakaumaan testin yllä kanssa yhtäpitävästi (jakso 7.2.2). □

Jos havaintoja on hyvin vähän, mieleen voi tulla perustaa testi Y:n binomijakaumaan $\mathsf{Bin}(n, \pi)$ (esim. Agresti ja Finlay 2009, 172–173). Testisuureen hylkäysalue eli lukumäärät k_1 ja k_2 valitaan niin, että $\mathsf{P}(Y \leq k_1 \text{ tai } Y \geq k_2) = \sum_{i=0}^{k_1} \binom{n}{i} \pi_0^i (1-\pi_0)^{n-i} + \sum_{i=k_2}^{n} \binom{n}{i} \pi_0^i (1-\pi_0)^{n-i} \geq \alpha$, jossa α on testin merkitsevyystaso. Koska jakauma on diskreetti, merkitsevyystasoa ei voida ylipäänsä asettaa täsmälleen α:ksi. Lähemmäksi haluttua merkitsevyystasoa päästään yleensä perustamalla päättely keski-p-arvoon (esim. Agresti 2007, 2013, 2019). Laskut voidaan tehdä Anna Gottardin R-koodilla.[132] Agresti (2013, 605) arvioi, että tällainen keski-p-korjattu testi on hyvä vaikkakin hieman konservatiivinen (jakso 11.3). Fagerland ym. (2019, 58–65) tuntuvat pitävän päätekstin testisuureeseen (12.1) perustuvaa testiä parhaana kaikilla otoskoilla. Hyvinä testeinä he pitävät myös Blakerin testiä ja juuri kuvattua keski-p-korjattua binomijakaumaan perustuvaa eksaktia testiä. Testisuureeseen (12.1) perustuvan testin koko voi olla jonkin verran epävakaampi kuin kahden viimeksi mainitun. Testisuure (12.1) pärjää varsin hyvin testejä vertailtaessa. Bilder ja Loughin (2015, 17) suosittelevat sen käyttöä.

12.1.2 Osuuksien erotuksen testi, jos osuudet ovat riippumattomia

Testataan, eroavatko osuudet π_1 ja π_2 populaatioissa. Nollahypoteesi on H_0: $\pi_1 - \pi_2 = 0$. Molemmista populaatioista on n_1:n ja n_2:n kokoiset riippumattomat otokset. Lasketaan havaitut osuudet $\hat{\pi}_1 = y_1/n_1$ ja $\hat{\pi}_2 = y_2/n_2$, joissa y_i ja n_i ovat tapahtumien ja havaintojen lukumäärät i. otoksessa, $i = 1, 2$. Testisuure on

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}}. \tag{12.2}$$

Siinä erotuksen $\hat{\pi}_1 - \hat{\pi}_2$ varianssi $\pi(1-\pi)(1/n_1 + 1/n_2)$ estimoidaan yhdistetystä otoksesta, koska nollahypoteesin mukaan osuus π on sama populaatioissa:

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2}.$$

Testisuure on standardinormaalijakautunut, jos havaintoja on paljon. Karkea ohjenuora riittävälle likiarvoistukselle on, että $y_i > 10$ ja $n_i - y_i > 10$ $i = 1, 2$. Jos testi tehdään kaksisuuntaisena riittää, että $y_i > 5$ ja $n_i - y_i > 5$. (Agresti ja Finlay 2009, 190.)

Esimerkki. Palo-Repo (2015) tutki Helsingin hovioikeuden 2003–2006 ratkaisemia riitoja lasten huoltajuudesta tai asumisesta.[133] Toisesta vanhemmasta tai hänen uudesta puolisostaan tehdyt syytökset väkivaltaan, päihteisiin tai huumeisiin tai mielenterveysongelmiin liittyen ovat yleisiä näissä riidoissa: Palo-Revon tutkimasta 198 riidasta 94:ssä eli 47.5 %:ssa oli tehty tällainen syytös.

Tutkitaan eroa syytösten toteennäyttämisosuuksissa osa-aineistossa, jossa vain toinen vanhempi tekee syytöksen ja osa-aineistossa, jossa molemmat vanhemmat tekevät syytöksen:

		toteennäytetty (lkm)			toteennäytetty (osuus)		
		kyllä	ei	Σ	kyllä	ei	Σ
syytös	vain toisesta	41	30	71	0.577	0.423	1
	molemmista	16	30	46	0.348	0.652	1
	Σ	57	60	117	0.487	0.513	1

Osa-aineistossa, jossa vain toinen vanhempi teki syytöksen, toteennäytettyjen syytösten osuus on $100 \times 41/71$ % = 57.74648 %. Osa-aineistossa, jossa molemmat vanhemmat tekivät syytöksen, osuus on $100 \times 16/46$ % = 34.78261 %.

Testataan merkitsevyystasolla 0.05 nollahypoteesia, että syytösten toteennäyttämisen todennäköisyydet eivät eroa, kun vain toinen vanhempi tekee syytöksen tai kun molemmat vanhemmat tekevät syytöksen.

Toteennäytettyjen syytösten osuus yhdistetyssä aineistossa on

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \frac{71 \times 0.5774648 + 46 \times 0.3478261}{71 + 46} = \frac{41 + 16}{117} \approx 0.4871795.$$

Testisuure on

$$\frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.5774648 - 0.3478261}{\sqrt{0.4871795 \times (1 - 0.4871795) \times (\frac{1}{71} + \frac{1}{46})}} \approx 2.427354.$$

2×2 -aineistotaulukon jokaisessa solussa on yli 5 havaintoa. Testisuureen normaalisuusapproksimaation ehto täyttyy. Testisuureen p-arvo on 0.0152 (`2*(1-pnorm(2.427354)`). Nollahypoteesi hylätään merkitsevyystasolla 0.05. Oikeus katsoo syytöksen toteennäytetyksi todennäköisemmin silloin, kun vain toinen oikeudenkäynnin osapuolista on tehnyt syytöksen.

R laskee osuudet ja testisuureen p-arvon käskyllä `prop.test (c(41,16), c(71,46),correct=F)` (ohje `correct=F` tarkoittaa, ettei tehdä ns. jatkuvuuskorjausta). R:n palautteessa on testisuureen arvon 2.427354 neliö. Selitys on jaksossa 12.2.2.

Bilder ja Loughin 2015, 35) suosittelevat tässä esitettyä testiä. Pienillä havaintomäärillä on silti parempi käyttää keski-p-korjattua Fisherin eksaktia testiä (ks. ohje jakson 12.2.2 lopussa). Fagerland ym. (2017, 174) suosittelevat sitä käytettäväksi kaikilla otoskoilla.

12.1.3 Osuuksien erotuksen testi, jos osuudet eivät ole riippumattomia

Palataan tilanteeseen, jossa havainnot ovat kaltaistettuja pareja, noudattavat multinomijakaumaa $\mathsf{Mul}(1, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$

		Y		
		y_1	y_2	Σ
X	x_1	π_{11}	π_{12}	π_{1+}
	x_2	π_{21}	π_{22}	π_{2+}
	Σ	π_{+1}	π_{+2}	1

ja muodostavat paritaulukon

		Y		
		y_1	y_2	Σ
X	x_1	n_{11}	n_{12}	n_{1+}
	x_2	n_{21}	n_{22}	n_{2+}
	Σ	n_{+1}	n_{+2}	n

(jakso 10.2.3). Tällaisen aineiston yhteydessä mielenkiintoinen nollahypoteesi voi olla, päteekö *reunahomogeenisuus* (*marginal homogeneity*) $\pi_{1+} = \pi_{+1}$ ja $\pi_{2+} = \pi_{+2}$. Reunahomogeenisuutta voidaan testata tutkimalla vain toisen yhtälön pitävyyttä, koska reunatodennäköisyydet summautuvat 1:ksi ja yhtälöt seuraavat siksi toisistaan.

Reunahomogeenisuuden pätiessä $\pi_{12} = \pi_{21}$:

$$\pi_{1+} = \pi_{+1} \;\Leftrightarrow\; \pi_{11} + \pi_{12} = \pi_{11} + \pi_{21} \;\Leftrightarrow\; \pi_{12} = \pi_{21}.$$

Tällöin aineistossa n_{12}:n ja n_{21}:n tulisi olla satunnaisvaihtelun puitteissa yhtäsuuret.

Ajatellaan $(1, 2)$-solun havaintoja tapahtumina, ja oletetaan, että $n_{12}+n_{21} > 0$. Lukumäärää n_{12} vastaava satunnaismuuttuja N_{12} noudattaa binomijakaumaa $\mathsf{Bin}(n_{12}+n_{21}, 0.5)$, jos nollahypoteesi reunahomogeenisuudesta pätee. Testisuure muodostetaan binomijakautuneen satunnaismuuttujan normaalisuusapproksimaatiosta:

$$\frac{n_{12} - (n_{12} + n_{21}) \times 0.5}{\sqrt{(n_{12} + n_{21}) \times 0.5 \times 0.5}} = \frac{0.5(n_{12} - n_{21})}{0.5\sqrt{n_{12} + n_{21}}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

(jakso 7.4.3). Testisuuretta verrataan standardinormaalijakauman kriittisiin arvoihin. Testiä kutsutaan *McNemarin testiksi*. R-komento `mcnemar.test` laskee testisuureen neliön ja vertaa sitä $\chi^2(1)$-jakaumaan.

Agrestin (2013, 416) mukaan normaalisuusapproksimaatio toimii, jos $n_{12} + n_{21} > 10$. Agresti ja Finlay (2009, 202) esittävät tiukemman ehdon $n_{12} + n_{21} >$

20. Fagerlandin ym:iden (2013, 2014) simulointikokeissa testin merkitsevyystaso on varsin lähellä oikeaa, jos reunahomogeenisuushypoteesiehdon $\pi_{+1} = \pi_{1+}$ reunatodennäköisyydet ovat välillä $0.1 - 0.9$ ja $n_{12} + n_{21} \geq 15$.

Esimerkki. Parisuhdeväkivalta (jatkoa). Testataan merkitsevyystasolla 0.01, kertovatko lapset eri todennäköisyyksillä äitiinsä ja isäänsä kohdistuvaa parisuhdeväkivaltaa. Havaitut todennäköisyydet ovat 0.088 ja 0.056 ja havaittujen lukumäärien $n_{12} = 674$ ja $n_{21} = 231$ summa ylittää 20 (jakso 10.2.3). Jakaumalikiarvoistus on käytettävissä. McNemarin testisuure on

$$\frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{674 - 231}{\sqrt{674 + 231}} = 14.72582.$$

Sen p-arvo on R:n raportointitarkkuudella 0 (`2*(1-pnorm(14.72582))`). R-koodi

```
nahnyt <- matrix(c(516,231,674,12038),nrow=2)
mcnemar.test(nahnyt,correct=F))
```

palauttaa yhtäpitävästi testisuureen arvon $216.85 \approx 14.72582^2$ ja nollasta 16. desimaalissa poikkeavan p-arvon. (Jälleen `correct=F` tarkoittaa, ettei tehdä ns. jatkuvuuskorjausta.) Ero osuuksissa 0.088 ja 0.056 on tilastollisesti merkitsevä. Lapset ilmaisevat enemmän isän äitiin kuin äidin isään kohdistamaa väkivaltaa. □

Fagerland ym. (2013 ja 2017, 383–384) suosittelevat päätekstissä kuvattua normaalisuuslikiarvoistukseen perustuvan testin käyttöä kaikissa tilanteissa. Sen koko on yleensä lähellä oikeaa, jos $n \geq 15$. Pienemmillä havaintomäärillä sen koko on oikeampi kuin muiden testien. Se on kaikissa tilanteissa voimakkaampi tai yhtä voimakas kuin muut testit. Agresti ja Finlay (2009, 202) neuvovat käyttämään binomijakauman kriittisiä arvoja, jos ehto $n_{12} + n_{21} > 20$ ei täyty. Ohje ei ole hyvä. Fagerland ym. (2013) arvioivat, että tällaisen testin todellinen merkitsevyystaso on niin paljon tarkoitettua pienempi ja testi niin heikko, että sitä ei tulisi käyttää.

12.2 χ^2-testejä

Vuonna 1900 Karl Pearson julkaisi χ^2-testin. Sitä on sanottu yhdeksi 20. vuosisadan tärkeimmistä keksinnöistä. Moni pukertaa sellaisen kandidaatintutkielmaansa. Samaa testiä käytetään Euroopan hiukkasfysiikan tutkimuskeskus CERNissä vahvistamaan ydinfysiikan viimeisimpiä saavutuksia. Testiä voidaan käyttää monenlaisten hypoteesien testauksessa.

12.2.1 Otosjakauman ja teoreettisen jakauman yhteensopivuustesti

Tutkitaan satunnaismuuttujia N_i, jotka noudattavat multinomijakaumaa $\mathsf{Mul}(n, \pi_1, \ldots, \pi_c)$. Solutodennäköisyydet π_i määräytyvät teoriasta tai hypoteesista, jonka pitävyyttä halutaan koetella. Taustalla voi olla jatkuva jakauma, jonka arvot on luokiteltu. Nollahypoteesi asettaa parametreille arvot $\pi_1 = \pi_{10}$, $\ldots, \pi_c = \pi_{c0}$.

Nollahypoteesin voimassaollessa *i.* solun odotettu lukumäärä on

$$e_i \equiv \mathsf{E}(N_i) = n\pi_{i0}$$

(jakso 7.1.4). Havaittujen lukumäärien n_i ja odotettujen lukumäärien $n\pi_{i0}$ ei tulisi poiketa suuresti toisistaan nollahypoteesin pätiessä. Testisuure

$$X^2 = \sum_{i=1}^{c} \frac{(N_i - e_i)^2}{e_i} \overset{n \text{ suuri}}{\sim} \chi^2(c-1) \tag{12.3}$$

perustuu tälle ajatukselle. Mikäli nollahypoteesi ei päde, havaittujen ja odotettujen lukumäärien poikkeamat ja testisuure X^2 suurenevat. Nollahypoteesin pätiessä X^2 noudattaa suurilla havaintomäärillä χ^2-jakaumaa $c-1$:llä vapausasteella. Suuret arvot johtavat nollahypoteesin hylkäämiseen. Testiä kutsutaan *χ^2-testiksi*.

Jakauma-approksimaation toimivuudelle on esitetty monia ohjenuoria kuten, että kaikki odotetut lukumäärät ovat vähintään yksi ($e_i \geq 1$) ja vähintään 80 % niistä on suurempia kuin viisi ($e_i > 5$). Lindgrenin (1976, 424) mukaan approksimaatio on melko hyvä, jos havaintojen lukumäärä on neljä–viisi kertaa solujen lukumäärä ($n > 4 \times c$) vaikka yksittäisiä yhtä pienempiä odotettuja lukumääriä olisi ($e_i < 1$). Jos sovellettava sääntö ei toteudu, tulee luokkia yhdistää niin, että se toteutuu.

Esimerkki.[134] Poissaolot. Yrityksen johto ajattelee, että sairastamisen tulisi jakautua tasaisesti viikonpäiville. Johto epäilee, että työntekijät ilmoittautuvat sairaaksi tavanomaista useammin viikonloppua ympäröivinä maanantaina ja perjantaina. Johto kerää poissaolotiedot seuraavilta neljältä viikolta:

ma	ti	ke	to	pe	Σ
49	35	32	39	45	200
40	40	40	40	40	200

Maanantaisin ja perjantaisin on muita päiviä enemmän poissaoloja. Alemmalla rivillä on nollahypoteesin tasaisesti jakautuneista sairaspäivistä mukaiset odotetut lukumäärät $n\pi_{i0} = 200 \times 1/5 = 40$. Ovatko poikkeamat odotetuista lukumääristä poikkeavia merkitsevyystasolla 0.05? Johto laskee R-käskyillä

```
lkm <- c(49,35,32,39,45)
chisq.test(lkm,p=c(1/5,1/5,1/5,1/5,1/5))
qchisq(0.95,4)
```

testisuureen arvoksi

$$\frac{(49-40)^2}{40} + \cdots + \frac{(45-40)^2}{40} = 4.900$$

ja $\chi^2(4)$-jakauman 0.95. kvantiiliksi 9.488. Jakauman vapausasteet ovat luokkien lukumäärä miinus yksi eli $5 - 1 = 4$. Nollahypoteesia ei ole syytä hylätä. Käsky `chisq.test(lkm,p=c(1/5,1/5,1/5,1/5,1/5))` tulostaa testisuureen p-arvoksi $0.298 = \mathsf{P}_{H_0}(X^2 > 4.9)$, jossa alaindeksi H_0 osoittaa todennäköisyyden laskemisen nollahypoteesin pätiessä. p-arvon saa myös käskyllä `1-pchisq(4.9,4)`. □

Luokittelu voi vaikuttaa testin tulokseen. Luokat kannattaa muodostaa mahdollisimman osuvaksi kysymyksenasettelun kannalta, koska se kasvattaa testin voimaa.

Esimerkki. Poissaolot (jatkoa). Poissaolojen epäillään keskittyvän viikonlopun ympärille perjantaille ja maanantaille. Nollahypoteesi diskreetistä tasaisesta jakaumasta on luonteva, mutta aineisto kannattaa luokitella poissaoloihin tiistaista torstaihin ja perjantaista maanantaihin, koska näiden luokkien väliseen eroon yrityksen johdon epäilyt kohdistuvat. Uudelleenluokiteltu aineisto ja vastaavat odotetut lukumäärät ovat:

ti–to	ma ja pe	Σ
106	94	200
120	80	200

Yllä $n\pi_{10} = 200 \times 3/5 = 120$ ja $n\pi_{20} = 200 \times 2/5 = 80$.

χ^2-testisuureen arvo on

$$\frac{(106-120)^2}{120} + \frac{(94-80)^2}{80} = 4.083333.$$

Vapausasteita on $2 - 1 = 1$. Tällöin merkitsevyystasoa 0.05 vastaava kriittinen arvo on 3.841 (`qchisq(0.95,1)`). Testisuureen arvo on sitä suurempi, joten nollahypoteesi hylätään merkitsevyystasolla 0.05. Testisuureen p-arvo on

$\mathsf{P}_{H_0}(X^2 > 4.083333) = 0.04330815$ (`1-pchisq(4.083333,1)`). Testin voi tehdä yhtälailla R-käskyillä alla:

```
lkm <- c(106,94)
chisq.test(lkm,p=c(3/5,2/5))
```

Poissaoloissa on viikonloppuvaikutus. Mielekkäällä luokittelulla nollahypoteesi hylätään. □

Voidaan osoittaa, että jakauman yhteensopivuus -testisuure on kahden luokan tilanteessa sama testisuure kuin osuuden testisuure (12.1) neliöitynä:

$$X^2 = \left(\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \right)^2 = z^2$$

(esim. Dudewicz ja Mishra 1988, 529–530). Molemmat noudattavat $\chi^2(1)$-jakaumaa suurilla havaintomäärillä. X^2-testisuure hälyttää yhtälailla liian pienistä kuin suurista lukumääristä odotettuun verrattuna. Jos testi on perusteltua tehdä yksisuuntaisena, testin voimaa voidaan kasvattaa tekemällä yksisuuntainen testi testisuureella z. Luopumalla voimasta yhteen suuntaan voitetaan voimaa toiseen suuntaan.

Esimerkki. Poissaolot (jatkoa). Yrityksen johto epäilee ylimääräisiä — ei muita päiviä vähäisempiä — poissaoloja perjantaisin ja maanantaisin. Testi on järkevintä tehdä yksisuuntaisena testisuureella (12.1).

Olkoon π perjantai- ja maanantaipoissaolojen osuus. Nollahypoteesi on, että $\pi = 2/5 = 0.4$. Vastahypoteesin mukaan $\pi > 2/5$. Havaittu osuus on nollahypoteesin esittämää suurempi:

$$\hat{\pi} = \frac{49 + 45}{200} = 0.47.$$

Testisuure on

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0 \times (1-\pi_0)/n}} = \frac{0.47 - 0.40}{\sqrt{0.40 \times (1-0.40)/200}} = 2.020726.$$

Testisuureen p-arvo on $\mathsf{P}_{H_0}(Z > 2.020726) = 0.0216$ (`1-pnorm(2.020726)`). Nollahypoteesi hylätään merkitsevyystasolla 0.05. Poissaolojen osuus 0.47 on suurempi kuin odotettu 0.4. Viikonlopun yhteydessä on keskimääräistä enemmän poissaoloja. Testisuureen p-arvo 0.02165407 on puolet vastaavan X^2-testisuureen p-arvosta 0.04330815. Puolittuminen saatiin aikaiseksi yksisuuntaisella

testillä. Todetaan lopuksi tarkistuksena, että $z^2 = 2.020726^2 = 4.083334 \approx 4.083333 = X^2$. □

Tyypillisempi on tilanne, että nollajakauman parametrit estimoidaan sovittaen nollajakauma aineistoon. Jos estimoitavia parametreja on s, niin suurilla havaintomäärillä X^2 noudattaa jakaumaa, jonka kriittiset arvot sijaitsevat $\chi^2(c-1-s)$- ja $\chi^2(c-1)$-jakaumien vastaavien kriittisten arvojen välissä (jälkimmäisen jakauman kvantiilit ovat suurempia kuin ensin mainitun). Estimoitu jakauma sopii aineistoon paremmin kuin todellinen jakauma. Vapausasteet ja kriittiset arvot pienenevät siksi nollajakaumassa. Jos nollajakaumana käyttää $\chi^2(c-1-s)$-jakaumaa, niin kriittiset arvot voivat olla liian pienet ja testin merkitsevyystaso liian suuri, jolloin nollahypoteesi hylätään liian usein. Likiarvoistus

$$X^2 \overset{n \text{ suuri}}{\sim} \chi^2(c-1-s)$$

voi silti olla hyvä. (Tang ym. 2012, 41, 151.)

Esimerkki. Perijättärien hedelmällisyys. Francis Galton pohti syitä, miksi suurmiesten suvut Iso-Britanniassa monesti kuolevat pois eli sukuun ei synny poikaa viemään sukunimeä eteenpäin. Taulukossa on syntyneiden poikien lukumäärä aatelisen miehen avioliitossa luokiteltuna sen mukaan, onko puolisonsa ollut perijätär vai ei. (Tyttöjen lukumäärät eivät ole esimerkissä merkityksellisiä.)[135]

Galton tulkitsi taulukkoa tähän tapaan: Avioliitosta perijättären kanssa syntyy vähemmän poikia kuin muista avioliitoista. Perijättäret ovat siten vähemmän hedelmällisiä kuin muut naiset. Syy hedelmättömyyteen on ilmeinen, koska naisesta tulee perijätär vain perheestä, jossa ei ole poikia (aikakautensa lainsäädännön mukaan), eli hedelmättömyys on perijättärien perinnöllinen ominaisuus.

	avioliittojen lkm, joissa äiti	
syntyneitä poikia	perijätär	ei-perijätär
0	11	1
1	8	5
2	11	7
3	11	17
4	5	10
5	3	4
6	1	4
7	0	2
>7	0	0
avioliittojen lkm	50	50
poikien lkm (Σ)	104	168
tyttöjen lkm (Σ)	103	142

Sovitetaan $\mathsf{Poi}(\mu_j)$-jakauma poikien lukumäärille perijätär- ja ei-perijätär-avioliitoissa ($j = 1, 2$). Odotusarvojen (μ_j) estimaatit ovat poikien lukumäärien keskiarvot $\hat{\mu}_1 = 2.08$ ja $\hat{\mu}_2 = 3.36$ (jakso 9.5). Havaitut sekä Poisson-jakaumasta estimoidut pistetodennäköisyydet ja odotetut lukumäärät luokille 0–7 saadaan kaavoista

$$\frac{n_{ij}}{n}, \quad e^{-\hat{\mu}_j}\frac{\hat{\mu}_j^i}{i!} \quad \text{ja} \quad ne^{-\hat{\mu}_j}\frac{\hat{\mu}_j^i}{i!}.$$

Yllä $n = 50$, n_{ij} on havaittu lukumäärä ja $i = 0, \ldots, 7$ ja $j = 1, 2$. Esimerkiksi perijätär-aineistossa havaittu ja estimoitu pistetodennäköisyys ja estimoitu odotettu lukumäärä avioliitoille, joissa on täsmälleen yksi poika, ovat

$$\frac{8}{50} = 0.16, \quad e^{-2.08}\frac{2.08^1}{1} \approx 0.2598548 \quad \text{ja} \quad 50e^{-2.08}\frac{2.08^1}{1} \approx 12.99274.$$

R laskee estimoidut pistetodennäköisyydet kätevästi käskyillä `dpois(0:7,2.08)` ja `dpois(0:7,3.36)`. Luokan "> 7" estimoitu todennäköisyys ryhmälle j on

$$1 - \sum_{i=0}^{7} e^{-\hat{\mu}_j}\frac{\hat{\mu}_j^i}{i!}$$

(R-käskyt `1-ppois(7,2.08)` ja `1-ppois(7,3.36)`).

Näin lasketut havaitut ja estimoidut pistetodennäköisyydet sekä havaitut ja odotetut lukumäärät ovat alla:

	havaittu todennäköisyys		estimoitu todennäköisyys	
	perijätär	ei-perijätär	perijätär	ei-perijätär
0	0.22	0.02	0.124930212	0.03473526
1	0.16	0.10	0.259854841	0.11671047
2	0.22	0.14	0.270249035	0.19607359
3	0.22	0.34	0.187372664	0.21960242
4	0.10	0.20	0.097433785	0.18446603
5	0.06	0.08	0.040532455	0.12396117
6	0.02	0.08	0.014051251	0.06941826
7	0.00	0.04	0.004175229	0.03332076
>7	0.00	0.00	0.001400527	0.02171203

	havaittu lukumäärä		odotettu lukumäärä	
	perijätär	ei-perijätär	perijätär	ei-perijätär
0	11	1	6.2465106	1.736763
1	8	5	12.9927421	5.835524
2	11	7	13.5124518	9.803679
3	11	17	9.3686332	10.980121
4	5	10	4.8716893	9.223302
5	3	4	2.0266227	6.198059
6	1	4	0.7025625	3.470913
7	0	2	0.2087614	1.666038
>7	0	0	0.07002636	1.085602

Koeponnistetaan χ^2-testillä nollahypoteesia, että poikien lukumäärät noudattavat estimoituja Poisson-jakaumia. Luokkia on yhdeksän, joista vain kolmessa odotettu lukumäärä on viittä suurempi. Kummassakaan aineistossa χ^2-jakauma-approksimaation ehto, että 80 % havaituista lukumääristä tulisi olla viittä suurempia, ei täyty. Yhdistetään perijätär-aineistossa viisi viimeistä luokkaa, jotta ehto täyttyy:

	havaittu lukumäärä	odotettu lukumäärä	estimoitu todennäköisyys
0	11	6.2465106	0.124930212
1	8	12.9927421	0.259854841
2	11	13.5124518	0.270249035
3	11	9.3686332	0.187372664
>3	5	7.879662	0.1575932

Luokkia on nyt 5, joista yhdessäkään odotettu lukumäärä ei ole alle 5. Jakauma-approksimaation ehto toteutuu. Estimoituja parametreja on yksi ($\hat{\mu}_1$). X^2-testisuure noudattaa siten jakaumaa, jonka kriittisten arvojen alaraja on $\chi^2(3)$-jakauman $(5 - 1 - 1 = 3)$ kriittiset arvot.

X^2-testisuure uudelleenluokitellulle perijätär-aineistolle on

$$\frac{(11-6.2465106)^2}{6.2465106}+\cdots+\frac{(5-7.879662)^2}{7.879662}=6.4464.$$

Sen p-arvo on 0.092 (`1-pchisq(6.4464,3)`). Nollahypoteesia, että perijättärien poikien lukumäärä noudattaa Poi(2.08)-jakaumaa, ei hylätä. (p-arvo `1-pchisq(6.4464,4)` olisi 0.168, jos nollajakaumana käytettäisiin $\chi^2(4)$-jakaumaa.)

Ei-perijätär aineistossa yhdistetään 2 ensimmäistä ja 3 viimeistä luokkaa, minkä jälkeen luokkia on 6. Tällöin $X^2=5.2816$. Sitä verrataan $\chi^2(4)$-jakaumaan ($6-1-1=4$). Testisuureen p-arvo on 0.260 (`1-pchisq(5.2816,4)`). ($\chi^2(5)$-jakaumasta p-arvoksi `1-pchisq(5.2816,5)` saataisiin 0.382.) Nollahypoteesi Poi(3.36)-jakaumasta jää voimaan.

Testien mukaan Poisson-jakauma vaikuttaa kelvolliselta kuvaukselta poikien lukumäärille perijätär- ja ei-perijätär avioliitoissa. Harjoitustehtävässä pohditaan, voidaanko poikien lukumäärän katsoa noudattavan samaa jakaumaa perijätär- ja ei-perijätär-avioliitoissa. Jaksossa 12.3.1 osoitetaan Poisson-jakautuneisuuden testaamiseen räätälöity testi. □

12.2.2 Otosjakaumien yhteensopivuus- ja satunnaismuuttujien riippumattomuustesti

Aineistona on I:hin luokkaan jaoteltuja havaintoja J-luokkaisesta satunnaismuuttujasta Y ($I\geq 2$, $J\geq 2$):

		Y			
		y_1	$\cdots$	y_J	Σ
X	x_1	n_{11}	$\cdots$	n_{1J}	n_{1+}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	n_{I1}	$\cdots$	n_{IJ}	n_{I+}
	Σ	n_{+1}	$\cdots$	n_{+J}	n

Riviluokkamuuttujaa merkitään yllä X:llä. Havaintoja on yhteensä n, joka on kiinteä luku.[136] Havaittujen lukumäärien n_{ij} taustalla on satunnaismuuttuja N_{ij}. Tällaista aineistoa kutsutaan *ristitaulukoksi* (*cross table*, *contingency table*). On hyvä hahmottaa kaksi tapaa, joilla aineisto on voinut syntyä.

Aineisto on voitu koostaa I:stä riippumattomasta otoksesta multinomijakautuneesta satunnaismuuttujasta Y (*riippumaton multinomiaalinen otanta*).

Otosten koot n_{i+} ovat kiinteitä. Taulukon kullakin rivillä Y on jakautunut tavalla, joka saattaa vaihdella luokkamuuttujan X arvon mukaan. Merkitään Y:n ehdollisia solutodennäköisyyksiä $\pi_{j|i}$:llä. Y:n ehdolliset luokittaiset jakaumat ovat riveillä alla:

		Y			
		y_1	$\cdots$	y_J	Σ
X	x_1	$\pi_{1\|1}$	$\cdots$	$\pi_{J\|1}$	1
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	$\pi_{1\|I}$	$\cdots$	$\pi_{J\|I}$	1

Ehdollisten solutodennäköisyyksien estimaatit ovat $\hat{\pi}_{j|i} = n_{ij}/n_{i+}$ (jakso 9.4).

Vaihtoehtoisesti aineisto on voinut muodostua yhtenä multinomiaalisena otoksena (*multinomiaalinen otanta*), jossa yhden havainnon todennäköisyys osua (i, j)-soluun on π_{ij}:

		Y			
		y_1	$\cdots$	y_J	Σ
X	x_1	π_{11}	$\cdots$	π_{1J}	π_{1+}
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	π_{I1}	$\cdots$	π_{IJ}	π_{I+}
	Σ	π_{+1}	$\cdots$	π_{+J}	1

Tässä molemmat luokkamuuttujat X ja Y ajatellaan satunnaismuuttujiksi. Kaikki reunalukumäärät n_{i+} ja n_{+j} ovat satunnaisia. Solutodennäköisyyksien estimaatit ovat $\hat{\pi}_{ij} = n_{ij}/n$ (jakso 9.4).

Tutkijaa kutkuttava hypoteesi voi olla, että riveittäiset jakaumat ovat identtisiä ehdollisten jakaumien taulukossa: $\pi_{j|1} = \cdots = \pi_{j|I}$. Tällöin j. sarakkeen ehdollisten todennäköisyyksien estimaatit ja estimoidut odotetut lukumäärät ovat

$$\hat{\pi}_{j|i0} = \frac{n_{+j}}{n} \quad \text{ja} \quad n_{i+}\hat{\pi}_{j|i0} = \frac{n_{i+}n_{+j}}{n}.$$

Niissä on merkitty alaindeksillä 0 nollahypoteesin pätiessä estimoitua solutodennäköisyyttä.

Jos $J = 2$, X^2 testaa I:n osuuden yhtäsuuruutta (alempana on numeerinen esimerkki). Näin ollen X^2 on kahden osuuden yhtäsuuruutta testaavan testisuureen (12.2) yleistys.

Multinomijakautuneen yhden otoksen tilanteessa monesti kiinnostava hypoteesi on, että satunnaismuuttujat X ja Y ovat riippumattomia. Tällöin solutodennäköisyys π_{ij} on reunatodennäköisyyksien π_{i+} ja π_{+j} tulo:

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

(kaava (4.10)). Hypoteesin pätiessä estimoitu solutodennäköisyys ja estimoitu odotettu solulukumäärä ovat

$$\hat{\pi}_{ij0} = \hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n^2} \quad \text{ja} \quad n\hat{\pi}_{ij0} = n\frac{n_{i+}n_{+j}}{n^2} = \frac{n_{i+}n_{+j}}{n}.$$

Odotetut estimoidut solutodennäköisyydet ovat $n_{i+}n_{+j}/n$ molemmilla aineiston muodostumistavoilla. Riveittäisten jakaumien samuutta tai satunnaismuuttujien X ja Y riippumattomuutta voidaan testata samalla χ^2-testisuureella

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N_{ij} - e_{ij})^2}{e_{ij}} \overset{n \text{ suuri}}{\sim} \chi^2((I-1)\times(J-1)). \tag{12.4}$$

Siinä odotetut solulukumäärät on laskettu kumman vaan edellä kuvatun nollahypoteesin mukaisesti:

$$e_{ij} = \frac{n_{i+}n_{+j}}{n} = n_{i+}\hat{\pi}_{j|i0} = n\hat{\pi}_{ij0}.$$

Vapausasteiden lukumäärän kaavassa (12.4) voi hahmottaa näin: Riippumattomassa multinomiaalisessa otannassa taulukon jokaisella rivillä on $J-1$ vapaata parametria, sillä yksi parametreista määräytyy ehdosta $\sum_{j=1}^{J}\pi_{j|i} = 1$. Vapaita parametreja on yhteensä $I(J-1)$, koska rivejä on I. Kukin estimoitava parametri vie vapausasteen. Nollahypoteesin pätiessä on vain yksi estimoitava jakauma eli saraketodennäköisyydet. Niitä täytyy estimoida $J-1$ — yksi niistä määräytyy ehdosta $\sum_{j=1}^{J}\pi_{j|i0} = 1$. Vapausasteita on

$$I(J-1) - (J-1) = (I-1)(J-1).$$

Multinomiaalisessa otannassa vapaita parametreja on $IJ-1$ — viimeinen parametri määräytyy ehdosta $\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1$. Riippumattomuushypoteesin pätiessä estimoituja parametreja on $I-1+J-1 = I+J-2$, sillä reunatodennäköisyydet π_{i+} ja π_{+j} summautuvat molemmat 1:ksi. Vapausasteita on tällöinkin

$$IJ - 1 - (I+J-2) = (I-1)(J-1).$$

Jakaumalikiarvoistuksen toimivuudelle on peukalosääntöjä, jotka eivät ole aivan yhtäpitäviä. Jos $I = J = 2$, kaikkien odotettujen lukumäärien pitää olla suurempia kuin 5 (Agresti 2019, 37). Vastaavalle testille jaksossa 12.1.2 esitettiin ehto, että lukumäärien pitää olla suurempia kuin 5 (Agresti ja Finlay 2009, 190). Muille taulukoille Cochranin (1954) ohje on: Likiarvoistus toimii, vaikka yksi odotettu lukumäärä olisi noin 1, jos yli 80 % odotetuista frekvensseistä on ainakin 5 ja vapausasteita on vähintään 2. Toinen ohje on, että kaikkien sarakkeiden ja rivien odotettujen lukumäärien keskiarvon tulisi olla suurempia kuin 5 ja kaikkien odotettujen lukumäärien olla vähintään 1. Kaikkia tilanteita kattavaa ohjetta on vaikea luoda. Agresti (2013, 78) varoittaa, että likiarvoistus saattaa olla huono, jos taulukossa on sekä hyvin pieniä että kohtuullisen suuria odotettuja lukumääriä. Kroonenberg ja Verbeek (2018) arvioivat, että Cochranin ohje ei ole luotettava 2×3 -taulukoille mutta on luotettava, jos vapausasteita on enemmän kuin 2.

Esimerkki. Lääketeollisuuden tutkimukset. *Nature*-lehdessä kerrottiin 2013 tilastotieteen väärinkäytöksestä lääketieteellisessä tutkimuksessa, joka johti lääkeyhtiön toimitusjohtajan tuomioon. Myös arkisemmissa medioissa on kyseenalaistettu lääketieteellisten tutkimusten luotettavuutta. Alla on Iltalehden 8.9.2014, Helsingin Sanomien 25.5.2015, Suomen kuvalehden 1.1.2017 ja Helsingin Sanomien 31.7.2018 uutisointia.[137]

> Professori vertaa lääketeollisuutta järjestäytyneeseen rikollisuuteen. Professori Peter C. Gøtzsche on kohauttanut kirjallaan – – .[138] – – Gøtzsche puhui – – Helsingin yliopistolla – – . – – lääkeyritykset pyrkivät ja useimmiten myös pystyvät kontrolloimaan lääkkeiden kehitystä ja testausta alusta loppuun. – – Niillä on valtava intressi manipuloida tuloksia niin, että tuotteella voitaisiin osoittaa olevan positiivista vaikutusta tai peitellä uuden lääkkeen sivuvaikutuksia. Houkutus vilppiin on liian suuri, kun kukaan ulkopuolinen ei pysty tarkistamaan kokeiden tuloksia ja kun pelissä on helposti miljardien voitot.
>
> — moderneihin sairauksiin tuhlataan aikaa ja rahaa. – – Lääkkeiden vaikutus lonkkamurtumien ehkäisemisessä on niin vähäinen, ettei ehkäisevä lääkehoito ole perusteltua, todetaan British Medical Journalissa — julkaistavassa katsaustutkimuksessa. Sen on tehnyt Helsingin yliopiston professorin Teppo Järvisen johtama kansainvälinen ryhmä. – – katsaus antaa lääketutkimuksista huolestuttavan kuvan. “Yleistrendi oli, että mitä enemmän tutkimuksessa oli puutteita, sitä varmemmin lääkkeellä havaittiin positiivisia vaikutuksia – – . Niitä [moderneja sairauksia] tehdään tarkoitushakuisilla tutkimuksilla”, sanoo Järvinen. Potilasryhmiä ja aineistoa käsitellään niin, että tulokset saadaan lääkkeen kannalta positiivisiksi.
>
> Tiedelehti Lancet julkaisi helmikuussa 2011 mullistavan tutkimuksen. Krooninen väsymysoireyhtymä näytti helpottavan joko kognitiivisella psykoterapialla tai asteittain lisätyllä liikunnalla. – – Tekijöiden kytkökset vakuutuslääketieteeseen arveluttivat– – tilastotieteen professorien avustuksella tehdyt uudelleenanalyysit

> – – julkaistiin syyskuussa – – . – – tulos poikkesi täysin Lancetin julkaisemasta versiosta. [kognitiivisella psykoterapialla] ja – – [asteittain lisätyllä liikunnalla] parani potilaista enää muutama prosentti.
>
> "Jopa puolet lääketieteen nimissä tehtävistä toimista on turhia tai niiden haitat ovat suuremmat kuin hyödyt", sanoo professori Teppo Järvinen Helsingin yliopistosta.

Ylen uutinen 11.8.2010 on esimerkin varsinainen aihe:

> Lääketeollisuuden omat tutkimukset päätyvät positiivisiin tuloksiin selvästi useammin kuin julkisten tai muiden tahojen rahoittamat. – – nyt vinouma havaittiin myös ClinicalTrials-tutkimusrekisterissä, joka perustettiin julkaisuharhan vähentämiseksi.
>
> Julkaisuharha syntyy, kun negatiiviset tutkimustulokset pimitetään ja vain suotuisat havainnot julkaistaan. – – Annals of Internal Medicine -lehdessä julkaistu 546 lääketutkimuksen selvitys paljasti, että 85 prosenttia lääketeollisuuden tutkimuksista päätyi tutkitun lääkkeen kannalta positiiviseen tulokseen, kun niin kävi noin 50 prosentissa viranomaisten rahoittamista. Järjestöjen tai muiden tahojen tutkimuksista 72 prosenttia päätyi suotuisaan tulokseen, mutta osuus suureni selvästi, jos yhtenä rahoittajana oli lääkeyhtiö.
>
> Tutkijat muistuttavat, että julkaisuharha on vain yksi monista seikoista, jotka selittävät rahoittajan ja tulosten yhteyttä. Tuloksia voi muokata itselleen suotuisaksi muun muassa viilaamalla tutkimusasetelmaa tai valitsemalla sopivia potilaita tutkittavaksi – – . Lisäksi lääkeyhtiöt ovat tarkkoja siitä mitä tutkimuksia ne rahoittavat, mikä osaltaan selittää positiivisten tulosten määrää.

Taulukossa alla on osa-aineisto uutisessa viitatusta tutkimuksesta.[139] Lääketeollisuuden tutkimukset päätyivät positiiviseen tulokseen lääkkeen vaikutuksesta $100 \times 188/220 \approx 85.4$ %:ssa tutkimuksista. Viranomaisten tutkimuksissa osuus oli $100 \times 18/36 = 50.0$ %. Tutkitaan χ^2-testillä, eroavatko osuudet 0.854 ja 0.5 tilastollisesti merkitsevästi (otosjakaumien yhteensopivuustesti).

Muodostetaan taulukko havaituista (n_{ij}) ja odotetuista (e_{ij}) solulukumääristä:

	tulos		
rahoitus	+	−	Σ
lääketeollisuudelta	188 (177.03125)	32 (42.96875)	220
viranomaisilta	18 (28.96875)	18 (7.03125)	36
Σ	206	50	256

Odotetut lukumäärät (suluissa) on laskettu kaavalla

$$e_{ij} = \frac{n_{i+} n_{+j}}{n}.$$

Esimerkiksi $e_{11} = 220 \times 206/256 = 177.03125$. Testisuure on

$$\begin{aligned} X^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx \frac{(188 - 177.03125)^2}{177.03125} + \cdots + \frac{(18 - 7.03125)^2}{7.03125} \\ &\approx 24.744. \end{aligned}$$

Testisuure noudattaa suurissa otoksissa $\chi^2((2-1)\times(2-1))$- eli $\chi^2(1)$-jakaumaa. Sen kriittinen arvo merkitsevyystasolla 0.001 on 10.828 (`qchisq(0.999,1)`). Testisuureen p-arvo on noin 6.54×10^{-7} (`1-pchisq(24.744,1)`). Nollahypoteesi hylätään merkitsevyystasolla 0.001, sillä $24.744 > 10.828$. Testi on nopsaan tehty R-komennoilla alla (`correct=F`: ei tehdä ns. jatkuvuuskorjausta):

```
taulukko <- matrix(c(188,18,32,18),nrow=2)
chisq.test(taulukko,correct=F)
chisq.test(taulukko)$expected
```

Kolmas rivi tuottaa ylle kirjatut odotetut lukumäärät.

Ero on tilastollisesti merkitsevä. Lääketeollisuuden tutkimukset päätyvät positiiviseen tulokseen useammin kuin viranomaisten rahoittamat tutkimukset. □

Voidaan osoittaa, että 2×2 -taulukon tilanteessa X^2-testisuure on sama kuin testisuure (12.2) neliöitynä:

$$X^2 = \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})} = z^2.$$

Molemmat noudattavat $\chi^2(1)$-jakaumaa suurissa otoksissa. X^2-testisuureella on voimaa, poikkeavat osuudet suuntaan tai toiseen toisistaan. Jos poikkeama vain toiseen suuntaan on mielekäs, testi kannattaa tehdä yksisuuntaisena testisuureella z.

Esimerkki. Lääketeollisuuden tutkimukset (jatkoa). Lääketeollisuuden rahoittamissa tutkimuksissa ei ole syytä epäillä, että lääkkeet osoittautuisivat keskimääräistä harvemmin toimiviksi. Testi on perusteltua tehdä yksisuuntaisena.

Mielletään aineisto kerätyksi erikseen lääketeollisuuden tutkimuksista ja viranomaisten tekemistä tutkimuksista (riippumaton multinomiaalinen otanta eli analyysi ehdollistetaan rivilukumäärille). Verrataan siis ehdollisia jakaumia:

	tulos		
rahoitus	+	−	Σ
lääketeollisuudelta	0.8545455	0.1454545	1
viranomaisilta	0.5	0.5	1
Σ	0.8046875	0.1953125	1

Nollahypoteesin mukainen positiivisen tutkimustuloksen tuottaneiden tutkimusten osuus on

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2} = \frac{188 + 18}{220 + 36} = 0.8046875.$$

Testisuure on

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{0.8545455 - 0.5}{\sqrt{0.8046875 \times 0.1953125 \times (\frac{1}{220} + \frac{1}{36})}}$$
$$\approx 4.974345.$$

Testisuure noudattaa standardinormaalijakaumaa suurilla havaintomäärillä. Kriittinen arvo yksisuuntaisessa testauksessa merkitsevyystasolla 0.01 on 2.326 (`qnorm(0.99)`). Havaittu testisuure on kriittistä arvoa suurempi, joten nollahypoteesi hylätään merkitsevyystasolla 0.01. Testisuureen p-arvo on puolittunut noin 3.27×10^{-7}:ksi (`1-pnorm(4.974345)`). Lääketeollisuus saa lääkkeen toimivuudesta positiivisen tuloksen useammin kuin viranomaiset.

Tarkistus: $z^2 = 4.974345^2 \approx 24.744 = X^2$. Tässä lasketun testisuureen neliö on edellisessä esimerkissä laskettu X^2-testisuure. □

Joskus ei ole selvää, onko otos muodostunut riippumattomalla multinomiaalisella vai multinomiaalisella otannalla. Ero ei ole välttämättä tärkeä. Testisuureen arvo (kaava (12.4)) ja sen nollajakauma eivät riipu siitä, kumpi otantamenetelmä on ollut, ja testin johtopäätös esimerkiksi nollahypoteesin voimaan jäämisestä voi tällöin olla mielekäs yhtälailla tulkittuna jakaumien yhteensopivuutena tai satunnaismuuttujien riippumattomuutena.

Jos 2×2 -taulukko koostuu kaltaisista pareista (jaksot 10.2.3 ja 12.1.3), otos on multinomiaalinen ja χ^2-testi on tulkittavissa vain riippumattomuustestinä. Silti kaltaisista pareista lasketusta χ^2-testisuureesta tehdään toisinaan virheellisesti päätelmiä osuuksista eli jakaumien yhteensopivuudesta (harjoitustehtävä).

Mikäli jakaumalikiarvoistuksen pätevyyden ohje ei toteudu, voidaan voidaan laskea hypergeometrista jakaumaa eksaktisti noudattava Fisherin testisuure ja sen keski-p-arvo (2×2

-taulukko) tai tämän Fisherin eksaktin testin yleistävä testi. (Agresti 2007, 2013, 2019.) 2×2 -taulukon tilanteeseen E. S. Pearson (1947) ja Campbell (2007) suosittelevat testisuureen kertomista $(n-1)/n$:llä.

Jos nollahypoteesi on diskreetti tasainen jakauma, niin χ^2-testiä saatetaan voidaan käyttää, vaikka otanta olisi tehty palauttamatta (jakso 7.1.5). Testisuuretta korjataan tällöin sopivalla kertoimella. (McCullagh ja Nelder, 1989, 191–192, Joe 1993.)

12.3 Jakaumatestejä

Luokittelemalla aineisto, laskemalla odotetut lukumäärät kussakin luokassa ja vertaamalla niiden eroa havaittuihin lukumääriin voidaan χ^2-testin avulla tutkia, onko aineisto sopusoinnussa tietyn jakauman kanssa. Voimakkaampi testi hyödyntää yleensä tietoa, mikä on tutkittava jakauma. Tieto jakaumasta voi olla itsessään mielenkiintoinen, tai siitä voi olla käytännön hyötyä esimerkiksi epätavallisten havaintojen seulonnassa. Jakaumatesti saatetaan tehdä myös koettelemaan jakaumaoletusta, johon muu tilastollinen analyysi perustuu. Koetteleminen ei ole aina tarpeen, sillä muu tilastollinen analyysi voi olla kelvollista, vaikka jakaumaoletus ei täsmälleen pätisi. Jaksossa esitetään testit Poisson- ja normaalijakautuneisuuden tutkimiseen. Jakaumatestien nollahypoteesi on, että satunnaismuuttuja noudattaa testattavaa jakaumaa.

12.3.1 Testi Poisson-jakautuneisuudelle

Jos Poisson-jakauma $\mathsf{Poi}(\mu)$ kuvaa aineistoa, tulisi *hajontaindeksin* (*index of dispersion*) $s^2/\hat{\mu}$ olla karkeasti 1. Voidaan osoittaa, että suurilla havaintomäärillä pätee

$$\frac{(n-1)s^2}{\hat{\mu}} \sim \chi^2(n-1). \tag{12.5}$$

Intuitiota kaavalle saa normaalijakauman varianssin estimaattorin jakaumasta (9.3). Kaavojen (9.3) ja (12.5) osoittajissa on varianssin estimaattori. Poisson-jakauma muistuttaa suurilla havaintomäärillä normaalijakaumaa (jakso 7.4.5), joten osoittajissa oleellisesti estimoidaan normaalijakautuneen satunnaismuuttujan varianssia. Poisson-satunnaismuuttujan varianssi on μ (kaava (7.10)), eli kaavojen (9.3) ja (12.5) nimittäjissä on tarkentuva estimaattori normaalijakautuneen satunnaismuuttujan varianssille (jaksot 9.5 ja 9.6).

Likiarvoistus (12.5) on hyvä, jos $\hat{\mu} > 5$ ja vielä monesti käyttökelpoinen, jos $\hat{\mu} > 2$ ja $n > 15$ (Armitage ym. 2002, 235). Testisuureen $(n-1)s^2/\hat{\mu}$ suuret arvot

johtavat nollahypoteesin Poisson-jakautuneisuudesta hylkäämiseen. Testisuure on helppo laskea ja testi kätevä tehdä.

Pienillä havaintomäärillä voidaan soveltaa Fisherin (1950) kehittämää testiä. Poisson-jakautuneisuuden testaamisesta yleispätevällä χ^2-testillä on esimerkkejä jaksossa 12.2.1 ja Krishnamoorthyn (2016, 92–94) kirjassa. Ylipäänsä on suositeltavampaa soveltaa tiettyyn tarkoitukseen räätälöityä testiä, jos sellainen on käytettävissä.

12.3.2 Testi normaalijakautuneisuudelle

Normaalijakautuneisuuden testaamiseksi on lukuisia testejä. Jakson 12.2 jakoavaimenluonteinen χ^2-testi olisi mahdollinen muttei ylipäänsä hyvä. Thoden (2002, 153) mukaan sen käyttöä tulisi välttää normaalisuuden testauksessa. Beran ym:iden (2016) simulointikokeissa se hävisi voimakkuudessa monille normaalisuustesteille. Ylipäänsä kannattaa käyttää nimenomaista tarkoitusta varten laadittua testiä, jos sellainen on olemassa. Jaksossa opitaan yksi eniten käytetyistä normaalisuustesteistä. Tieto normaalisuudesta voi olla erittäin hyödyllinen.

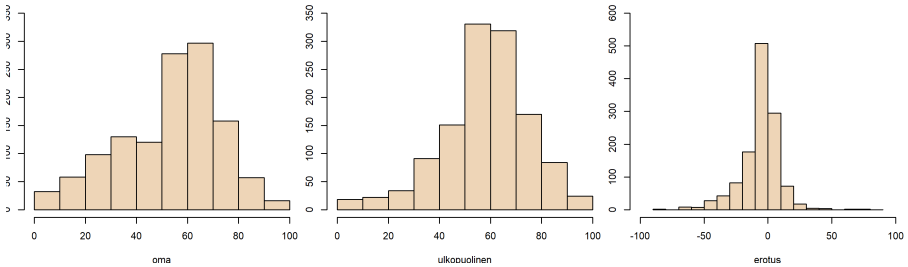
Esimerkki. Seulonta. Erityisryhmiä seulotaan huippulahjakkaiden valmennusryhmään, tukiopetukseen, seurantariskiryhmään, jatkotutkimuksiin mahdollisesta sairaudesta jne. Tieto jakaumasta ja sopivasta seulontarajasta tarvitaan, jotta voidaan ohjata seuloa oikeat ihmiset, ennustaa seulottujen lukumäärää ja arvioida ennusteen luotettavuutta. Ennuste tarvitaan tarjottavan palvelun laadun takaamiseksi ja kustannusten kontrolloimiseksi. Normaalijakauma on monesti luonteva lähtökohta seulonnalle. □

Järjestetään aineisto $x_1, \ldots, x_n$ suuruusjärjestykseen $x_{(1)}, \ldots, x_{(n)}$. Tässä $x_{(i)}$ on aineiston suuruusjärjestyksessä i. havainto ($i = 1, \ldots, n$). Merkitään $o_{(i)}$:llä $x_{(i)}$:n odotusarvoa nollahypoteesin normaalijakautuneisuudesta pätiessä. Shapiro–Wilk-testisuure (SW) on likimain $x_{(i)}$:n ja $o_{(i)}$:n otoskorrelaation neliö:

$$\mathsf{SW} \approx \hat{\rho}^2_{x_{(i)}, o_{(i)}}.$$

Mitä suurempi korrelaation neliö on, sitä paremmin aineisto istuu normaalijakaumaan. Pienet korrelaation neliön arvot johtavat nollahypoteesin hylkäämiseen. R-komento `shapiro.test(x)` laskee testisuureen ja sen likimääräisen p-arvon (aineisto on luettu muuttujaan `x`).

Esimerkki.[140] Kuvassa 12.1 on histogrammit 1 244 suomalaisen omasta arviosta viehättävyydestään ja ulkopuolisen arviosta siitä asteikolla 0–100 (100 on viehättävin mahdollinen) sekä arvioiden erotus. Ulkopuoliset vaikuttavat arvioivan koehenkilöt viehättävämmäksi kuin he itse: Oma arvio on keskimäärin 53.7, ulkopuolisen 58.7 ja arvioiden erotuksen keskiarvo on -5.0.



Kuva 12.1: Koehenkilöiden arvio viehättävyydestään, ulkopuolisen arvio ja arvioiden erotus.

SW-testisuureiden arvot ovat 0.96724, 0.97438 ja 0.90572 vastaten noin korrelaatioita $\sqrt{0.96724} \approx 0.983$, $\sqrt{0.97438} \approx 0.987$ ja $\sqrt{0.90572} \approx 0.952$. Korrelaatiot ovat melkoisia, mutta suurten otoskokojen takia testillä on voimaa ja testisuureiden p-arvojen 13–15 ensimmäistä desimaalia ovat nollia. Nollahypoteesi normaalijakaumasta tulee kirkkaasti hylätyksi kunkin muuttujan kohdalla. Tulos on odotettu varsinkin kahdelle ensin mainitulle muuttujalle, koska millekään välille rajoitetut satunnaismuuttujat eivät voi olla normaalijakautuneita. □

Normaalisuustestejä käytetään joskus esitesteinä ennen varsinaista normaalisuuden olettavaa testiä. Tällaiset menettelyt eivät ole välttämättä toimivia. Normaalisuustesti ei aina hylkää normaalisuusnollahypoteesia, vaikka poikkeama normaalisuudesta riittäisi vääristämään seuraavaksi sovellettavaa testiä (Wilcox 2012, 319).

Tieto tai testi havaintojen normaalijakautuneisuudesta ei ole aina tarpeen. Keskeisen raja-arvolauseen (jakso 7.3) perusteella suurilla havaintomäärillä monen testisuureen voi olettaa noudattavan normaalijakaumaa, vaikka havainnot eivät noudattaisi.

12.4 Odotusarvon ja odotusarvojen erotuksen testaus satunnaismuuttujien ollessa normaalijakautuneita

Jakson testit seuraavat suoraviivaisesti luottamusväleistä jaksossa 10.4. Samaan tapaan kuin siellä tässä jaksossa edetään yksinkertaisimmasta ja empiirisesti epärelevanteimmasta tilanteesta monimutkaisimpaan ja empiirisesti relevanteimpaan tilanteeseen. Kiireisimmät tutustuvat vain kolmeen jaksoon 12.4.2, 12.4.6 ja 12.4.7.

Alla kuvataan kaksisuuntaiset testit merkitsevyystasolla α. Yksisuuntaiset testit voidaan muodostaa vastaavalla tavalla kuin aiemmissa yhteyksissä. Yksinkertaisuuden vuoksi varianssien vertailutesti kuvataan alla yksisuuntaisena (jakso 12.5.2). Jaksossa havainnot on saatu yksinkertaisella satunnaisotannalla. Kahta ryhmää vertailtaessa havainnot ovat lisäksi riippumattomia toisen ryhmän havainnoista paitsi jaksossa 12.4.7.

12.4.1 Testi normaalijakauman odotusarvolle, jos varianssi tunnetaan

Jos $X_i \sim \mathsf{N}(\mu, \sigma^2)$, σ^2 tunnetaan ja nollahypoteesin mukaan $\mu = \mu_0$, niin

$$\frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{N}(0, 1) \tag{12.6}$$

ja

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} < \mathsf{z}_{1-\alpha/2}\right) = 1 - \alpha.$$

Yllä $\hat{\mu} = \sum_{i=1}^{n} X_i/n$ ja $\mathsf{z}_{\alpha/2} = -\mathsf{z}_{1-\alpha/2}$ on standardinormaalijakauman $\alpha/2$. kvantiili (esim. $\mathsf{z}_{0.025} = -\mathsf{z}_{0.975} = -1.960$). Nollahypoteesi hylätään, jos testisuure (12.6) osuu hylkäysalueelle eli $\mid (\hat{\mu} - \mu_0)/(\sigma/\sqrt{n}) \mid > \mathsf{z}_{1-\alpha/2}$.

Esimerkki. R:n `rnorm`-käskyn luotettavuus. R:llä voi tuottaa näennäisesti satunnaisia lukuja standardinormaalijakaumasta `rnorm`-käskyllä. Luvut ovat näennäisesti satunnaisia, koska antamalla sama siemenluku (*seed*), R tuottaa aina samat luvut. Alla oleva koodi tuottaa 100 000 näennäisesti satunnaista lukua normaalijakaumasta. Testataan, imitoiko `rnorm`-käsky hyvin normaalijakaumaa, jonka odotusarvo on $\mu_0 = 1$ (nollahypoteesi). Koodi alla asettaa lisäksi normaalijakauman keskihajonnaksi 1, tuottaa näennäiset satunnaisluvut ja laskee niiden keskiarvoksi 0.9968141:

```
set.seed(21042016) # luku suluissa on mielivaltainen siemenluku
x <- rnorm(n=100000,mean=1,sd=1)
mean(x)
```

Testisuureen arvo on

$$\frac{\hat{\mu}-\mu_0}{\sigma/\sqrt{n}} = \frac{0.9968141-1}{1/\sqrt{100000}} = -1.007481.$$

Se on $\mathsf{N}(0,1)$-jakauman 0.1568518. kvantiili (`pnorm(-1.007481)`). Kaksisuuntaisessa testauksessa p-arvo on noin 0.314. Ei ole syytä hylätä nollahypoteesia, että `rnorm`-komennolla tuotetut luvut matkivat satunnaismuuttujaa, jonka odotusarvo on 1. □

12.4.2 Testi normaalijakauman odotusarvolle, jos varianssia ei tunneta

Jos varianssia σ^2 ei tunneta, kaavaa (12.6) vastaava testisuure on

$$\frac{\hat{\mu}-\mu_0}{s/\sqrt{n}} \sim \mathsf{t}(n-1)$$

($s^2 = \sum_{i=1}^n (X_i-\hat{\mu})^2/(n-1)$). Pätee

$$\mathsf{P}\left(\mathsf{t}_{\alpha/2}(n-1) < \frac{\hat{\mu}-\mu_0}{s/\sqrt{n}} < \mathsf{t}_{1-\alpha/2}(n-1)\right) = 1-\alpha.$$

Tässä $\mathsf{t}_{\alpha/2}(n-1)$ ja $\mathsf{t}_{1-\alpha/2}(n-1)$ ovat $\mathsf{t}(n-1)$-jakauman $\alpha/2$. ja $(1-\alpha/2)$. kvantiilit. Nollahypoteesi H_0: $\mu=\mu_0$ hylätään, jos $|\ (\hat{\mu}-\mu_0)/(s/\sqrt{n}\ )| > \mathsf{t}_{1-\alpha/2}(n-1)$. Testiä kutsutaan *Studentin t-testiksi*. William Gosset julkaisi sen 1908 nimimerkillä Student.

Esimerkkejä testisuureen käytöstä on jaksoissa 11.2 ja 12.4.7. Testin voi tehdä `t.test(x,mu=1)` tapaisen komennon avulla, kun aineisto `x` on ensin luettu R:ään (μ_0 asetetaan `mu`-ohjeella).

12.4.3 Testi normaalijakaumien odotusarvojen erotukselle, jos varianssit yhtäsuuria ja tunnetaan

Oletetaan, että $X_{1j} \sim \mathsf{N}(\mu_1,\sigma^2)$, $X_{2j} \sim \mathsf{N}(\mu_2,\sigma^2)$ ja että niiden yhteinen varianssi σ^2 on tiedossa. Nollahypoteesi on H_0: $\mu_1-\mu_2=(\mu_1-\mu_2)_0$, jossa alaindeksillä on merkitty nollahypoteesin mukaista arvoa odotusarvojen erotukselle

(tyypillisesti 0). Koska

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{\sigma\sqrt{1/n_1 + 1/n_2}} < \mathsf{z}_{1-\alpha/2}\right) = 1 - \alpha,$$

niin nollahypoteesi hylätään, jos $| [\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0]/(\sigma\sqrt{1/n_1 + 1/n_2}) | > \mathsf{z}_{1-\alpha/2}$.

12.4.4 Testi normaalijakaumien odotusarvojen erotukselle, jos varianssit erisuuria ja tunnetaan

Olkoot satunnaismuuttujien $X_{1j} \sim \mathsf{N}(\mu_1, \sigma_1^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma_2^2)$ varianssit erisuuria ja tiedossa. Yhtälöstä

$$\mathsf{P}\left(\mathsf{z}_{\alpha/2} < \frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < \mathsf{z}_{1-\alpha/2}\right) = 1 - \alpha$$

seuraa, että H_0: $\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ hylätään, jos $| [\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0]/(\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}) | > \mathsf{z}_{1-\alpha/2}$.

12.4.5 Testi normaalijakaumien odotusarvojen erotukselle, jos varianssit yhtäsuuria ja tuntemattomia

Jos tiedetään, että satunnaismuuttujien $X_{1j} \sim \mathsf{N}(\mu_1, \sigma^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma^2)$ varianssit ovat yhtäsuuria, niin varianssi estimoidaan kaavalla

$$s^2 = \frac{\sum_{j=1}^{n_1}(X_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(X_{2j} - \hat{\mu}_2)^2}{n_1 + n_2 - 2}.$$

Jos testisuureen

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{s\sqrt{1/n_1 + 1/n_2}} \sim \mathsf{t}(n_1 + n_2 - 2) \tag{12.7}$$

itseisarvoksi tulee $\mathsf{t}_{1-\alpha/2}(n_1+n_2-2)$:tä suurempi arvo, H_0: $\mu_1-\mu_2 = (\mu_1-\mu_2)_0$ hylätään. R-komento `t.test(x1,x2,var.equal=TRUE)` tekee testin laskutoimitukset (`x1`:n ja `x2`:n sisältäessä otosten havainnot).

Testiä tulee käyttää vain, jos on selkeä peruste, että varianssit ovat yhtäsuuria. Ei ole harvinaista, että ensin testataan, ovatko varianssit yhtäsuuria

(jakso 12.5.2), ja jos nollahypoteesi varianssien yhtäsuuruudesta jää voimaan, jatketaan testaamaan odotusarvojen yhtäsuuruutta tässä esitetyllä tavalla. Tällainen menettely voi johtaa aivan väärään testin merkitsevyystasoon (Wilcox 2012, 319, Rasch ja Schott 2018, xiii, 127, Rasch, Verdooren ja Pilz 2020, 67).

12.4.6 Testi normaalijakaumien odotusarvojen erotukselle, jos varianssit erisuuria ja tuntemattomia

Jos $X_{1j} \sim \mathsf{N}(\mu_1, \sigma_1^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma_2^2)$ ja varianssit ovat (mahdollisesti) erisuuria ja tuntemattomia, testisuureen

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

jakauma riippuu suhteesta σ_1^2/σ_2^2 sekä otoskoista n_1 ja n_2 ($s_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \hat{\mu}_i)^2/(n_i - 1)$, $i = 1, 2$). Jakauma on approksimatiivisesti t-jakauma ν vapausasteella, jossa vapausasteet ν määritellään kaavalla (10.10). Testin merkitsevyystaso on osapuilleen $1-\alpha$. Testiä kutsutaan tässä *Smithin–Welchin–Satterthwaiten testiksi*. Kirjallisuudessa nimitys vaihtelee (esim. Gorroochurn 2016, 480). Jos molemmissa otoksissa havaintoja on enemmän kuin 30, nollajakaumana voidaan käyttää standardinormaalijakaumaa (Ramachandran ja Tsokos 2020, 280).

Esimerkki. Kännykkään puhuminen ja reaktioaika.[141] Koehenkilöt ($n_1 = n_2 = 32$) ajoivat autosimulaattoria. Simulaattorissa välähti sattumanvaraisesti punainen tai vihreä valo. Koehenkilöiden tuli painaa jarrupoljinta heti nähtyään punaisen valon. Ensimmäisessä kokeessa koehenkilöt puhuivat puhelimessa ajaessaan. Toisessa kokeessa he kuuntelivat radio-ohjelmaa tai äänikirjaa kun ajoivat. Kokeissa kirjattiin koehenkilöiden reaktioajat punaisen valon välähdykseen millisekunneissa. Keskimääräinen reaktioaika ensimmäisessä kokeessa 585.1875 oli pidempi kuin toisessa kokeessa 534.5625. Vastaavat otosvarianssit olivat 8036.415 ja 4415.286. Testattava nollahypoteesi on H_0: $\mu_1 - \mu_2 = 0$, ja testisuure on

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{585.1875 - 534.5625}{\sqrt{8036.415/32 + 4415.286/32}} = 2.566408.$$

Nollajakauma on t-jakauma vapausasteilla

$$\nu = \mathsf{int}\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right] = \mathsf{int}\left[\frac{\left(\frac{8036.415}{32}+\frac{4415.286}{32}\right)^2}{\frac{\left(\frac{8036.415}{32}\right)^2}{32-1}+\frac{\left(\frac{4415.286}{32}\right)^2}{32-1}}\right]$$
$$= \mathsf{int}[57.16537] = 57.$$

Testisuureen p-arvo 0.013 saadaan t(57)-jakaumasta (`2*(1-pt(2.566408, 57))`). Peukalosäännön edellä mukaan nollajakaumana voi käyttää standardinormaalijakaumaa, jos otosten koot ovat yli 30. Sääntö täyttyy esimerkissä. Standardinormaalijakaumasta laskettu p-arvo 0.010 on varsin sama kuin juuri laskettu. p-arvot ovat evidenssiä nollahypoteesia vastaan. Kännykkään puhuminen vaikuttaa pidentävän reaktioaikaa radion kuuntelua enemmän.

Huom! Otosten koehenkilöt olivat samoja, eli otokset eivät olleet riippumattomia eikä käytetty testi sopiva. Laskut edellä vain havainnollistavat testin käyttöä. Jaksoissa 12.4.7 ja 15.1.1 sovelletaan tällaiseen testausasetelmaan sopivaa testiä esimerkin aineistoon. Vertailu testiasetelman tässä ja jaksossa 12.4.7 välillä tuo jälkimmäisen etua esille. □

Esimerkki. Poikien ja tyttöjen suorituserot. Lapsiasiavaltuutetun vuosikirjasta 2014:[142]

> Valtaosa suomalaislapsista selviytyy PISA-lukutaitotestistä hyvin, mutta viime vuosina aiempaa suurempi osa 15-vuotiaista on saanut testistä heikkoa lukutaitoa ilmentävän pistemäärän. Etenkin pojista yhä suurempi osa suoriutuu testistä heikosti. Vuonna 2012 pojista 18 prosenttia oli luokiteltavissa heikkoihin lukijoihin. Tytöistä heikosti luki viisi prosenttia. Vuosituhannen vaihteessa vastaavat prosenttiosuudet olivat pojilla yksitoista ja tytöillä kolme. – – Suomessa sukupuolten välinen ero on OECD-maiden suurin. Lukutaidon eriarvoisuus näyttää lisääntyneen, sillä lukutaitotestipistemäärän aiemmin alhainen keskihajonta vastaa Suomessa nyt OECD-maiden keskiarvoa.

PISA-tutkimuksessa verrataan 15-vuotiaiden koulutaitoja OECD-maissa. Taulukossa on suomalaisten poikien ($n_1 = 2\,954$) ja tyttöjen ($n_2 = 2\,856$) matematiikan ja lukemisen koepistemäärien keskiarvot ja -hajonnat vuoden 2012 tutkimuksessa.[143] Oletetaan, että pistemäärät ovat normaalijakautuneita.

	keskiarvo		keskihajonta	
aine/sukupuoli	poika	tyttö	poika	tyttö
matematiikka	541.79	539.21	80.00	73.02
lukeminen	508.39	563.48	83.96	72.45

Testataan merkitsevyystasolla 0.005 (kaksisuuntainen testi) nollahypoteeseja, että matematiikan ja lukemisen kokeiden pistemäärien odotusarvot ovat pojilla ja tytöillä samat (kummankin aineen kohdalla nollahypoteesi on H_0: $\mu_1 - \mu_2 = 0$).

Matematiikan pistemäärille testisuure on

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{541.79 - 539.21}{\sqrt{\frac{80.00^2}{2954} + \frac{73.02^2}{2856}}} \approx 1.284636.$$

Otoskoot ovat niin suuria, että nollajakaumana voidaan käyttää standardinormaalijakaumaa. Kaksisuuntaisessa testissä merkitsevyystasolla 0.005 kriittiset arvot standardinormaalijakaumasta ovat -2.807 (`qnorm(0.0025)`) ja 2.807. Koska $-2.807 < 1.285 < 2.807$, niin nollahypoteesia ei hylätä. Testisuureen p-arvo on noin 0.199 (`2*(1-pnorm(1.284636))`). Matematiikan kokeen pistemäärän ero on pieni aineistossa eikä ole tilastollisesti merkitsevä. Nollahypoteesia samoista odotusarvoista ei hylätä. Tämä on esimerkki kuvan 11.4 täysin kielteisestä tuloksesta (jakson 11.3 perusteella vastaava luottamusväli peittää nollan).

Lukemispistemääriin sovellettuna testisuure saa arvon -26.804:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{508.39 - 563.48}{\sqrt{\frac{83.96^2}{2954} + \frac{72.45^2}{2856}}} \approx -26.804.$$

Sen p-arvo on 3×10^{-158} (`2*pnorm(-26.804)`). Lukemiskokeen pistemäärän odotusarvo on pojilla pienempi kuin tytöillä. Ero on suuri sekä poikien ja tyttöjen osaamisen kannalta että tilastollisesti. Tämä on esimerkki kuvan 11.4 selkeästi tärkeästä tuloksesta. □

R-komento `t.test(x1,x2, var.equal=FALSE)` suorittaa testin laskutoimitukset.

12.4.7 Testi normaalijakaumien odotusarvojen erotukselle, jos havainnot parittaisia

Edellisissä jaksoissa verrattiin kahden ryhmän odotusarvoja. Kummassakin — toisistaan riippumattomasti kerätyssä — otoksessa oli satunnaisvaihtelua. Se johtui osin siitä, millaisia havaintoja otoksiin oli tullut. Suurissa ryhmissä sattuman vaikutus häviää ja havainnon poikkeuksellisuus kumoutuu toisen poikkeuksellisuudella toiseen suuntaan. Pienillä havaintomäärillä niin ei välttämättä käy.

Jos aineisto voidaan muodostaa *kaltaistetuista pareista* (jakso 10.2.2), sattuman vaikutusta voidaan yleensä pienentää ja testin voimaa kasvattaa. Edelleen oletetaan havaintojen normaalisuus molemmissa tutkittavissa populaatioissa: $X_{1j} \sim \mathsf{N}(\mu_1, \sigma_1^2)$ ja $X_{2j} \sim \mathsf{N}(\mu_2, \sigma_2^2)$, $i = 1, \ldots, n$. Kukin j. havainto ryhmissä 1 ja 2 kytkeytyvät toisiinsa, niin, että niitä on luontevaa ajatella pareina (X_{1j}, X_{2j}). Odotusarvojen vertailutesti muodostetaan erotuksille

$$D_j = X_{1j} - X_{2j}.$$

Nollahypoteesin pätiessä erotusten odotusarvo on μ_{D0} — tyypillisesti 0. Testaustilanne on jaksossa 12.4.2 kuvatunlainen mutta pitäen tutkittavina satunnaismuuttujina D_j:tä. Testisuure on

$$\frac{\hat{\mu}_D - \mu_{D0}}{s_D/\sqrt{n}} \sim \mathsf{t}(n-1), \tag{12.8}$$

jossa $\hat{\mu}_D = \sum_{j=1}^{n} D_j/n$ ja $s_D^2 = \sum_{j=1}^{n}(D_j - \hat{\mu}_D)^2/(n-1)$. Testiä kutsutaan *parittaiseksi t-testiksi*. Testin voi tehdä R-komennon `t.test(x1,x2, paired=TRUE)` avulla.

Havaintojen X_{1j} ja X_{2j} ollessa riippumattomia erotuksen $X_{1j} - X_{2j}$ varianssi on $\mathsf{V}(X_1) + \mathsf{V}(X_2)$. Voidaan osoittaa, että muulloin erotuksen varianssi on

$$\mathsf{V}(X_1) + \mathsf{V}(X_2) - 2\mathsf{C}(X_1, X_2), \tag{12.9}$$

jossa $\mathsf{C}(X_1, X_2)$ on X_1:n ja X_2:n kovarianssi eli korrelaatio kerrottuna muuttujien keskihajonnoilla (jakso 6.3). Kaltaistetuille pareille havaintojen korrelaatio on yleensä positiivinen. Tällöin erotuksen D_j:n varianssi on pienempi kuin riippumattomien havaintojen tilanteessa. Pienempi varianssi johtuu havaintoja sitovien yhteisten tekijöiden kumoutumisesta erotuksessa. Intuitiivisesti pienemmän varianssin tulisi johtaa voimakkaampaan testiin.

Kaksisuuntainen testi voi olla toimiva, vaikka normaalijakaumaoletus ei pätisi. Testi toimii kelvollisesti, jos aineisto ei ole pieni, sen jakauma ei ole hyvin vino eikä todella erikoisia oudokkeja ole. (Agresti ja Finlay 2009, 196.)

Parittaista t-testiä voi käyttää, vaikka otokset olisivat riippumattomia, jos niissä on yhtä paljon havaintoja. Nollahypoteesin odotusarvojen yhtäsuuruudesta pätiessä erotuksen odotusarvo olisi 0 ja testi tulokseen (12.8) perustuen mahdollinen. Järkevämpää on käyttää otosten riippumattomuuden olettavaa testisuuretta (12.7) (tai versiota, joka sallii erisuuret varianssit populaatioissa). Se noudattaa nollahypoteesin pätiessä $\mathtt{t}(n+n-2)$-jakaumaa. Sen vapausasteet ovat kaksinkertaiset parittaisten erojen t-testisuureen jakaumaan $\mathtt{t}(n-1)$ verrattuna. Pienemmät vapausasteet tarkoittavat, että testisuureen varianssi on suurempi, joten parittaisten erojen t-testi on ilmeisesti heikompi.

Jos otokset on mahdollista tuottaa omalla koejärjestelyllä, riittäisikö parien komponenttien pienikin positiivinen korrelaatio motivoimaan parittaisen koejärjestelyn ja parittaisen t-testin käytön? Jos parien havaintojen korrelaatio olisi hyvin pieni, oltaisiin lähellä edellä kuvattua tilannetta, jossa parittainen t-testi on heikompi. Parittaiseen koejärjestelyyn ei ilmeisesti kannattaisi lähteä. Karkea sääntö on, että parittainen t-testi on voimakkaampi, jos parien havaintojen korrelaatio on suurempi kuin 0.25 (Wilcox 2012, 402). Tällöin kannattaisi koejärjestely muotoilla kaltaistetuiksi pareiksi.

Parittainen t-testi voi olla heikompi kuin vastaava riippumattomuuden olettava testi. Niin käy, jos havaintojen korrelaatio on negatiivinen, jolloin erotuksen $X_{1j} - X_{2j}$ varianssi on suurempi kuin riippumattomassa tilanteessa (kaava (12.9)). Tällaiseen koejärjestelyyn ei kannata ryhtyä. Jos jo olemassa oleva aineisto tiedetään tämäntapaiseksi, täytyy parittaista t-testiä käyttää, koska riippumattomat otokset olettava testi ei ole käyttökelpoinen.

Joskus osa havainnoista on parittaisia ja osa ei (esim. koehenkilöitä on osallistunut vain ensimmäiseen tai toiseen mittaukseen). Odotusarvojen eroa voidaan tällöin testata yhdistämällä informaatio parittaisesta osa-aineistosta ja lopuista havainnoista. (Esim. Derrick ja White 2022, Grabchak 2023, Guo ja Yuan 2017, Samawi ja Vogel 2014, Uddin ja Hasan 2020.)

Esimerkki. Lääkkeen teho. Uuden lääkkeen tehoa voitaisiin tutkia muodostamalla arpomalla kaksi ryhmää, joista toinen saisi uutta ja toinen vanhaa lääkettä. Lääkkeellä voitaisiin esimerkiksi pyrkiä pienentämään verenpainetautia sairastavien verenpainetta. Tutkimusjakson jälkeen mitattaisiin ryhmistä, kuinka hyvin lääke on toiminut. Keskiarvojen vertailutestillä (jakso 12.4.6) pääteltäisiin, onko lääkkeiden tehossa eroa.

Voimakkaampaan testiin päästäisiin, jos lääkkeen tehoa verrattaisiin samojen ihmisten välillä. Koehenkilöt käyttäisivät ensin vanhaa lääkettä ja sen jälkeen uutta lääkettä. Mikäli uuden lääkkeen käytön jälkeen verenpaine olisi keskimäärin laskenut, se viittaisi uuden lääkkeen olevan vanhaa tehokkaampi. Tällaisen testin pitäisi olla voimakkaampi kuin edellä kuvatun, koska verenpaineiden koehenkilökohtaisessa vertailussa eliminoituu monia sekoittavia tekijöitä ensimmäiseen tutkimusasetelmaan verrattuna. Jos koehenkilö syö paljon verenpainetta nostavaa suolaa, testin tulokseen vaikuttaa, kumpaan koeryhmään hän päätyy ensimmäisessä tutkimusjärjestelyssä (runsas suolan käyttö kenties kumoaa lääkkeen vaikutuksen), mikä kasvattaa tutkittavan erotuksen varianssia.

Toisessa tutkimusjärjestelyssä ero lääkkeiden toimivuuden välillä selviää, vaikka koehenkilö olisi erityisen suolaisen ruoan ystävä. □

Esimerkki. Kaksoskokeet. Jos identtisille kaksosille tehdään erilaiset kokeet, tulosten ero johtuu muista kuin geneettisistä tekijöistä. Sattumanvaraisten geneettisten tekijöiden vaikutus on eliminoitu. □

Esimerkki. Kännykkään puhuminen ja reaktioaika (jatkoa). Koehenkilöt olivat samoja jakson 12.4.6 esimerkin molemmissa kokeissa. Aineisto tulisi analysoida parittaisena. Reaktioajat ja reaktioaikojen erotus kullakin koehenkilöllä on taulukoitu alla. Koehenkilö 28:n reaktioajat ovat olleet ylivoimaisesti hitaimmat. Reaktioaikojen erotus on hänellä silti vasta kolmanneksi suurin. Parittainen vertailu on palauttanut koehenkilön tulokset samaan suuruusluokkaan muiden kanssa.

koehenk.	1	2	3	4	5	6	7	8	9	10	11	12
känny	636	623	615	672	601	600	542	554	543	520	609	559
radio	604	556	540	522	459	544	513	470	556	531	599	537
erotus	32	67	75	150	142	56	29	84	-13	-11	10	22
koehenk.	13	14	15	16	17	18	19	20	21	22	23	24
känny	595	565	573	554	626	501	574	468	578	560	525	647
radio	619	536	554	467	525	508	529	470	512	487	515	499
erotus	-24	29	19	87	101	-7	45	-2	66	73	10	148
koehenk.	25	26	27	28	29	30	31	32				
känny	456	688	679	960	558	482	527	536				
radio	448	558	589	814	519	462	521	543				
erotus	8	130	90	146	39	20	6	-7				

R-koodi alla laskee erotuksen keskiarvon 50.625 ja otoskeskihajonnan 52.48579:

```
x1 <- c(636,623,615,672,601,600,542,554,543,520,609,559,595,565,573,554,
        626,501,574,468,578,560,525,647,456,688,679,960,558,482,527,536)
x2 <- c(604,556,540,522,459,544,513,470,556,531,599,537,619,536,554,467,
        525,508,529,470,512,487,515,499,448,558,589,814,519,462,521,543)
d <- x1-x2
mean(d)
sd(d)
```

Testisuureen arvo on

$$\frac{\hat{\mu}_D}{s_D/\sqrt{n}} = \frac{50.625}{52.48579/\sqrt{32}} = 5.456301.$$

Sitä verrataan t(31)-jakaumaan. Testisuureen p-arvo on noin 6×10^{-6} (`2*(1-pt(5.456301,31))`). Samat tulokset saadaan R-käskyllä `t.test(x1,x2,paired=TRUE)`. Nollahypoteesi hylätään kaikilla tavanomaisilla merkitsevyystasoilla. Puhelimeen puhuminen hidastaa reaktioaikaa enemmän kuin radion kuuntelu.

p-arvo parittaisten erotusten t-testissä on pienempi kuin samasta aineistosta jaksossa 12.4.6 (oletusten vastaisesti) laskettu odotusarvojen erotuksen testisuureen p-arvo. Se on intuitiivista, sillä parittaisessa vertailussa karsiutuu satunnaistekijöitä pois ja testin voiman kasvaminen on odotettua.

Testi on nyt tehty oikein. Agresti ja Finlay (2009, 196) varoittavat, että aineistosta voidaan silti tehdä vain tunnustelevia päätelmiä, koska se on ilmeisesti kerätty itsevalikoituneella otannalla (jakso 8.4). Erotus d ei välttämättä ole normaalijakautunut (käsky shapiro.test(d) laskee jakson 12.3.2 Shapiro–Wilk-testisuureen p-arvoksi 0.023), mutta Agresti ja Finlay ovat katsoneet testin soveltamisen mahdolliseksi. □

12.5 Varianssin testaus satunnaismuuttujien ollessa normaalijakautuneita

12.5.1 Yhden varianssin testaus

Jaksossa 9.6 todettiin, että

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

(kaava (9.3)). Nollahypoteesi H_0: $\sigma^2 = \sigma_0^2$ hylätään merkitsevyystasolla α, jos $(n-1)s^2/\sigma_0^2 < \chi^2_{\alpha/2}(n-1)$ tai $(n-1)s^2/\sigma_0^2 > \chi^2_{1-\alpha/2}(n-1)$. Tässä $\chi^2_{\alpha/2}(n-1)$ ja $\chi^2_{1-\alpha/2}(n-1)$ ovat $\chi^2(n-1)$-jakauman $\alpha/2$. ja $(1-\alpha/2)$. kvantiilit.

Oletus normaalisuudesta on tärkeä. Jos havainnot eivät ole normaalijakautuneita, testin merkitsevyystaso voi poiketa paljon α:sta.

Esimerkki. R:n `rnorm`-käskyn luotettavuus (jatkoa). Luodaan 100 000 näennäissatunnaislukua koodilla alla. Tutkitaan, kestääkö nollahypoteesi, että luvut olisivat (normaali)jakaumasta, jonka varianssi $\sigma_0^2 = 1$. Koodi tulostaa otosvariansiksi 1.007285:

```
set.seed(21042016)
x <- rnorm(n=100000, mean=1, sd=1)
var(x)
```

Testisuureen arvo on

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{99999 \times 1.007285}{1} = 100727.5.$$

Se on $\chi^2(99\,999)$-jakauman 0.948082. kvantiili (`pchisq(100727.5,99999)`).

Yksisuuntaisessa testauksessa (pidettäessä 1:stä suurempia arvoja mahdollisina) testisuureen p-arvo olisi 0.051918 (`(1-pchisq(100727.5,99999))`). Jos jakauma on epäsymmetrinen, ei ole yksikäsitteistä, miten p-arvo tulisi määritellä kaksisuuntaisessa testauksessa. χ^2-jakauma on näin suurilla vapausasteilla hyvin symmetrinen, joten kerrotaan yksisuuntaisen testin testisuureen p-arvo kahdella. p-arvoksi saadaan näin 0.103836.

Jakauman symmetrisyyden tarkistusta: $\chi^2(99999)$-jakauman 0.05. ja 0.95. kvantiilit ovat 99264.54 ja 100735.7 (`qchisq(0.05,099999)` ja `qchisq(0.95, 99999)`). Ne ovat yhtä kaukana 99999:stä ($99999 - 99264.54 = 736.7325 = 100735.7 - 99999$). Jakauma vaikuttaa symmetriseltä.

Otosvarianssin poikkeama teoreettisesta varianssista ei anna aihetta hylätä nollahypoteesia. R:n `rnorm`-käsky vaikuttaa toimivan tarkoitetulla tavalla varianssilla mitattuna. □

12.5.2 Kahden varianssin testaus

Edellistä tyypillisempi testaustilanne on kahden varianssin yhtäsuuruuden testaus. Havaintojen normaalisuus on edelleen oleellinen oletus jakaumateorian pätemiselle. Yksinkertaisuuden vuoksi jaksossa kuvataan yksisuuntainen testi. Esimerkiksi Armitage ym. (2002, 150–153) selostavat kaksisuuntaisen testauksen.

On laskettu n_1:n ja n_2:n kokoisista riippumattomista otoksista otosvarianssit $s_1^2 > s_2^2$. Nollahypoteesi on, että vastaavat varianssit ovat samat (H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$). Tällöin testisuure

$$\frac{s_1^2}{s_2^2} = \frac{[(n_1-1)s_1^2/\sigma^2]/(n_1-1)}{[(n_2-1)s_2^2/\sigma^2]/(n_2-1)} \sim \mathsf{F}(n_1-1, n_2-1)$$

noudattaa F-jakaumaa vapausasteilla n_1-1 ja n_2-1. Perustelu: Keskimmäisestä muodosta, jakaumatuloksesta (9.3) ja otosten riippumattomuudesta seuraa, että s_1^2/s_2^2 on nollahypoteesin pätiessä suhde riippumattomista $\chi^2(n_1-1)$- ja $\chi^2(n_2-1)$-satunnaismuuttujista vapausasteillaan jaettuna. Tällainen suhde on $\mathsf{F}(n_1-1, n_2-1)$-jakautunut (jakso 7.2.4). Suuret testisuureen arvot johtavat nollahypoteesin hylkäämiseen.

Esimerkki.[144] Poikien ja tyttöjen suorituserot (jatkoa). Harvardin yliopiston taloustieteen professori (1983–1991, 2011–), Maailman pankin pääekonomisti (1991–1993), Yhdysvaltojen valtiovarainministeri (1999–2001), Harvardin yliopiston rehtori (2001–2006), presidentti Barack Obaman nimittämän kansallisen talousneuvoston johtaja (2009–2010) Lawrence Summers käyttää ajoittain terävämpää kieltä kuin tieteessä on tavallista (esim. Summers 1991). Vuonna 2005 hän pohti tieteellisessä seminaarissa, miksi huippuyliopistoissa ja -tutkimuslaitoksissa on vähän naisia. Summers esitti selityksenä, että monet ominaisuudet kuten pituus, paino, taipumus rikollisuuteen, älykkyysosamäärä, matemaattinen lahjakkuus ja tieteellinen kyvykkyys vaihtelevat miehillä enemmän kuin naisilla ja että pienetkin erot keskihajonnoissa sukupuolten välillä johtavat suuriin eroihin sukupuolten välillä poikkeuksellisen lahjakkaiden yksilöiden lukumäärissä. Summersiä syytettiin seksismistä, ja hän joutui pyytämään anteeksi sanomaansa. Summers erosi rehtorin tehtävästään seuraavana vuonna riitauduttuaan yliopiston henkilökunnan kanssa. Summers on sittemmin kritisoinut "absurdia poliittista korrektiutta" ja "totalitarismin hivuttautumista yliopistoihin" mielessä, mistä yliopistoissa on sallittua keskustella.

Eroavatko suomalaisten 15-vuotiaiden poikien ja tyttöjen matematiikan ja lukemisen koepistemäärien varianssit vuoden 2012 PISA-tutkimuksessa? Testataan nollahypoteesia varianssien yhtäsuuruudesta H_0: $\sigma_1^2 = \sigma_2^2$. Testisuureen arvo koepistemäärien variansseille matematiikassa on 1.200:

$$\frac{s_1^2}{s_2^2} = \frac{80.00^2}{73.02^2} \approx 1.200318.$$

F-jakauman vapausasteilla $n_1 - 1 = 2953$ ja $n_2 - 1 = 2855$ kriittinen arvo merkitsevyystasolla 0.001 on 1.122 (`qf(0.999,2953,2855)`). Testisuureen p-arvo on kuuden desimaalin tarkkuudella nolla (`1-pf(1.200318,2953,2855)`). Koska $1.200 > 1.122$, niin nollahypoteesi hylätään merkitsevyystasolla 0.001. Matematiikan koepistemäärän varianssi on suurempi pojilla kuin tytöillä.

Lukemisen otosvariansseista laskettuna testisuure saa arvon 1.343:

$$\frac{s_1^2}{s_2^2} = \frac{83.96^2}{72.45^2} \approx 1.342975.$$

Sen p-arvo on 1×10^{-15} (`1-pf(1.342975,2953,2855)`). Nollahypoteesi varianssien yhtäsuuruudesta kaatuu merkitsevyystasolla 0.001. Myös lukemisen koepistemäärän varianssi on suurempi pojilla kuin tytöillä.

Ukkola ym. (2020, 41, 48) raportoivat vastaavia tuloksia suomalaisille 7-vuotiaille pojille ja tytöille (harjoitustehtävä). He esittävät sekä genetiikkaan että kulttuuriin pohjautuvan selityksen tuloksille. □

12.6 Testejä ilman jakaumaoletusta

Jotkin jakson 12.4 testisuureista voivat olla käyttökelpoisia, vaikkeivät havainnot tulisi normaalijakaumasta.

12.6.1 Odotusarvon testaus, jos jakauma on tuntematon

Keskeisen raja-arvolauseen (jakso 7.3) perusteella

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0, 1),$$

vaikka havainnoista ei tiedettäisi juuri muuta kuin, että ne ovat riippumattomia ja että niillä on odotusarvo ja varianssi. Yllä voidaan varianssi korvata estimaatilla s^2. Odotusarvoa voidaan testata tähän tulokseen tukeutuen.

Kuten approksimoitaessa luottamusvälejä (jakso 10.5.1), likimääräisessä merkitsevyystestauksessa on luontevampaa käyttää standardinormaalijakaumaa paksuhäntäisempää t-jakaumaa nollajakaumana. Muuten sama testisuureen arvo voisi olla johtamatta nollahypoteesin hylkäykseen, jos jakauma oletettaisiin tunnetuksi mutta johtaa hylkäykseen, jos jakauma olisi tuntematon. Tietämättömyyden jakaumasta ei tulisi vahvistaa testiä. Likimääräisenä merkitsevyystestinä sovelletaan siis jakson 12.4.2 t-testiä.

Agresti ja Finlay (2009, 122, 155, 196) mainitsevat otoskoon 30 karkeana rajana, jota suuremmilla havaintomäärillä odotusarvon kaksisuuntainen testaaminen on mahdollista keskeiseen raja-arvolauseen perustuen. He toteavat, että yksisuuntainen testi on epäluotetettava, jos havaintojen jakauma on hyvin vino tai aineistossa on erityisen poikkeavia havaintoja. Wilcoxin (2012, jakso 5.5) mukaan perinteinen viisaus on, että 100 havaintoa riittää takaamaan testin oikean koon.

Wilcox (2012, jakso 5.5) on tehnyt tarkentavia varoittavia simulointeja. Jos havainnot tulevat normaalijakaumaa paksuhäntäisemmästä symmetrisestä jakaumasta, varianssin estimaatti tapaa suureta, mikä pienentää t-testin todellista merkitsevyystasoa ja heikentää testiä. Melko pienetkin poikkeamat normaalisuudesta riittävät kuvattuun. Oudokeilla on vastaava vaikutus. Jos jakauma on vino, testin todellinen merkitsevyystaso voi olla pienempi tai suurempi kuin tarkoitettu. Wilcoxin esimerkit koskevat melko pieniä aineistoja (esim. $n = 20$).

12.6.2 Kahden odotusarvon erotuksen testaus, jos jakauma on tuntematon

Kahden odotusarvon erotusta testattaessa voidaan niinikään tukeutua keskeiseen raja-arvolauseen. Riittävän suurilla havaintomäärillä Smithin–Welchin–Satterthwaiten testin (jakso 12.4.6) todellinen merkitsevyystaso voi olla likimain oikea, jos verrattavien ryhmien satunnaismuuttujien jakaumat ovat identtisiä vaikkeivät normaalijakaumia. Jos jakaumat ovat erilailla vinoja tai niiden varianssit eroavat (eikä normaalisuus päde), testi ei välttämättä toimi hyvin. (Wilcox 2012, 327–328.) Myös oudokit vääristävät testiä (Chihara ja Hesterberg 2019, 250).

12.6.3 Varianssien testaus, jos jakauma on tuntematon

Jaksojen 12.5.1–12.5.2 varianssitestisuureiden jakaumat voivat riippua suuresti siitä, ovatko havainnot normaalijakaumasta vai eivät. Jos eivät ole, kyseisiä varianssitestejä ei tule käyttää.

Varianssien yhtäsuuruutta voi testata, vaikka normaalisuus ei pätisi. Esimerkiksi Ugarte ym. (2016, jakso 11.5.3) osoittavat testausmenetelmän.

12.7 Korrelaation testaaminen

Jaksossa satunnaismuuttujat ovat binormaalijakautuneita (jakso 7.5) ja havaintoparit $(X_1, Y_1), \ldots, (X_n, Y_n)$ ovat satunnaisotos tutkittavasta populaatiosta. Alajaksossa 12.7.2 tutkitaan kahta tällaista toisistaan riippumatonta otosta.

12.7.1 Yhden korrelaation testaaminen

Useimmiten kiinnostava nollahypoteesi on, että satunnaismuuttujat eivät ole korreloituneita (H_0: $\rho = 0$). Nollahypoteesin pätiessä testisuure

$$t = \sqrt{n-2}\frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \sim \mathsf{t}(n-2) \tag{12.10}$$

noudattaa t-jakaumaa $(n-2)$:lla vapausasteella. Yllä $\hat{\rho}$ on kaavan (8.2) määrittelemä otoskorrelaatio. Nollahypoteesi hylätään, mikäli aineistosta laskettu t on hylkäysalueella.

Esimerkki. Kuinka suuri tulee otoskorrelaation olla, jotta nollahypoteesi H_0: $\rho = 0$ voidaan hylätä kaksisuuntaisessa testauksessa merkitsevyystasolla 0.05? Vastaus riippuu havaintomäärästä. Jos havaintoja on 100, pitää otoskorrelaation itseisarvon olla vähintään 0.20. Pienemmillä/suuremmilla havaintomäärillä tarvitaan suurempi/pienempi otoskorrelaation itseisarvo taulukon alla mukaisesti.

n	5	10	25	50	100	200	1000
$\lvert\hat{\rho}\rvert$	0.88	0.63	0.40	0.28	0.20	0.14	0.06

(Taulukon lasku on harjoitustehtävä.) □

Nollahypoteesia puuttuvasta lineaarisesta yhteydestä voidaan testata edellä kuvatulla tavalla myös, jos toinen satunnaismuuttujista on normaalijakautunut ja toinen noudattaa muuta jakaumaa tai on peräti kiinteä (kaava (13.9)).

Joskus halutaan testata nollahypoteesia tietynsuuruisesta korrelaatiosta, jolloin H_0: $\rho = \rho_0$. Jos $n > 50$, testi voidaan perustaa *Fisherin z-muunnokseen*

$$Z = \frac{1}{2} \log \frac{1 + \hat{\rho}}{1 - \hat{\rho}} \tag{12.11}$$

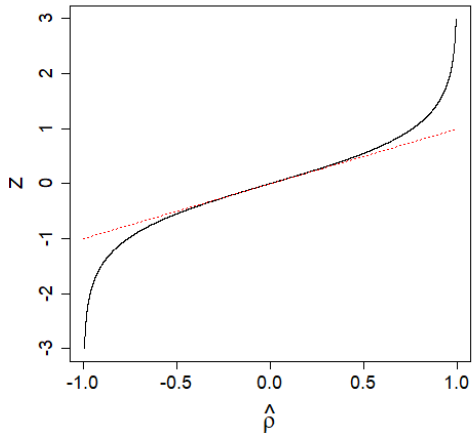
(Lindgren 1976, 478, Stuart ja Ord 1991, 980). Yllä $\hat{\rho}$ tulee mieltää korrelaatiokertoimen estimaattorina eli satunnaismuuttujana.

Kuvassa 12.2 peilataan z-muunnosta ja $\hat{\rho}$:n estimaattia eli toteumaa. Jos $|\hat{\rho}| < 0.4$, niin z ja $\hat{\rho}$ ovat lähes yhtäsuuria. Mitä lähempänä estimaatti $\hat{\rho}$ on ± 1:htä — eli karkeasti mitä vinompi estimaattorin $\hat{\rho}$ jakauma on — sitä enemmän z poikkeaa estimaatista $\hat{\rho}$. Voidaan osoittaa, että muunnos tepsii eli tuottaa estimaattorin $\hat{\rho}$ jakaumaa symmetrisemmän jakauman.

Nollahypoteesin pätiessä Fisherin z-muunnos on likimäärin normaalijakautunut odotusarvolla ja varianssilla

$$\mathsf{E}(Z) = \frac{1}{2} \log \frac{1 + \rho_0}{1 - \rho_0} \quad \text{ja} \quad \mathsf{V}(Z) = \frac{1}{n - 3}.$$

Varianssin riippumattomuus korrelaatiosta ρ_0 liittyy jakauman symmetrisoitumiseen. Jakauma symmetrisoituu jo, kun $n = 10$, jos $-0.8 < \rho < 0.8$ (Cramer



Kuva 12.2: Fisherin z-muunnos (- - - = yhtäsuuruussuora).

1946, 401). Standardoidun testisuureen

$$\frac{z - \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}}{\frac{1}{\sqrt{n-3}}} = \sqrt{n-3}\left(z - \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}\right)$$

arvoa verrataan tavalliseen tapaan standardinormaalijakaumaan (z on Z:n toteuma).

Andersonin (2003, 134) mukaan viritetty testisuure

$$\sqrt{n-3}\left(z - \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0/2}{N-1}\right)$$

on parempi. Suurilla havaintomäärillä se noudattaa nollahypoteesin pätiessä standardinormaalijakaumaa.

12.7.2 Kahden korrelaation testaaminen

Kahden korrelaation ρ_1:n ja ρ_2:n yhtäsuuruutta voidaan testata Fisherin z-muunnosten standardinormaaliuden perusteella:

$$\frac{Z_1 - Z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}}.$$

Z_i on otoskorrelaation $\hat{\rho}_i$ Fisherin z-muunnos (12.11), n_i on otoskoko ja otokset ovat riippumattomia satunnaisotoksia populaatioista i ($i = 1, 2$). Testisuureen nollajakauma on standardinormaali suurilla havaintomäärillä.

Luku 13

Regressio

> Kaikki mallit ovat vääriä, mutta jotkut ovat hyödyllisiä.
> George Box (1919–2013)

13.1 Regressio kohti odotusarvoa

Francis Galtonin hahmottelema ensimmäinen *regressio* vuodelta 1877 on kuvassa 13.1 ("herneen siemen -vanhemmat" ja "herneen siemen -jälkipolvi"). Oleellisesti sama ilmiö on kuvassa 13.2 — ilmeisesti toisessa koskaan tehdyssä regressiossa (Pearson 1930, 13) — jossa on Galtonin vuonna 1886 havaitsema vastaava yhteys vanhempien pituuksien painotetun keskiarvon (*mid-parent*; "keskivanhempi") ja heidän lastensa pituuden välillä.[145] Kuviosta nähdään, että vanhempien keskipituus on ollut keskimäärin runsas 68 tuumaa. Keskivanhempi-suora kuvaa vanhempien keskipituuden poikkeamaa keskimääräisestä pituudesta (suoran kulmakerroin on yksi). Kuvion mukaan

- keskimääräistä pidempien vanhempien lapsi on myös keskimääräistä pidempi muttei yhtä paljon kuin vanhempansa (suoran *children* kulmakerroin on 0:n ja 1:n välillä).
- keskimääräistä lyhyempien vanhempien lapsi on myös keskimääräistä lyhyempi muttei yhtä paljon kuin vanhempansa.
- pituus regressoituu (palautuu, taantuu) eli pyrkii palaamaan kohti odotusarvoansa (yllä runsas 68 tuumaa). (*Regression toward the mean*, *regression to the mean*.) Pitkien vanhempien lapset ovat keskimääräistä pidem-

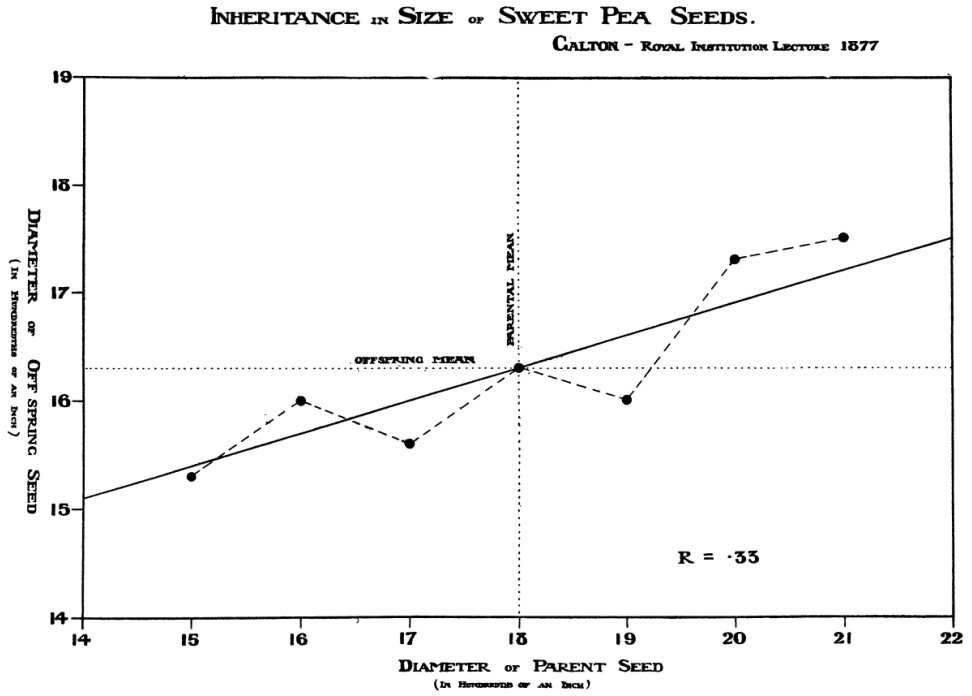
> piä ja lyhyempien vanhempien lapset keskimääräistä lyhyempiä, mutteivät yhtä paljon pidempiä tai lyhyempiä keskipituuteen nähden kuin vanhempansa.

Mieleen saattaisi tulla — kuten Galtonille aikoinaan — että regressiosta keskipituutta kohti seuraisi sukupolvi sukupolvelta pituuden vaihtelun pieneneminen niin, että lopulta kaikki olisivat keskipituisia. Niin ei käy, vaikka lasten pituus keskimäärin regressoituukin vanhempiensa pituudesta. Galton havainnollisti asiaa kuvalla 13.3 vuonna 1901. Kussakin vanhempien pituusluokassa lasten pituuden vaihtelu on pienempää kuin vanhempien pituuden jakaumassa. Silti lasten pituuden jakauma yhtyy vanhempien pituuden jakaumaan. Demidenko (2020, 467–468) vahvistaa pituuden jakauman muuttumattomuuden modernisti simuloimalla. Galtonin keskivanhempi-käsitettä käytetään kasvututkimuksessa edelleen (esim. Saari ym. 2012).

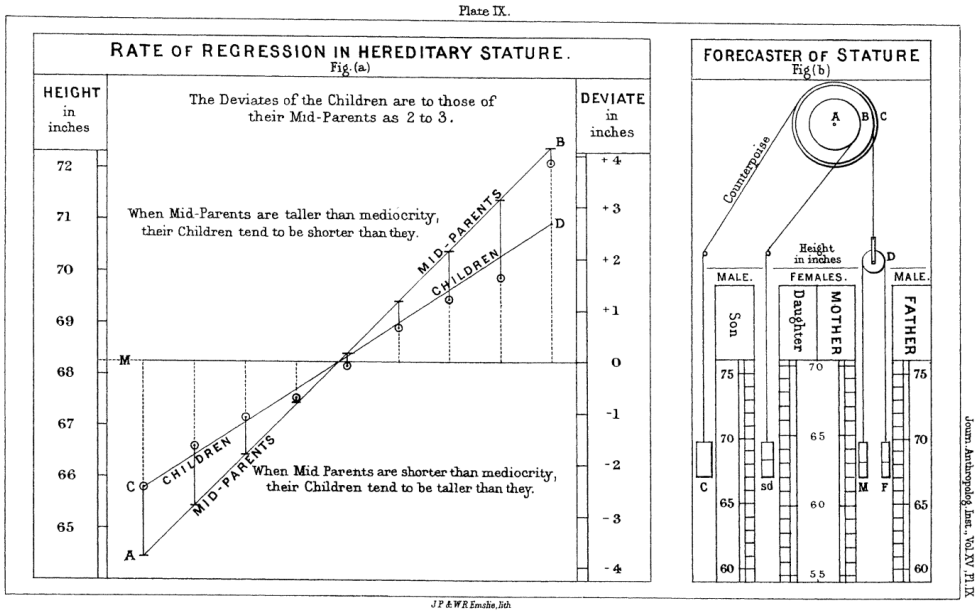
Regressiossa odotusarvoa kohti on edellä kyse kahdesta samoinjakautuneesta muuttujasta, joiden yhteyden summeeraavan suoran kulmakerroin on alle yhden. Simuloitu esimerkki on jaksossa 13.10.4. Ääritilanteessa muuttujien välillä ei ole mitään yhteyttä: Kuvitteelliseen kuvioon piirretyn summeeraavan sovitteen kulmakerroin on nolla, ja poikkeamat odotusarvosta pyrkivät keskimäärin "korjaantumaan" täysin seuraavassa havainnossa. Nykypäivään ja yhteiskuntatieteisiin liittyviä esimerkkejä on helppo keksiä. Regressio odotusarvoa kohti on usein yksinkertaisin selitys.

Esimerkki. Sosiaalitukien määrä. Verrataan sosiaalitukia saavien (tai rikosten, avioliittojen, syntyneiden lasten jne.) lukumäärää suomalaisissa kaupungeissa vuosina 2021 (y-akseli) ja 2020 (x-akseli). Tällöin poikkeuksellisen suuri sosiaalitukea saavien määrä tietyssä kaupungissa tasoittuu lähemmäksi odotusarvoa seuraavana vuonna. Vastaavasti tavanomaista pienemmästä sosiaalitukea nauttivien lukumäärästä vuonna 2020 ilahtuneet kaupunginjohtajat joutuvat tyy-

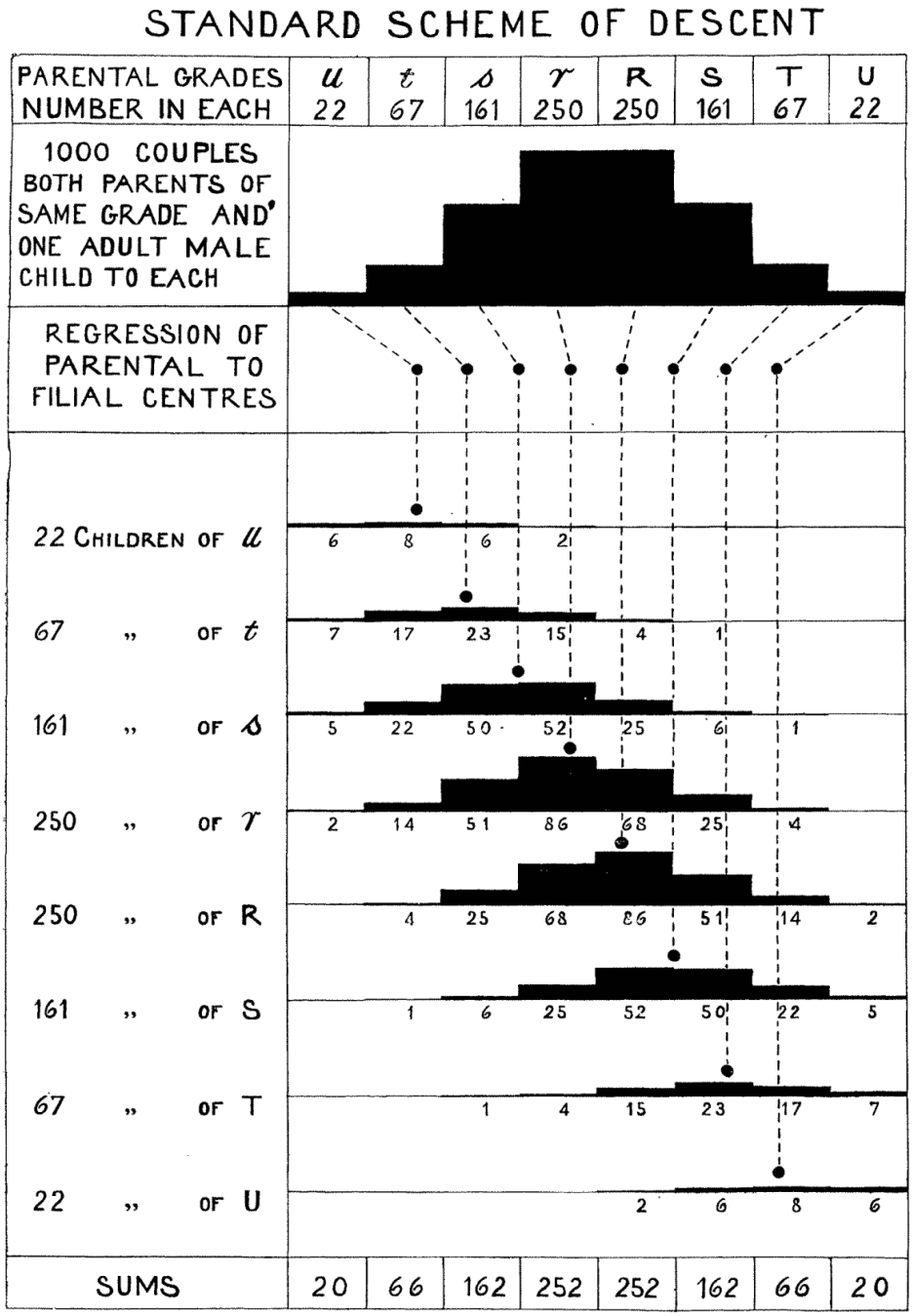
[1]Kuva 13.1 on Pearsonin (1920) artikkelista. Kuva löytyy myös Pearsonin (1930, 4) kirjasta. Regressiosuora on Pearsonin uusiksi laskema ja ilmeisesti hänen apulaisensa (A. Davinin) piirtämä Galtonin muistiinpanojen vuodelta 1875 avulla. (Pearson 1920, 34, 1930, 4.) Kuvan otoksessa "herneen siemen -vanhemmat" ovat keskimäärin selvästi pidempiä kuin "herneen siemen -jälkipolvi". Se lienee otokseen liittyvä vääristymä (Pearson 1930, 3). Kuva 13.2 on Galtonin (1886) artikkelista. Se on painettu uudelleen Pearsonin (1930, 16) kirjassa. Tällaisia kuvia on myös Galtonin (1889, 96 ja 107) kirjassa. Kuva 13.3 on artikkelista Galton (1901a) mutta löytyy myös Galtonin (1901b) artikkelista. Kiitän Oxford University Pressiä, joka on Biometrika Trustin puolesta myöntänyt luvan kuvan 13.1 julkaisemiseen. Kuvat 13.2 ja 13.3 olen poiminut sivustolle `galton.org` kootuista artikkeleista. Sivuston ylläpitäjä Gavan Tredoux ilmoittaa aineiston olevan vapaasti käytettävissä (ko. sivusto sekä henkilökohtainen vahvistus 17.3.2020). Kiitän ylläpitäjä Tredouxia.



Kuva 13.1: Galtonin ensimmäinen regressio 1877.[1]



Kuva 13.2: Galtonin toinen regressio 1886.[1]



Kuva 13.3: Galtonin havainnollistus vanhempien ja lasten pituuden jakaumista 1901.[1]

pillisesti pettymään, kun sosiaalitukea haetaan vuonna 2021 edellistä vuotta enemmän. □

Esimerkki. Voittajan kirous. Edelläkuvatunkaltainen ilmiö on "voittajan kirous" (*winner's curse*): Kun suuresta joukosta esimerkiksi työpaikan tai urheilujoukkueen jäsenyyden hakijoista poimitaan suorituksiltaan paras, ei valittu yllä aivan parhaimpien suoritustensa mukaiseen tulokseen. □

Edellä vertailtiin kahden satunnaismuuttujan yhteyttä, kun ne ovat samoinjakautuneita ja niiden sirontakuvioon piirretyn muuttujien välisen systemaattisen komponentin summeeraavan suoran kulmakerroin on (itseisarvoltaan) alle yksi (tyypillisesti oleellisesti sama satunnaismuuttuja jossain mielessä kahdesti mitattuna). Regressio kohti odotusarvoa ei rajoitu tällaisiin tilanteisiin. Asiaan palataan yhden selittäjän tilannetta koskien jaksossa 13.4.1.

Regressioanalyysissa (jakso 13.3) voidaan sallia useampia muuttujia, jotka voivat olla erilailla jakautuneita tai kiinteitä, eikä systemaattisen vaikutuksen suuruutta tarvitse rajoittaa edelliseen tapaan. Tällöinkin havaitaan regressio odotusarvoa kohti kahden muuttujan välillä mutta puhdistettuna muiden muuttujien vaikutuksesta.

Regressioanalyysissa on aina kyse pohjimmiltaan samasta ilmiöstä kuin edellä eli että osa havaintojen käyttäytymisestä on systemaattista ja osa sattumaa. Regressioanalyysillä pyritään selvittämään systemaattisuus yhden muuttujan ja muiden muuttujien välillä. Sattuman vaikutus tulisi regressoida "pois" muuttujien välistä yhteyttä arvioitaessa. Esimerkiksi Galtonin tutkimusaineistossa lasten ja vanhempien pituuksien suhteella on geneettinen (systemaattinen) selitys, mutta osin lasten pituudet johtuvat (tutkijan näkökulmasta) sattumanvaraisista seikoista kuten lapsen perimistä geeneistä, lapsen ruoan ravinnepitoisuudesta tai lapsen sairastamista taudeista, kellonajasta, jolloin lapsi on mitattu (aamulla lapsi on pidempi) ja niin edelleen.

On vain hieman liioiteltua sanoa, että lähes asiassa kuin asiassa on regressiota. Campbellin ja Kennyn (1999, ix) mukaan regressio odotusarvoa kohti on yhtä väistämätön asia kuin verot tai kuolema.

13.2 Regressiovirhepäätelmä

Regressiovirhepäätelmä (*regression fallacy*) tehdään, kun satunnaisvaihtelussa regressiota odotusarvoa kohti kuvaavan suoran ympärillä kuvitellaan kausaalisuutta kuten että regressio johtaisi jakauman typistymiseen. Yhteiskuntatieteilijät ovat joskus hahmottavinaan kausaalisuutta tilanteista, joissa sitä ei ole.

Vaikka ongelma on tunnettu, edelleen tehdään virheellisiä tulkintoja.

Esimerkki. Yritysten keskiarvoistuminen. Kuuluisa esimerkki on tilastotieteen(!) professori Horace Secrist. Hän julkaisi 1933 massiivisen empiirisen tutkimuksen amerikkalaisten yritysten liikevoittojen kehityksestä 1920–1930. Hän havaitsi, että yritysten, jotka pärjäsivät parhaimmin tai huonoimmin 1920, liikevoitot olivat lähestyneet 1930 kaikkien yritysten liikevoittojen keskiarvoa. Secrist päätteli, että taloudellinen kilpailu pakotti yritykset "keskiarvoistumaan" ajan myötä. Löytönsä korostamiseksi Secrist antoi kirjalleen nimeksi *The Triumph of Mediocrity in Business*. Todellisuudessa yritysten liikevoittojen jakauma ei ollut muuttunut, ja Secristin havainnot selittyvät regressiolla odotusarvoa kohti.[146] □

Esimerkki. Yrityskirjallisuus. Kahneman (2011, 204–208) kritisoi yrityskirjallisuutta, jossa perehdytään menestyneiden yhtiöiden strategioihin, yrityskulttuureihin ja johtamistapoihin. Esimerkkinä hän mainitsee Collinsin ja Porrasin (2000) kirjan. Sen viesti on, että jokaisen toimitusjohtajan, johtajan ja yrittäjän tulisi lukea se, jotta muutkin yritykset osaisivat noudattaa menestyneiden yritysten toimintamalleja ja pärjäisivät. Kahnemanin mukaan Collinsin ja Porrasin ylistämät yritykset eivät pian tutkimuksen julkaisemisen jälkeen enää pärjänneet juurikaan kilpailijoitansa paremmin. Kahneman viittaa muihin vastaaviin tapauksiin, joissa tutkimuksessa hehkutettujen yritysten kukoistus lopahtaa tutkimuksen julkaisemisen jälkeen. Regressio odotusarvoa kohti voi olla luonteva tulkinta tällaisille tapahtumille. Ihaillut yritykset olivat erityisen menestyviä tutkimushetkellä sattumalta. □

Esimerkki. Hävittäjälentäjät. Kahneman (mts. 174) kertoo mainion esimerkin, kuinka ihmiset voivat kuvitella kausaalisuutta siellä, missä on pelkkää sattumaa (lyhennetty käännös):

> Sain yhden elämäni tyydyttävimmistä eureka-kokemuksistani opettaessani Israelin ilmavoimien lentokouluttajille tehokkaan opettamisen psykologiaa. Olin kertonut kouluttajille, kuinka hyvän suorituksen palkitseminen toimii paremmin kuin virheistä rangaitseminen. Yksi vanhemmista kouluttajista arveli, että hyvästä suorituksesta palkitseminen sopii ehkä linnuille muttei hävittäjälentäjäkadeteille: "Olen monesti kehunut kadetteja puhtaasta suorituksesta vaikeassa lentoliikkeessä. Seuraavalla kerralla he järjestään suoriutuvat samasta liikkeestä huonommin. Toisaalta olen monesti huutanut kadetin korvakuulokkeeseen haukkuen häntä huonosta suorituksesta. Ylipäänsä haukkumani kadetit pärjäävät seuraavalla yrityksellä paremmin. Olkaa siis hyvä, älkääkä kertoko meille, että kehuminen toimii ja rangaistus ei, koska asia on juuri päinvastoin."

Vanhemman kouluttajan kokemukset selittyvät sattumalla: Erityisen hyvin pärjänneen kadetin suoritus regressoitui seuraavalla lennolla kohti odotusarvosuo-

ritustaan ja erityisen heikosti suoriutuneen kadetin suoritus samoin. Kouluttaja virheellisesti liitti muutoksiin kuvittelemansa syy-seuraussuhteen kehuistaan ja karjumisistaan.[147] □

13.3 Regressioanalyysi

Regressioanalyysi on käytetyimpiä tilastotieteellisiä menetelmiä. Ei liene olemassa kaupallista tilasto-ohjelmistoa, joka ei sisältäisi regressioanalyysia. Yksi ilmeinen syy on, että sillä voi arvioida muuttujan vaikutuksen suuruutta toiseen muuttujaan tai vaikutuksen olemassaoloa ylipäätään. Regressioanalyysi on tässä mielessä usein hyvin antoisaa ja tuloksellista. Joskus regressioanalyysista innostutaan päättelemään asioita, joita siitä ei välttämättä seuraa. Regressiolla estimoidaan tilastollinen yhteys. Vaikutustulkinta seuraa sovellusalan teoriasta; ei tilastotieteestä. (Jaksot 13.10 ja 13.11.2 sekä luku **??**.) Tässä ja alla kirjoitetaan vaikutuksesta, kun se on luonnollinen tulkinta. Välttämättä ei ole kuitenkaan kyse syy-seuraussuhteesta.

Regressioanalyysia käytettäessä tulisi huolella tutustua dataan, piirtää siitä kuvia, testata mallin oletuksia jne. Tällaiseen syventymiseen ei ole tässä mahdollisuutta. Omiin sovelluksiin tähtäävän lukijan kannattaa täydentää tietojaan esimerkiksi Farawayn (2015), Foxin (2016) tai Weisbergin (2014) oppikirjoista.

R:n

```
malli <- lm(y~x1+x2+x3, data=aineisto)
summary(malli)
```

tapaisilla käskyillä on helppo tehdä luvussa kuvattavat estimoinnit ja testit (kolme selittäjää; muuttujat y, x1, x2 ja x3 sisältävä "aineisto" luettuna valmiiksi R:ään). Käskyyn voi lisätä jaksossa 13.11.3 kuvatut painot (`weights`-määre, aja R:ssä käsky `help(lm)`).

13.4 Yhden selittäjän lineaarinen regressiomalli

Tarkastellaan kahta muuttujaa Y ja x. Edellisen pitää olla välimatka-asteikollinen; jälkimmäinen voi olla myös luokka-asteikollinen. Muuttuja Y määräytyy lineaarisen regressiomallin

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{13.1}$$

mukaisesti x:n arvoista. Muuttujaa Y kutsutaan *selitettäväksi muuttujaksi*, *vastemuuttujaksi* tai *vasteeksi* (*response variable*, *regressand*) ja muuttujaa x *selittäväksi muuttujaksi* tai *selittäjäksi* (*explanatory variable*, *regressor*). Viimeinen

termi ε on *satunnaistermi* (*random error*), jonka odotusarvo on 0 ja varianssi on σ^2. Satunnaistermi poimii selitettävän Y vaihtelun, joka ei selity selittäjän x vaihtelulla. Mallin parametrit ovat kiinteitä lukuja (esim. $\beta_0 = 90.5$ ja $\beta_1 = 0.5$), joiden suuruudet (tyypillisesti) ovat tuntemattomia ja joiden selvittämiseen regressioanalyysillä pyritään. Parametria β_0 kutsutaan usein *vakiotermiksi* (*intercept*) ja parametria β_1 *regressiokertoimeksi* (*regression coefficient*).

Jaksossa oletetaan yksinkertaisuuden vuoksi, että selittäjä x on ei-satunnainen (kiinteä) — ja jaksossa 13.5, että selittäjät x_i ovat ei-satunnaisia. Empiirisessä tutkimuksessa oletus ei usein ole uskottava. Analyysit luvussa pätevät, jos selittäjät ja satunnaistermit ovat toisistaan riippumattomia satunnaismuuttujia ja tehdään sopivia oletuksia ja merkintöjä tuunataan. Regressiomallia voi soveltaa, vaikka selittäjät olisivat satunnaismuuttujia (jakso 13.10).

Jaksossa 13.1 haettu systemaattinen komponentti on yhden selittäjän regressiossa selitettävän (selittäjän x arvosta riippuva) odotusarvo

$$\mathsf{E}(Y) = \mathsf{E}(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x. \tag{13.2}$$

Yllä E on odotusarvo-operaattori (jakso 6.2) ja on käytetty oletusta $\mathsf{E}(\varepsilon) = 0$.

Tavalliseen x, y-koordinaatistoon funktio $y = \beta_0 + \beta_1 x$ piirtyy suorana, jonka kulmakerroin on β_1 ja joka leikkaa y-akselin kohdassa β_0. Periaatteessa β_0 kertoo siten selitettävän odotusarvon, kun selittäjän arvo on 0:

$$\mathsf{E}(Y) = \mathsf{E}(\beta_0 + \beta_1 \times 0 + \varepsilon) = \beta_0.$$

Empiirisessä analyysissa tämä tulkinta ei ole aina järkevä, mihin palataan myöhemmin (jaksot 13.4.1 ja 13.5.2).

Mallin mukaan Y:n suuruus riippuu x:n suuruudesta lineaarisesti parametrin β_1 välityksellä: Jos x muuttuu yksikön verran, niin Y muuttuu β_1:n verran. Esimerkiksi jos Y on lapsen pituus, x on isän pituus ja $\beta_1 = 0.5$, niin mallin mukaan lapsen pituus tapaa olla 0.5 senttimetriä pidempi, jos isä on senttimetrin pidempi. Vakiotermi β_0 asettaa mallin kuvaaman suoran sopivalle korkeudelle. Satunnaistermi ε kuvaa Y:n vaihtelua, joka ei selity x:n vaihtelulla. Esimerkiksi lapsen pituuteen vaikuttaa muitakin tekijöitä kuin isän pituus (jakso 13.1). Ne jäävät mallissa huomioimatta ja puristetaan ε:iin.

Erikoistapaus on $\beta_1 = 0$. Tällöin malli (13.1) typistyy niin, että Y on satunnaisesti jakautunut vakiotermin β_0 ympärillä:

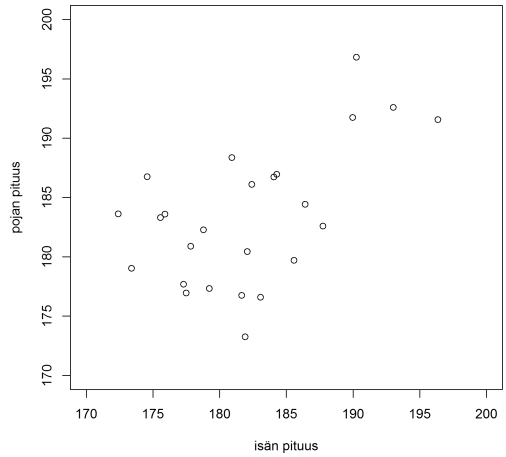
$$Y = \beta_0 + 0 \times x + \varepsilon = \beta_0 + \varepsilon. \tag{13.3}$$

Kiinnostavin asia mallissa (13.1) onkin tyypillisimmin parametrin β_1 suuruus — esimerkiksi poikkeaako se nollasta eli päteekö malli (13.1) vai (13.3).

Regressioanalyysi on keino arvioida parametrien suuruutta ja systemaattista komponenttia, kun mallin kuvaamasta ilmiöstä on havaintoaineisto. Mallin (13.1) kohdalla aineisto koostuisi havaintopareista $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Tässä $n \geq 2$ on havaintojen lukumäärä.

Galtonin herneen siemen -vanhemmat ja niiden jälkipolvi sekä vanhempien ja lasten pituudet -esimerkit (jakso 13.1) havainnollistavat regressiota odotusarvoa kohti mallin (13.1) mukaisesti, kun $0 < \beta_1 < 1$. Kuvitteellinen sosiaalitukien saajat -esimerkki (jakso 13.1) vastaa mallia (13.3): Sosiaalitukien saajien lukumäärä edellisenä vuonna (x) ei auta ennustamaan heidän lukumääräänsä kuluvana vuonna (Y): Kerroin $\beta_1 = 0$, ja lukumäärät pyrkivät palautumaan kohti odotusarvoaan β_0.

Kuvaan 13.4 on piirretty keinotekoinen aineisto ($n = 25$) — vaikkapa tamperelaisten isien (x) ja heidän poikiensa (y) pituuksista aikuisina.[148] Kukin piste vastaa yhtä havaintoparia (x_i, y_i). Mitä pidempi isä, sitä pidempi tapaa poika olla. Mutta kuinka paljon? Voitaisiinko yhteys tiivistää suoraksi, jonka parametreista voitaisiin päätellä vaikutuksen keskimääräinen suuruus?



Kuva 13.4: Isien ja poikien pituudet.

13.4.1 Yhden selittäjän lineaarisen regressiomallin estimointi ja selityskyky

Yhden selittäjän regressiossa aineistoon sovitetaan regressiosuora, joka summeeraa muuttujien välisen riippuvuuden eli systemaattisen osan. Regressiosuoran parametriarvot ovat vastaus edellä esitetyn tapaisiin kysymyksiin. Oletetaan, että käytössä on havaintoparit $(x_1, y_1), \ldots, (x_n, y_n)$.

Sovittaminen voidaan tehdä monella tavalla. Ylivoimaisesti käytetyin tapa on *pienimmän neliösumman (PNS) menetelmä* (*ordinary least squares*). Siinä parametrit β_0 ja β_1 valitaan niin, että y_i-havaintojen poikkeamat sovitettavasta suorasta neliöidään ja neliöiden summa minimoidaan:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Merkintä "min" tarkoittaa, että sen oikealla puolella oleva lauseke minimoidaan min-merkinnän alapuolelle merkittyjen suureiden suhteen. Minimoinnin voi ajatella tapahtuvan ikään kuin kokeilemalla eri lukuarvoja β_0:lle ja β_1:lle ja valitsemalla sellainen β_0, β_1-pari, että lauseke $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ ei voi saada pienempiä arvoja. (Todellisuudessa tilasto-ohjelmisto ratkaisee minimointitehtävän yhdellä laskutoimituksella eikä kokeile eri arvoja.) Poikkeamien suoralta $y_i - \beta_0 - \beta_1 x_i$ kasvaessa (itseisarvoltaan) kasvaa neliösumma $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ nopeasti. PNS-menetelmä pyrkii siten tuottamaan regressiosuoran, joka ei koskaan sijoittuisi kovin kauas yhdestäkään havaintopisteestä. Termien $(y_i - \beta_0 - \beta_1 x_i)$ neliöinnin takia minimoinnin kannalta ei ole väliä, onko y_i suurempi tai pienempi kuin mallin mukainen arvo $\beta_0 - \beta_1 x_i$. Kaikkia poikkeamia kohdellaan tässä mielessä samanarvoisesti.

Neliösumman minimoivia parametriarvoja kutsutaan *PNS-estimaateiksi* ja niitä merkitään $\hat{\beta}_0$:lla ja $\hat{\beta}_1$:lla. Suureita

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

kutsutaan *sovitteiksi* (*fitted value*) ja suureita

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

jäännöksiksi (*residual*, $i = 1, \ldots, n$). *Regressiosuora* $\hat{\beta}_0 + \hat{\beta}_1 x_i$ saadaan piirtämällä sirontakuvioon suora sovitteiden $(\hat{y}_i)$ kautta. Jäännökset $(\hat{\varepsilon}_i)$ ovat satunnaistermien (ε_i) estimaatteja.

Vakiotermin PNS-estimaatin voidaan osoittaa olevan

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

jossa $\bar{y} = \sum_{i=1}^{n} y_i/n$ ja $\bar{x} = \sum_{i=1}^{n} x_i/n$. Näin ollen sovite $\hat{y}_i$ on ilmaistavissa muuttujien keskiarvojen avulla ja pätee, että

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) \quad \Leftrightarrow \quad \hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}). \tag{13.4}$$

Sovite poikkeaa $\bar{y}$-keskiarvosta $\hat{\beta}_1$ kertaa x_i:n poikkeaman omasta keskiarvostaan verran. Sijoittamalla $x_i = \bar{x}$ havaitaan, että regressiosuora kulkee pisteen $(\bar{x}, \bar{y})$ kautta.

Regressiokertoimen PNS-estimaatti punoutuu y:n ja x:n otoskorrelaatiokertoimeen $\hat{\rho}$ ja niiden otoskeskihajontoihin s_y ja s_x:

$$\hat{\beta}_1 = \hat{\rho}\frac{s_y}{s_x} \tag{13.5}$$

(harjoitustehtävä). Yhtälöistä (13.4) ja (13.5) seuraa, että

$$\frac{\hat{y}_i - \bar{y}}{s_y} = \hat{\rho}\frac{x_i - \bar{x}}{s_x}.$$

Koska $-1 < \hat{\rho} < 1$, niin sovite poikkeaa selitettävän keskiarvosta vähemmän kuin selittäjä poikkeaa keskiarvostaan, kun poikkeamat on standardoitu vastaavilla otoskeskihajonnoilla. Tässä mielessä regressiomallissa on aina regressiota kohti odotusarvoa.

Tärkeä käsite on *jäännösneliösumma* (*residual sum of squares*)

$$\mathsf{JNS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Nimityksensä mukaisesti se on summa jäännösten neliöistä. Se on sitä suurempi, mitä enemmän y_i-havainnot poikkeavat sovitteista $\hat{y}_i$. Jäännösneliösummasta saadaan helposti estimaatti satunnaistermin varianssille *jäännösvarianssi* (*residual variance*):

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Estimaatti s^2 mittaa jäännöksen neliön keskimääräistä suuruutta eli vaihtelevuutta aineistossa.[149] Yleensä toivotaan, että s^2 olisi pieni, koska silloin malli

selittää hyvin y_i-havaintojen vaihtelun. Monesti raportoidaan satunnaistermin estimoitu keskihajonta $\sqrt{s^2} = s$, koska se on samassa mittayksikössä kuin selitettävä muuttuja ja siksi helpompi hahmottaa.

Määritellään vastaavasti *kokonaisneliösumma* (*total sum of squares*)

$$\mathsf{KNS} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{13.6}$$

Se kuvaa, kuinka suurta on y_i-havaintojen vaihtelu keskiarvonsa ympärillä.

Neliösummista saadaan mallin selityskyvylle mittari *selitysaste* (*coefficient of determination*)

$$\mathsf{R}^2 = \frac{\mathsf{KNS} - \mathsf{JNS}}{\mathsf{KNS}} = 1 - \frac{\mathsf{JNS}}{\mathsf{KNS}}. \tag{13.7}$$

Selitysaste saa lähellä yhtä olevia arvoja, mikäli jäännösneliösumma on pieni suhteessa selitettävän kokonaisneliösummaan ($\mathsf{JNS}/\mathsf{KNS} \approx 0$). Tällöin y selittyy hyvin x:llä. Mikäli x:llä ei ole selityskykyä, jäännösneliösumma ei eroa paljoa kokonaisneliösummasta ($\mathsf{JNS}/\mathsf{KNS} \approx 1$). Tällöin selitysaste on lähellä nollaa.

Selitysaste on hyvin intuitiivinen mittari mallin hyvyydelle. Nyt esillä olevassa yhden selittäjän regression tilanteessa se on otoskorrelaatiokertoimen ($\hat{\rho}$) neliö:

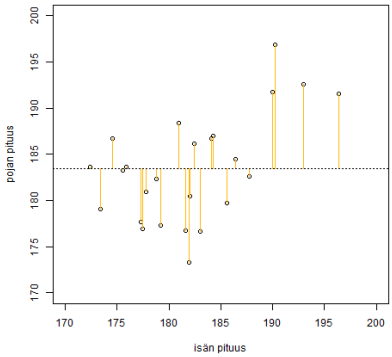
$$\mathsf{R}^2 = \hat{\rho}^2. \tag{13.8}$$

Yhtälö pätee, kunhan mallissa on vakiotermi.

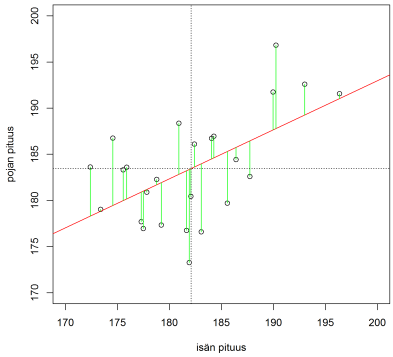
Kuvat 13.5 ja 13.6 havainnollistavat käsitteitä isä-poika-aineiston avulla. Kuvaan 13.5 on piirretty poikien pituuksien poikkeamat poikien pituuksien keskiarvosta ($(y_i - \bar{y})$:t). Poikkeamien neliöiden summa on kokonaisneliösumma (13.6). Kuvassa 13.6 regressiosuora määrittää sovitteen kunkin x_i-havainnon kohdalla. Jäännökset ovat (x_i, y_i)-havainnoista regressiosuoraan pystysuorasti meneviä viivoja ($(y_i - \hat{y}_i)$:t). Toinen suora tuottaisi toiset jäännökset. Kuvion jäännösten neliöiden summa on pienin mahdollinen. Mallin (13.1) PNS-estimointi tuottaa tästä aineistosta tulokset

$$\begin{aligned} y &= 86.79 + 0.531x + \hat{\varepsilon} \\ &= 183.4 + 0.531(x - 182.1) + \hat{\varepsilon}, \\ s &= 4.96, \quad \mathsf{R}^2 = 0.313, \quad n = 25. \end{aligned}$$

Malli ennustaa pojalle lisää pituutta 0.531 eli noin 0.5 senttimetriä isän pituuden kasvaessa senttimetrillä ja selittää noin 31 % poikien pituuden vaihtelusta aineistossa. Toinen rivi yllä esittää mallin kaavan (13.4) muodossa isien ja poikien pituuksien keskiarvojen ($\bar{x} = 182.1$ ja $\bar{y} = 183.4$) avulla ($86.79 \approx$



Kuva 13.5: Poikien pituuksien poikkeamat poikien keskipituudesta.



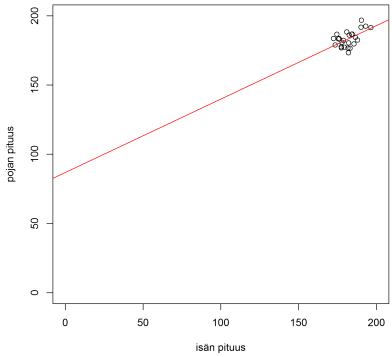
Kuva 13.6: Isien ja poikien pituuden regressiosuora ja jäännökset.

$183.4 - 0.531 \times 182.1$). Kaavasta (13.8) ja regressiokertoimen positiivisuudesta seuraa, että otoskorrelaatio on selitysasteen neliöjuuri: $\hat{\rho} = \sqrt{\mathsf{R}^2}$. Pituuksien otoskorrelaatio on siten $\sqrt{0.313} \approx 0.559$. Satunnaistermin estimoitu keskihajonta on 4.96.

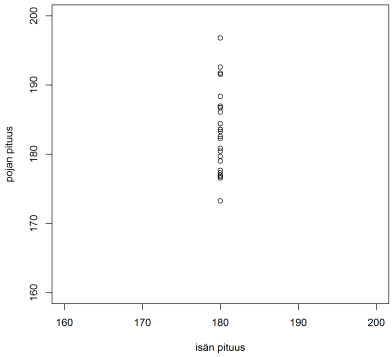
Koska aineisto oli keinotekoinen, estimointituloksia voidaan verrata aineiston tuottaneeseen todelliseen malliin. Suureiden todelliset arvot ovat $\beta_0 = 90.5$, $\beta_1 = 0.5$, $\sigma \approx 5.25$ (seuraa tehdyistä oletuksista tavalla, jota ei tässä selitetä), korrelaatio populaatiossa $\rho = 0.5$ ja selitysaste populaatiossa $\mathsf{R}^2 = \rho^2 = (0.5)^2 = 0.25$.[150] Kaikki suureet tulivat estimoiduksi varsin hyvin.

Kuten usein on, estimoidulla vakiotermillä ei ole mielekästä tulkintaa. Estimoidun mallin ja vakiotermin mukaan pojan pituus olisi noin 87 senttimetriä, jos isän pituus olisi 0 senttimetriä (kuvio 13.7), mikä on järjetön ajatus. Malli ei välttämättä antaisi luotettavaa ennustetta edes periaatteessa mahdollisen mutta poikkeuksellisen lyhyen isän (esim. $x = 155$) pojan pituudelle. Regressiomalleja ei ylipäänsä kannata soveltaa aineiston vaihteluvälin ulkopuolella (ekstrapoloida).

Edellä implisiittisesti oletettiin, että kaikki x_i-havainnot eivät ole yhtäsuuria. Jos ne olisivat, β_1-parametria ei voisi estimoida, mitä kuva 13.8 havainnollistaa. Seuraavassa jaksossa tarvitaan muitakin oletuksia.



Kuva 13.7: Regressiosuoralla ekstrapolointi.



Kuva 13.8: Kaikki isät ovat samanpituisia.

13.4.2 Yhden selittäjän lineaarisen regressiomallin testaus

Ei ole harvinaista, että tutkijan päämielenkiinto on estimoinnissa. Vaikka se ei olisi, pääsääntöisesti regressiomallin tarkastelussa ei tulisi rajoittua estimointituloksiin. Mallia tulisi aina testata. Filosofia on sama kuin muutenkin tilastotieteessä: Pelkkä estimaatin tai tilastollisen tunnusluvun subjektiivinen arviointi ei ole riittävää; tulee myös laskea väliestimaatti tai testata, poikkeaako tunnusluku nollasta tai muusta oleelliseksi katsotusta arvosta tilastollisesti merkitsevästi. Hedelmällisen tilastotieteen soveltamisen tunnusmerkkejä on, että on arvioitu sekä tunnuslukujen merkityksellisyyttä sovellusalan kannalta että niiden tilastollista merkitsevyyttä luottamusvälien tai testien avulla. Alla keskitytään testaukseen, koska niin tehdään valtaosassa empiiristä kirjallisuutta.

Regressioanalyysillä voidaan testata parametreihin liittyviä nollahypoteeseja (H_0). Sellaisia ovat esimerkiksi H_0: $\beta_1 = 0$ tai H_0: $\beta_1 = 1$. Vakiotermin suuruutta testataan harvoin, koska sillä ei ole usein selkeää sovellukseen liittyvää merkityksellistä tulkintaa. Esimerkki on isä-poika-malli edellä. Vakiotermin luontevasta tulkinnasta on esimerkki jakson 13.5.2 lopussa.

Hypoteesien testaukseen tarvitaan lisäoletuksia:

- Satunnaistermi noudattaa normaalijakaumaa odotusarvolla 0 ja variansilla σ^2: $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$, $i = 1, \ldots, n$.

- Satunnaistermit ε_i eivät korreloi keskenään eli ne ovat riippumattomia (normaalijakauman tilanteessa korreloimattomuudesta seuraa riippumattomuus).

Oleellista on ymmärtää, että $\hat{\beta}_1$ on satunnaismuuttuja. Se saa tietyn arvon tutkittavana olevassa aineistossa. Jos tutkittavana olisi toinen — esimerkiksi espoolainen 25 havainnon aineisto isien ja poikien pituuksista — saataisiin toisensuuruinen $\hat{\beta}_1$. Samoin estimaatti muuttuisi, jos tutkittaisiin 25 helsinkiläisen, 25 jyväskyläläisen jne. aineistoa isien ja poikien pituuksista. Koska $\hat{\beta}_1$ on satunnaismuuttuja, on sillä (ilmeisesti) myös keskihajonta, jota kutsutaan tässä yhteydessä keskivirheeksi (jakso 9.1).

Edellä lueteltujen oletuksien pätiessä PNS-estimaattorien jakaumat tunnetaan. Tavalla, jota tässä ei selitetä, voidaan laskea estimoitu keskivirhe $\hat{\beta}_1$:lle ($s_{\hat{\beta}_1}$) ja muodostaa t-testisuure eli t-arvo

$$t_{\beta_1=\beta_{10}} = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}} \sim \mathsf{t}(n-2).$$

Nollahypoteesin H_0: $\beta_1 = \beta_{10}$ pätiessä se noudattaa t-jakaumaa vapausasteilla $n-2$. Huomionarvoista on, että jakauma riippuu havaintojen lukumäärästä mutta tunnetaan kaikilla havaintomäärillä. Testisuure on hyvin intuitiivinen. Estimaatin $\hat{\beta}_1$ poikkeama nollahypoteesin mukaisesta arvosta β_{10} suhteutetaan estimaatin estimoituun keskivirheeseen. Suurikaan poikkeama ei ole tilastollisesti merkitsevä, jos $\hat{\beta}_1$:n keskivirhe on suuri. Toisaalta pienikin poikkeama on tilastollisesti merkitsevä, jos $\hat{\beta}_1$:n keskivirhe on hyvin pieni. Keskivirhe pienenee havaintojen lukumäärän kasvaessa. (Muutkin tekijät vaikuttavat keskivirheen suuruuteen.)

Tyypillisimmin testataan nollahypoteesia $\beta_1 = 0$. Tällöin testisuure on yksinkertaisesti β_1:n estimaatti jaettuna estimoidulla keskivirheellään:

$$t_{\beta_1=0} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim \mathsf{t}(n-2). \tag{13.9}$$

Monet tilasto-ohjelmistot tulostavat tämän testisuureen regressoitaessa selitettävää yhdellä selittävällä muuttujalla. Toiset ohjelmistot raportoivat PNS-estimaatin ja sen keskivirheen, jolloin käyttäjän tehtävä on muodostaa osamäärä $\hat{\beta}_1/s_{\hat{\beta}_1}$. Tieteellisissä artikkeleissa käytäntö vaihtelee: Joissain raportoidaan estimaatti ja t-arvo ja toisissa estimaatti ja sen estimoitu keskivirhe. Jälkimmäisessä tilanteessa lukijan tulee osata itse muodostaa t-arvo, jos haluaa tietää sen suuruuden.

Testaaminen etenee tämän jälkeen tavanomaiseen tapaan eli valitaan sopivaksi katsottu merkitsevyystaso, ja katsotaan, onko testisuureen itseisarvo suurempi kuin merkitsevyystasoon liittyvä kriittinen arvo (kaksisuuntainen testaus). Esimerkiksi 5 %:n merkitsevyystasoa käytettäessä kriittiset arvot olisivat isä-poika-esimerkissä t-jakauman $25-2=23$:lla vapausasteella 0.025. tai 0.975. kvantiilit.

Isä-poika-esimerkissä $\hat{\beta}_1$:n keskivirhe on 0.164, joten t-arvo on $0.531/0.164 \approx 3.238$. Esimerkin laskussa käytetty R-ohjelmisto raportoi sekä estimoidun keskivirheen että t-arvon, joka täsmää juuri lasketun kanssa. Ohjelmiston mukaan p-arvo on noin 0.004, joten nollahypoteesi $\beta_1 = 0$ hylätään merkitsevyystasolla 0.05 ja paljon pienemmilläkin merkitsevyystasoilla. Samaan tulokseen päädytään vertaamalla t-arvoa 3.238 t-jakauman 23 vapausasteella 0.975. kvantiiliin 2.069 (`qt(0.975,23)`).

Testin mukaan isien ja lasten pituus ovat yhteydessä ($\beta_1 \neq 0$). Tulos on odotettu. Mikäli tutkittava ilmiö olisi tuntemattomampi, keskeinen osa regressioanalyysia olisi testata, poikkeaako parametri β_1 nollasta. Mikäli nollahypoteesia ei hylättäisi (t-arvo olisi itseisarvoltaan pienempi kuin kriittiset arvot), pääteltäisiin, että muuttujien välillä ei ole yhteyttä tai että aineisto ei ainakaan ole ristiriidassa oletuksen yhteyden puuttumisesta kanssa. Mallin selitysaste olisi tällöin lähellä nollaa (mieti miksi!), ja kaavan (13.8) perusteella muuttujien välinen otoskorrelaatio olisi samoin lähellä nollaa. Regressiokerrointa koskevan testin tulos sanoitetaan monesti niin, että siihen liittyvä selittäjä — edellä isän pituus — on tilastollisesti merkitsevä tai ei-merkitsevä.

Esimerkki. Itsemurhat.[151] Daly ym. (2011) estimoivat PNS-menetelmällä yhtälön

$$\begin{array}{rcccccc} y & = & 24.912 & + & 8.255x & + & \hat{\varepsilon}, \\ & & (2.311) & & (3.992) & & \end{array}$$

$$s = 8.266,\ \mathsf{R}^2 = 0.248,\ n = 15.$$

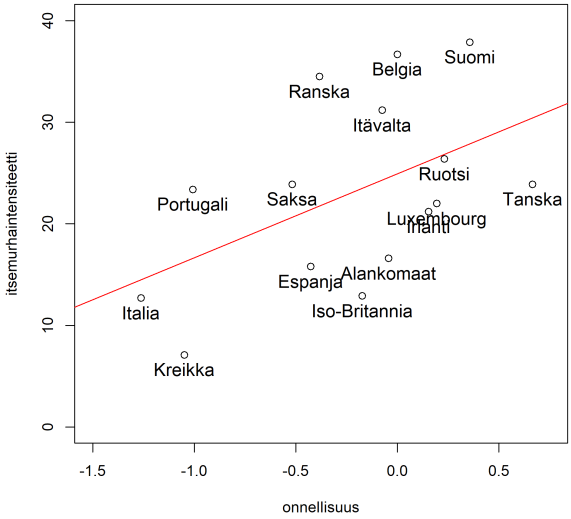
Yllä y on itsemurhien lukumäärä 100 000 kansalaista kohti, x on kansakunnan onnellisuutta mittaava indeksi, $\hat{\varepsilon}$ on jäännös, luvut suluissa ovat estimoituja keskivirheitä ja n on havaintojen lukumäärä. Satunnaistermien ε oletetaan noudattavan normaalijakaumaa $\mathsf{N}(0, \sigma^2)$ ja olevan keskenään korreloimattomia. Kukin havaintopari (x_i, y_i) liittyy eurooppalaiseen valtioon ($i = 1, \ldots, 15$). Havainnot ja niihin sovitettu regressiosuora ovat kuvassa 13.9. Mallin mukaan

- itsemurhaintensiteetti (itsemurhien lukumäärä 100 000 kansalaista kohti) on 24.912 (estimoitu vakiotermi), kun onnellisuusindeksi saa arvon 0.

- itsemurhaintensiteetti kasvaa onnellisuusindeksin kasvaessa. Kun jälkimmäinen suurenee yksiköllä, edellinen kasvaa 8.255:llä (estimoitu kerroin onnellisuusindeksille).
- 24.8 % itsemurhaintensiteetin vaihtelusta selittyy onnellisuusindeksin vaihtelulla (selitysasteen R^2 suuruus).

Onnellisuusindeksi saa toiseksi ja itsemurhaintensiteetti suurimman arvon Suomen kohdalla. Suomessa tehdään itsemurhia vielä enemmän kuin malli ennustaa — eniten koko aineistossa.

Korrelaatiokertoimen 0.4975:n neliö on mallin selitysaste 0.248, koska mallissa on vain yksi selittäjä (kaava (13.8)).



Kuva 13.9: Itsemurhien ja onnellisuuden yhteys 15 eurooppalaisessa valtiossa.

Koska satunnaistermit ε ovat normaalijakautuneita ja keskenään korreloimattomia, estimoidut kertoimet jaettuna estimoiduilla keskivirheillään ovat $\mathtt{t}$-jakautuneita. Koska mallissa on vain yksi selittäjä ja havaintoja on 15, on jakauma $\mathtt{t}(15-2)$ eli $\mathtt{t}(13)$ (kaava (13.9)). Testisuure on $8.255/3.992 \approx 2.068$. Ja-

kauman t(13) 0.975. kvantiili on 2.160 (`qt(0.975,13)`). Koska $|2.068| < |2.160|$, niin nollahypoteesi ei tule aivan hylätyksi 5 %:n merkitsevyystasolla kaksisuuntaisessa testauksessa. Onnellisuusindeksi ei ole tilastollisesti merkitsevä selittäjä eikä aineiston perusteella ole syytä luopua oletuksesta, että itsemurhaintensiteetti ja onnellisuusindeksi eivät korreloi.

Kuvion perusteella saattaisi veikata, että muuttujien välillä olisi todellinen yhteys. Selitys tilastolliselle ei-merkitsevyydelle saattaa olla aineiston pieni koko: Tilastollisesti merkitsevä positiivinen suhde itsemurhaintensiteetin ja osavaltioiden onnellisuusindeksien välillä pätee Yhdysvalloissa (mt.), ja osavaltioita on enemmän kuin eurooppalaisia valtioita regressiossa edellä. Mahdollisesti osavaltiot ovat myös homogeenisempia kuin eurooppalaiset valtiot, jolloin tutkittu suhde tulee selvemmin esiin osavaltioaineistossa (satunnaistermi sisältää vähemmän vaihtelevia tekijöitä). □

13.5 Monen selittäjän lineaarinen regressiomalli

Selitettävä muuttuja Y määräytyy nyt monen selittävän muuttujan x_i ($i = 1, \ldots, k$) lineaarisesta regressiomallista

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon. \tag{13.10}$$

Mallin tulkinta on samantapainen kuin mallin (13.1). Selitettävä Y on jatkuva-arvoinen, selittäjät x_i voivat olla myös luokka-asteikollisia ja ε on satunnaistermi odotusarvolla 0 ja varianssilla σ^2. Siihen tiivistyy Y:n vaihtelu, joka ei selity x_i:den vaihtelulla. Parametrit $\beta_0, \ldots, \beta_k$ ovat kiinteitä yleensä tuntemattomia lukuja, joiden suuruudet pyritään selvittämään regressioanalyysilla (eritoten β_1:stä β_k:hon). Parametria β_0 kutsutaan vakiotermiksi ja parametreja $\beta_1, \ldots, \beta_k$ (regressio)kertoimiksi. Yksinkertaisuuden vuoksi selittäjät x_i oletetaan kiinteiksi (niissä ei ole satunnaisuutta).

Mallin (13.10) systemaattinen komponentti on selitettävän (selittäjien x_i arvoista riippuva) odotusarvo

$$\mathsf{E}(Y) = \mathsf{E}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{13.11}$$

Vakiotermi kuvaa nyt selitettävän odotusarvoa, kun kaikki selittäjät saavat arvon 0:

$$\mathsf{E}(Y) = \mathsf{E}(\beta_0 + \beta_1 \times 0 + \cdots + \beta_k \times 0 + \varepsilon) = \beta_0.$$

Kuten yhden selittäjän regressiossa (jakso 13.4.1), tämä tulkinta ei ole aina järkevä.

Kerroin β_i kuvaa x_i:n yksikön suuruisen muutoksen vaikutuksen Y:hyn, kun muut selittäjät eivät muutu. Monesti mielenkiintoisin kysymys on, eroavatko β_i-kertoimet nollasta eli selittääkö x_i:den vaihtelu Y:n vaihtelua.

13.5.1 Monen selittäjän lineaarisen regressiomallin estimointi ja selityskyky

Monen selittäjän regressiomallin (13.10) systemaattisen komponentin ja parametrien selvittäminen edellyttää n:stä havaintovektorista $[x_{11} \ldots x_{1k}\ y_1], \ldots,$ $[x_{n1} \ldots x_{nk}\ y_n]$ koostuvaa aineistoa ($n \geq k$). Merkintä $[x_{i1} \ldots x_{ik}\ y_i]$ tarkoittaa, että selittävien muuttujien ja selitettävän muuttujan lukuarvot i. havainnon kohdalla on järjestetty luettelon tapaan jonoon (esim. [180.2 172.6 178.7], jos $k = 2$). Muuttujien ensimmäinen indeksi on havainnon numero ($i = 1, \ldots, n$) ja jälkimmäinen kertoo, mistä selittäjästä havaintoarvo x_{ij} on ($j = 1, \ldots, k$). Yhdenkään selittäjän x_i arvot eivät saa riippua täydellisesti lineaarisesti muiden selittäjien x_j, $j \neq i$, arvoista.[152]

Monen selittäjän regressiossa aineistoon sovitetaan selitettävän ja selittäjien välisen riippuvuuden summeeraava lineaarinen funktio eli mallin (13.10) systemaattinen osa (13.11).[153] Sovittaminen tehdään yleisimmin PNS-menetelmällä jaksossa 13.4.1 esitettyyn tapaan. Parametrien $\beta_0, \ldots, \beta_k$ lukuarvot valitaan minimoimaan y_i-havaintojen poikkeamien systemaattisesta komponentista neliöiden summa:

$$\begin{aligned} & \min_{\beta_0,\ldots,\beta_k} \textstyle\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^2 \\ = \quad & \min_{\beta_0,\ldots,\beta_k} \textstyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2. \end{aligned}$$

(Hakasulut eivät tässä liity vektorimerkintään edellä. Hakasulut voisi tässä korvata kaarisuluilla.) Neliösumman minimoivat parametriarvot ovat PNS-estimaatteja $\hat{\beta}_0, \ldots, \hat{\beta}_k$. Jaksossa 13.4.1 selitetyt käsitteet yleistyvät muutenkin suoraviivaisesti k:n selittäjän tilanteeseen. Sovitteet ($\hat{y}_i$), jäännökset ($\hat{\varepsilon}_i$), jäännösneliösumma ja satunnaistermin varianssin estimaatti ovat nyt

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik},$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik},$$

$$\mathsf{JNS} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2$$

ja

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2. \tag{13.12}$$

Kokonaisneliösumman ja selitysasteen kaavat (13.6) ja (13.7) eivät muutu. Tärkeä ero on, että selitysasteen ja otoskorrelaation neliön sitova kaava (13.8) pätee nyt, kun otoskorrelaatiokerroin $\hat{\rho}$ on laskettu selitettävän muuttujan y_i ja sen sovitteen $\hat{y}_i$ välille. (Tämä tulkinta on mahdollinen myös yhden selittäjän regression mallin kohdalla.) Korrelaatiokerrointa kutsutaan tässä yhteydessä *yhteiskorrelaatiokertoimeksi* (*multiple correlation coefficient*).

13.5.2 Monen selittäjän lineaarisen regressiomallin testaus

Tärkein ja useimmin testattu monen selittäjän regressiomallin (13.10) β_i-kertoimia koskeva nollahypoteesi on, että ne ovat kaikki nollia (H_0: $\beta_1 = \cdots = \beta_k = 0$). Nollahypoteesin mukaan selittäjät x_i eivät kykene selittämään selitettävän y:n vaihtelua. Tämän hypoteesin päteminen tai pätemättömyys on tutkijalle usein keskeisimpiä kysymyksiä. Testaamista varten satunnaistermit oletetaan normaalijakautuneiksi ja riippumattomiksi jakson 13.4.2 tapaan.

Nollahypoteesia H_0: $\beta_1 = \cdots = \beta_k = 0$ testaava F-testisuure on hyvin yksinkertainen:

$$F = \frac{\mathsf{R}^2/k}{(1-\mathsf{R}^2)/(n-k-1)} \sim \mathsf{F}(k, n-k-1). \tag{13.13}$$

Se noudattaa nollahypoteesin pätiessä F-jakaumaa k:lla ja $n-k-1$:llä vapausasteella. Jälleen (vrt. jakso 13.4.2) jakauma riippuu havaintojen lukumäärästä tunnetaan kaikilla havaintomäärillä.

Mikäli selittäviä muuttujia on vain yksi ($k = 1$), niin F-testisuure on sama kuin yhden selittäjän regressiomallin yhteydessä esitetyn t-testisuureen (13.9) neliö tai korrelaatiokertoimen t-testisuureen neliö:

$$F = \frac{\mathsf{R}^2/1}{(1-\mathsf{R}^2)/(n-1-1)} = (n-2)\frac{\mathsf{R}^2}{1-\mathsf{R}^2} = (n-2)\frac{\hat{\rho}^2}{1-\hat{\rho}^2} = t^2.$$

Kolmas yhtäsuuruus seuraa yhtälöstä (13.8) ja neljäs yhtälöstä (12.10). Toiseksi viimeinen muoto tekee selväksi, että testisuure regressiokertoimen β_1 nolluudelle nivoutuu x- ja y-muuttujien otoskorrelaatioon.[154] Kaksisuuntaisessa testauk-

sessa F- ja t-testit johtavat täsmälleen samoihin johtopäätöksiin, koska $\mathsf{F}(1,d)$-jakauman kvantiilit ovat samat kuin $\mathsf{t}(d)$-jakauman kvantiilit neliöitynä (jakso 7.2.4).

Testisuureen suuret arvot ovat testin kannalta hälyttäviä. Useimmat tilasto-ohjelmistot laskevat F-testisuureen automaattisesti regression yhteydessä.

F-testisuureen intuitio on selkeä. Mikäli selittäjät x_i kykenevät selittämään suuren osan selitettävän y vaihtelusta (KNS), niin jäljelle jäävä selittämätön vaihtelu (JNS) muodostuu pieneksi ja selitysaste R^2 suureksi (kaava (13.7)). Kaavasta (13.13) nähdään, että mitä suurempi R^2 on, sitä suurempi on F-testisuure. F-testi siis hälyttää, kun selittäjillä on aineistossa hyvä selityskyky. Myös havaintojen lukumäärän kasvattaminen pyrkii kasvattamaan F-testisuuretta ja todennäköisyyttä hylätä nollahypoteesi, kun se ei päde. Mikäli nollahypoteesi pätee, R^2 tapaa jäädä pieneksi ja F-testisuure samoin.

Yleisiä mallin (13.10) parametreja koskevia nollahypoteeseja ovat, että i:nnen selittäjän kerroin on nolla (H_0: $\beta_i = 0$) tai että se on tietyn suuruinen (H_0: $\beta_i = \beta_{i0}$). Edellisessä tilanteessa i:nnettä selittäjää ei tarvittaisi regressiossa (13.10). Näitä nollahypoteeseja voidaan testata jakson 13.4.2 tapaisilla t-testisuureilla:

$$t_{\beta_i=\beta_{i0}} = \frac{\hat{\beta}_i - \beta_{i0}}{s_{\hat{\beta}_i}} \sim \mathsf{t}(n-k-1)$$

ja

$$t_{\beta_i=0} = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim \mathsf{t}(n-k-1). \tag{13.14}$$

Vastaavan nollahypoteesin pätiessä ne noudattavat t-jakaumaa vapausasteilla $n-k-1$. Taas jakauma seuraa havaintojen lukumäärää ja tunnetaan. Tilasto-ohjelmisto raportoi yleensä automaattisesti jälkimmäisen t-arvon kaikkien selittäjien estimoiduille kertoimille tai niiden estimoidut keskivirheet $s_{\hat{\beta}_i}$, $i = 1, \ldots, k$. Testaus tapahtuu jaksossa 13.4.2 selitetyllä tavalla. Siellä kuvattiin myös t-testisuureiden intuitio.

Monesti on kiinnostavaa testata, olisivatko mallin (13.10) d ($0 < d \leq k$) oikeanpuoleisinta selittäjää tarpeettomia eli päteekö $\beta_{k_0+1} = \cdots = \beta_k = 0$, jossa $k_0 = k - d > 0$. (Oletus tarpeettomien selittäjien sijoittumisesta mallin oikeanpuoleisimmiksi tehdään merkintöjen yksinkertaistamiseksi.) Näin rajoitettu malli olisi

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k_0} x_{k_0} + \varepsilon. \tag{13.15}$$

Se saadaan mallista (13.10) erikoistapauksena asettamalla d kappaletta β_i-kertoimia nollaksi ($k_0 + 1$). selittäjästä lähtien.

Nollahypoteesia H_0: $\beta_{k_0+1} = \cdots = \beta_k = 0$ voidaan testata testisuureella

$$\frac{(\mathsf{R}^2 - \mathsf{R}_0^2)/d}{(1-\mathsf{R}^2)/(n-k-1)} \sim \mathsf{F}(d, n-k-1).$$

Testisuure vaatii sekä regression (13.10) että regression (13.15) laskemisen. Jälkimmäisen regression selitysastetta on merkitty yllä R_0^2:lla. Nollahypoteesin pätiessä testisuure noudattaa F-jakaumaa d:llä ja $n-k-1$:llä vapausasteella. Testisuureen suuret arvot ovat hälyttäviä.

Tämänkin testisuureen toimintaperiaate on hyvin ymmärrettävä. Mikäli kertoimet $\beta_{k_0+1}, \ldots, \beta_k$ poikkeavat tai osa niistä poikkeaa nollasta, mallien selitysasteiden tulisi erota selvästi. Tällöin erotus $\mathsf{R}^2 - \mathsf{R}_0^2$ testisuureen osoittajassa muodostuu suureksi ja testisuure samoin. Mikäli d. viimeisellä selittäjällä ei ole selitysvoimaa (nollahypoteesi pätee), erotus ja testisuure jäävät pieniksi.

Muunkinlaisia rajoituksia (esim. $\beta_1 = \beta_2$ tai $\beta_1 + \cdots + \beta_k = 1$) mallin (13.10) parametreille voidaan testata. Asia jätetään maininnan varaan.

Vaikka testit kohdistuvat selittäjien kertoimiin, monesti puhutaan selittäjien tilastollisesta merkitsevyydestä. Niin myös alla.

Esimerkki. Siivoojien tuntipalkat I.[155] Keinänen ja Pakarinen (2009) tutkivat siivoojien tuntipalkkoja ja mahdollista palkkasyrjintää suomalaisessa siivousyrityksessä vuonna 2007. He estimoivat vaihtoehtoisia malleja, jotka ovat kaikki palkkasyrjintää koskevalta tulokseltaan yhtäpitäviä. Yksi heidän PNS-menetelmällä estimoimistaan malleista on

$$\begin{array}{rcccccccccc} y & = & 8.430 & + & 0.114x_1 & - & 0.001x_2 & + & 0.169x_3 & + & 0.339x_4 & + & \hat{\varepsilon}. \\ & & (0.000) & & (0.840) & & (0.983) & & (0.747) & & (0.000) & & \end{array} \tag{13.16}$$

$$\mathsf{R}^2 = 0.256,\ F = 11.269,\ n = 137.$$

Yhtälössä y on tuntipalkka, x_1 on osoitinmuuttuja, joka saa arvon 1, kun siivooja on mies ja 0 muuten, x_2 on siivoojan ikä, x_3 on osoitinmuuttuja työsuhteen laadulle, joka saa arvon 1, kun työsuhde on toistaiseksi voimassa oleva ja 0 muutoin ja x_4 on työsuhteen kesto vuosina. Muut merkinnät (F-testisuureella täydennettynä) ja oletukset ovat kuten edellisessä esimerkissä. Kukin havaintovektori $[x_{i1} \ \ldots \ x_{i4} \ y_i]$ liittyy yhteen siivojaan ($i = 1, \ldots, 137$).

Satunnaistermin normaalisuusoletuksen perusteella t-ja F-testisuureet noudattavat t- ja F-jakaumia, kun testatut regressiokertoimet ovat nollia. Muuttujien selityskykyä yhdessä tutkaillaan testisuurella $F = 11.269$. Nollahypoteesin pätiessä se on $\mathsf{F}(4, 137-4-1)$- eli $\mathsf{F}(4, 132)$-jakautunut (kaava (13.13)). Tämän F-jakauman 0.95. kvantiili on 2.440 (`qf(0.95,4,132)`).[156] Tehdään testi

merkitsevyystasolla 0.05. Koska $11.269 > 2.440$, niin nollahypoteesi hylätään. Mallin selittäjillä on yhdessä selityskykyä.

Sukupuoliosoittimen t-arvo on $0.114/0.840 \approx 0.136$. Nollahypoteesin (regressiokerroin on 0) pätiessä se noudattaa $\mathtt{t}(137-4-1)$ eli $\mathtt{t}(132)$-jakaumaa (kaava (13.14)). Sen 0.95. kvantiili on 1.656 (`qt(0.95,132)`).[157] Koska $|0.136| < |1.656|$, niin nollahypoteesia ei hylätä kaksisuuntaisessa testauksessa merkitsevyystasolla 0.1. Aineiston mukaan ei ole syytä luopua oletuksesta, että miehet ja naiset saavat samaa palkkaa (kun muut palkkaan vaikuttavat tekijät on huomioitu) eli että palkkasyrjintää ei ole. Ikämuuttujan estimoitu kerroin on -0.001, ja sen t-arvo on $-0.001/0.983 \approx -0.001$. Testisuureen arvon perusteella on selvää, että ikämuuttuja ei voi olla tilastollisesti merkitsevä selittäjä millään järkevällä merkitsevyystasolla. Aineiston mukaan ikä ei vaikuta siivojan palkkaan. Samoin voidaan päätellä, että työsuhteen laatu ei näytä olevan yhteydessä palkkaan ($0.169/0.747 \approx 0.226$). Työsuhteen kesto on tilastollisesti merkitsevä selittäjä palkalle: t-testisuure on 0.339 jaettuna lähes nollalla (kertoimen estimaatin keskihajonta on raportointitarkkuuden yllä mukaan 0.000 mutta todellisuudessa hieman nollaa suurempi).

Selittäjistä ainoastaan työsuhteen kesto näyttää olevan yhteydessä siivoojien palkkaan. Kukin työvuosi nostaa palkkaa 0.339 euroa.

Tutkimuksessa seuraava vaihe voisi olla estimoida malli, jossa ainoa selittäjä on työsuhteen kesto. Tällöin luultavasti saataisiin hieman eri ja keskimäärin hieman tarkempi estimaatti lisätyövuoden palkkaa nostavalle vaikutukselle. Tällaisen mallin vakiotermi olisi palkka siivoojalle, joka on juuri aloittanut siivoojan työuransa.

Mallin (13.16) vakiotermillä ei ole järkevää tulkintaa. Kirjaimellisesti tulkiten se olisi palkka 0-vuotiaalle naissiivoojalle, jonka työsuhde ei ole toistaiseksi voimassa oleva ja jonka työsuhde on vasta alkanut. □

13.6 Varianssianalyysi

Varianssianalyysi (*analysis of variance*, ANOVA) voidaan tulkita regressioanalyysin erikoistapaukseksi, jossa kaikki selittäjät ovat luokkamuuttujia (tässä yhteydessä *factor*). Tutkimuskysymys on, onko jatkuva-arvoisen selitettävän muuttujan odotusarvo sama luokkien ryhmissä vai ei. Tyypillisesti varianssianalyysilla tutkaillaan vähintään kolmen tai useamman odotusarvon yhtäsuuruutta tai eroavuutta, mutta odotusarvoja kahdessakin luokassa voidaan verrata. Mikäli luokkamuuttujia on yksi, puhutaan *yksisuuntaisesta varianssianalyysista* (*one-way analysis of variance*). Jos luokkamuuttujia on kaksi, nimitys on

kaksisuuntainen varianssianalyysi (*two-way analysis of variance*). (Jne.) Mikäli selittäjinä on myös muita kuin luokkamuuttujia, muita selittäjiä saatetaan kutsua tässä yhteydessä *kovariaateiksi* (*covariate*) ja analyysia *kovarianssianalyysiksi* (*analysis of covariance*, *ANCOVA*). Jakson 13.5.2 esimerkissä siivoojien tuntipalkoista tehtiin kovarianssianalyysia.

Yksi- ja kaksisuuntainen varianssianalyysi-nimitykset ovat harhaanjohtavia. Odotusarvot, eivät varianssit, ovat suurennuslasin alla, eikä suunnistakaan ole kyse. Nimi juontaa Ronald Fisherin tutkimuksiin 1920-luvulla ja tuli tunnetuksi Fisherin (1925) soveltajille suuntaaman vallankumouksellisen oppikirjan myötä.

Varianssianalyysissa vasteen varianssi hajotetaan luokkamuuttujien variansseihin luokissa. Menettely selitetään alla yhden luokkamuuttujan tilanteessa. Vaikka varianssihajotelmat ovat omalla tavallaan mainioita, opiskelun tehokkuuden takia alla varianssianalyysi istutetaan jo opiskeltuun regressioanalyysiin. Varianssihajotelmat ja regressioanalyysi tuottavat yhtäpitävät vastaavat testit.

13.6.1 Yksisuuntainen varianssianalyysi

Oletetaan, että satunnaismuuttujat Y_{ij} ovat normaalijakautuneita $\mathsf{N}(\mu_i, \sigma^2)$ ja riippumattomia. Indeksi i viittaa luokkaan $i = 1, \ldots, m$ ja $j = 1, \ldots, n_i$ havainnon järjestysnumeroon luokassa i. Yksisuuntainen varianssianalyysi kiteytyy F-testiin nollahypoteesille H_0: $\mu_1 = \cdots = \mu_m$. Varianssianalyysilla testataan, eroavatko odotusarvot luokissa. Vastahypoteesi on, että ainakin yksi odotusarvoista poikkeaa muista.

Lasketaan luokkakohtaiset otoskeskiarvot

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

ja kaikkien havaintojen keskiarvo eli *kokonaiskeskiarvo* (*grand mean*)

$$\bar{y} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}}{n}.$$

Yllä $n = n_1 + \cdots + n_m$. Neliöimällä identiteetti

$$y_{ij} - \bar{y} = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

puolittain seuraa neliösummahajotelma

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$= \sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2.$$

(Neliöinnistä muodostuva ristitulotermi häviää summauksessa. Harjoitustehtävä.) Sen 1., 4. ja 5. kaksoissummaa kutsutaan kokonaisneliösummaksi (KNS), *luokkaneliösummaksi* (LNS) ja jäännösneliösummaksi (JNS):

$$\mathsf{KNS} = \mathsf{LNS} + \mathsf{JNS}.$$

Mikäli nollahypoteesi pätee, luokkakohtaiset otoskeskiarvot $\bar{y}_i$ eivät tapaa poiketa suuresti kokonaiskeskiarvosta $\bar{y}$ eikä luokkaneliösumma LNS tapaa muodostua suureksi. Testisuure vertaa *keskineliösummien* $\mathsf{LNS}/(m-1)$ ja $\mathsf{JNS}/(n-m)$ suhdetta ja noudattaa nollahypoteesin pätiessä $\mathsf{F}(m-1, n-m)$-jakaumaa:

$$F = \frac{\mathsf{LNS}/(m-1)}{\mathsf{JNS}/(n-m)} \sim \mathsf{F}(m-1, n-m). \tag{13.17}$$

Testisuure punnertaa hylkäysalueelle, jos $\mathsf{LNS}/(m-1)$ on suuri suhteessa normeeraavaan tekijään satunnaistermin varianssin estimaattiin $\mathsf{JNS}/(n-m)$.

Joskus varianssianalyysin tulos esitetään varianssianalyysitaulukkona:

vaihtelu	neliö-summa	vapaus-asteet	keskineliö-summa	F
luokka	LNS	$m-1$	$\mathsf{LNS}/(m-1)$	$[\mathsf{LNS}/(m-1)]/[\mathsf{JNS}/(n-m)]$
jäännös	JNS	$n-m$	$\mathsf{JNS}/(n-m)$	
kokonais	KNS	$n-1$	$\mathsf{KNS}/(n-1)$	

Varianssianalyysi-nimitys juontaa siitä, että havaintojen vaihtelu ositetaan luokkien ja jäännöksen vaihteluun, joita verrataan.

Testi voidaan pukea regressioanalyyttiseen kaapuun. Estimoidaan PNS:llä malli

$$y_{ij} = \beta_0 + \beta_2 x_{2j} + \cdots + \beta_m x_{mj} + \varepsilon_{ij}. \tag{13.18}$$

Siinä x_i on osoitinmuuttuja, joka saa arvon 1, jos havainto kuuluu luokkaan i, ja 0 muulloin. Ensimmäisen luokka on vertailuluokka, jolle ei ole osoitinmuuttujaa. Mallin vakiotermi tekeytyy ensimmäisen luokan odotusarvoksi. Regressiokertoimet nappaavat kunkin luokan odotusarvon poikkeaman ensimmäisen luokan odotusarvosta. Voidaan osoittaa, että mallista (13.18) laskettu F-testisuure

nollahypoteesille H_0: $\beta_2 = \cdots = \beta_m = 0$ on F-testisuure (13.17) nollahypoteesille H_0: $\mu_1 = \cdots = \mu_m$.

Yksinkertaisimmillaan testataan kahden luokan odotusarvon yhtäsuuruutta. Tällöin testi voidaan tehdä yhtäpitävästi myös regressiokertoimen nolluutta testaavalla t-testillä tai kahden odotusarvon erotusta yhtäsuurien varianssien tilanteessa testaavalla t-testillä (jakso 12.4.5).

Esimerkki. Uraseurantakysely.[158] Valtakunnallisella maistereiden uraseurantakyselyllä haetaan tietoa heidän sijoittumisestaan työelämään ja tyytyväisyydestä tutkintoonsa. Helsingin yliopiston 11 tiedekunnasta vuonna 2011 valmistuneille syksyllä 2016 suoritetun kyselyn miesten ja naisten tiedekunnittaiset vastausprosentit ovat kuvassa 13.10. Estimoidaan PNS:llä malli, jossa selitetään vastausprosentteja sukupuoliosoittimella (x: 0 = mies ja 1 = nainen):

$$\begin{array}{rcccccc} y & = & 33.571 & + & 6.595x & + & \hat{\varepsilon}. \\ & & (2.579) & & (3.647) & & \end{array}$$

$$s = 2.924,\ \mathsf{R}^2 = 0.141,\ F = 3.270,\ n = 22.$$

Parametrien estimaattoreiden estimoidut keskivirheet ovat suluissa.

Jäännöksen keskihajonta on 2.9-prosenttiyksikköä. Se kuvaa prosenttiosuuksien vaihtelun suuruutta aineistossa. Tässä mallissa vakiotermillä on selkeä tulkinta: Vakiotermin estimaatti on miesten estimoitu vastaustodennäköisyys 33.6 %. Mallin mukaan naisten vastaustodennäköisyys $33.571 + 6.595 \approx 40.2$ % on 6.6 %-yksikköä suurempi kuin miesten. Ero ei ole tilastollisesti merkitsevä: F-testisuureen p-arvo on 0.086 (`1-pf(3.27,1,20)`). F-testisuure on t-testisuureen neliö: $3.270 \approx 1.808314^2 \approx (6.595/3.647)^2$. Sama p-arvo saadaan `t`-jakaumasta (`2*(1-pt(1.808314,20))`).

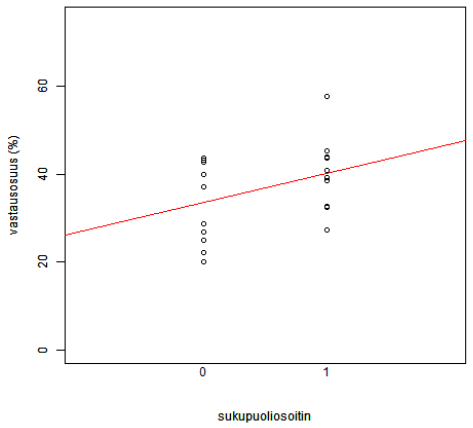
Testissä vertaillaan kahta odotusarvoa olettaen havainnot normaalisiksi niiden ympärillä. Testitilanne on sama kuin jaksossa 12.4.5. Siellä kuvatun t-testisuureen laskeva R-komento `t.test(x1,x2, var.equal=TRUE)` tulostaa niinikään p-arvon 0.086 (kun `x1` ja `x2` sisältävät ryhmien havainnot).

Testien nollahypoteesin mukaan havainnoilla on sama odotusarvo — esimerkin tilanteessa vastaustodennäköisyys — ja varianssi. Keskeisen raja-arvolauseen (7.3) perusteella kukin vastausosuushavainto on likimäärin $\mathsf{N}(\pi, \pi(1-\pi)/n^*)$-jakautunut, jos vastausastehavainnot ovat samankokoisista otoksista ($n_i = n^*$, $i = 1, \ldots, n$) ja π on kunkin kontaktoidun todennäköisyys vastata (jakso 9.3). Kuvatut kolme testiä (F-testi, regressiokertoimen t-testi ja odotusarvojen erotuksen t-testi) ovat yhtäpitävät. Nollahypoteesin pätiessä testisuureiden jakaumat pätevät likimäärin, kun vastausosuushavainnot on

laskettu suurista havaintomääristä n^*, jolloin vastausosuudet ovat likimäärin normaalijakautuneita.

Huom1! Jos vastaustodennäköisyys eroaa sukupuolilla, satunnaistermin varianssi eroaa selittäjän kahdella arvolla. PNS ei ole tällöin paras mahdollinen estimointimenetelmä. Asiaan palataan jaksoissa 13.7, 13.11.3 ja 14.1.

Huom2! Osoitinselittäjän tilanteessa mallia ei pidä soveltaa muille arvoille kuin 0 ja 1! Muilla arvoilla ei ole mielekästä tulkintaa. □



Kuva 13.10: Helsingin yliopiston 11 tiedekunnasta vuonna 2011 valmistuneiden maistereiden vastausosuudet syksyn 2016 uraseurantakyselyssä. 0 = mies. 1 = nainen.

Esimerkki. Nuoren tytön itsevarmuus.[159] Malli on

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

Selitettävä muuttuja on peruskoulun yläasteikäisen (7.–9.-luokkalaisen) nuoren tytön itsevarmuusindeksi. Se saa arvoja välillä $[1.0, 5.0]$ (suuri arvo, suuri itsevarmuus). Selittäjä on luokka-asteikollinen äidin kasvatustyyli. Kasvatustyylejä on neljä: laiminlyövä, autoritaarinen (x_2), salliva (x_3) ja auktoritatiivinen (x_4). Kolmea viimeksi mainittua kasvatustyyliä varten on luotu osoitinmuuttu-

ja. PNS-estimointi tuottaa mallin

$$\begin{array}{rcccccccc} y & = & 2.833 & + & 0.308x_2 & + & 0.510x_3 & + & 0.670x_4 & + & \hat{\varepsilon}. \\ & & (0.150) & & (0.250) & & (0.205) & & (0.171) & & \end{array}$$

$$s = 0.823,\ \mathsf{R}^2 = 0.0835,\ F = 5.435,\ n = 183.$$

Äidin laiminlyövää kasvatustyyliä kokeneen tytön itsevarmuusindeksi on mallin mukaan keskimäärin 2.83. Kaikilla muilla kasvatustyyleillä tytön itsevarmuusindeksi estimoituu suuremmaksi. Muiden kasvatustyylien ero on tilastollisesti merkitsevä: F-testisuure noudattaa nollahypoteesin pätiessä $\mathsf{F}(3, 179)$-jakaumaa. F-testisuureen p-arvo on 0.001 (`1-pf(5.435,3,179)`). Tytön itsevarmuus ja äidin kasvatustyyli ovat yhteydessä toisiinsa.

Regressio viestii, että äidin auktoritatiivisessa kasvatuksessa varttuneet tytöt ovat erityisen itsevarmoja. Heidän itsevarmuusindeksinsä on keskimäärin $2.83 + 0.67 = 3.50$. Ero äidin kasvatuksellisesti laiminlyömiin tyttöihin on tilastollisesti merkitsevä: $0.670/0.171 \approx 3.915$. Sen p-arvo on alle 0.001 (`2*(1-pt(3.915, 179))`). Myös äidin salliva kasvatustyyli vaikuttaa olevan yhteydessä tytön suurempaan itsevarmuuteen: Itsevarmuusindeksi on tällöin $2.83 + 0.51 = 3.34$. Estimoidun regressiokertoimen t- ja p-arvot ovat $0.510/0.205 \approx 2.490$ ja 0.014 (`2*(1-pt(2.490,179))`). □

13.6.2 Kaksisuuntainen varianssianalyysi

Kaksisuuntaisessa varianssianalyysissa tutkitaan riippumattomien satunnaismuuttujien $Y_{ijk} \sim \mathsf{N}(\mu_{ij}, \sigma^2)$ odotusarvojen μ_{ij} mahdollista riippuvuutta kahdesta luokkamuuttujasta x ja z. Indeksit i ja j nimeävät luokat $i = 1, \ldots, I$ ja $j = 1, \ldots, J$, ja $k = 1, \ldots, n_{ij}$ on havainnon järjestysnumero, kun havainto on luokkamuuttujien luokkien leikkauksessa eli solussa ij.

$I \times J$ -taulukko

		z			
		z_1	$\cdots$	z_J	
x	x_1	μ_{11}	$\cdots$	μ_{1J}	$\mu_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	μ_{I1}	$\cdots$	μ_{IJ}	$\mu_{I\cdot}$
		$\mu_{\cdot 1}$	$\cdots$	$\mu_{\cdot J}$	$\mu_{\cdot\cdot}$

havainnollistaa odotusarvojen määräytymistä. Kullakin x:n ja z:n luokan arvolla eli kussakin ij-solussa odotusarvo μ_{ij} voi olla eri. Oikeanpuolimmaisessa sarakkeessa on kunkin rivin odotusarvojen keskiarvo *reunakeskiarvo*

$$\mu_{i\cdot} = \frac{\sum_{j=1}^{J} \mu_{ij}}{J}$$

ja alimmalla rivillä kunkin sarakkeen reunakeskiarvo

$$\mu_{\cdot j} = \frac{\sum_{i=1}^{I} \mu_{ij}}{I}.$$

Kokonaiskeskiarvo on kaikkien odotusarvojen keskiarvo:

$$\mu_{\cdot\cdot} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} \mu_{ij}}{I \times J} = \frac{\sum_{i=1}^{I} \mu_{i\cdot}}{I} = \frac{\sum_{j=1}^{J} \mu_{\cdot j}}{J}.$$

Reunakeskiarvot (ja kokonaiskeskiarvo) summeeraavat huomion kohteena olevat soluodotusarvot μ_{ij} mutteivät välttämättä yhdy vastaaviin riveittäisiin ja sarakkeittaisiin Y_{ijk}:n odotusarvoihin.

Mikäli satunnaismuuttujan Y_{ij} odotusarvo määräytyy kummastakin luokkamuuttujasta itsenäisesti toisen luokkamuuttujan arvosta riippumatta, erotus

$$\mu_{ij} - \mu_{i^*j} = \mu_{i\cdot} - \mu_{i^*\cdot}. \tag{13.19}$$

ei riipu indeksistä j ($i \neq i^* = 1, \ldots, I$) eikä erotus

$$\mu_{ij} - \mu_{ij^*} = \mu_{\cdot j} - \mu_{\cdot j^*} \tag{13.20}$$

indeksistä i ($j \neq j^* = 1, \ldots, J$). Tällöin erotukset ovat samat eri sarakkeissa (yhtälö (13.19)) tai eri riveillä (yhtälö (13.20)). Erotukset (13.19) ovat luokkamuuttujan x ja erotukset (13.20) luokkamuuttujan z *päävaikutukset* (*main effect*).

Luokkamuuttujilla voi olla *yhdysvaikutuksia* (*interaction*) satunnaismuuttujan Y_{ij} odotusarvoon. Silloin yhtälöt (13.19) ja (13.20) eivät ylipäänsä päde ja vastaavien solujen odotusarvojen erotukset voivat olla erisuuria eri sarakkeilla tai riveillä.

Kuva (13.11) havainnollistaa tilannetta $I = J = 2$:

		z		
		z_1	z_2	
x	x_1	μ_{11}	μ_{12}	$\mu_{1\cdot}$
	x_2	μ_{21}	μ_{22}	$\mu_{2\cdot}$
		$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	

Kuvan vasemmassa puoliskossa satunnaismuuttujan Y_{ij} odotusarvo kasvaa päävaikutuksen $\mu_{\cdot 2} - \mu_{\cdot 1}$ verran luokkamuuttujan z arvon muuttuessa z_1:stä z_2:hteen riippumatta luokkamuuttujan x arvosta (x_1 tai x_2). Kuvan oikeassa puoliskossa yhdysvaikutus tekee tekosiaan. Luokkamuuttujan z arvon muutoksen vaikutus satunnaismuuttujan Y_{ij} odotusarvoon riippuu siitä, kumpi luokkamuuttujan x arvoista pätee. Erotus $\mu_{\cdot 2} - \mu_{\cdot 1}$ ei enää piirrä satunnaismuuttujan Y_{ij} odotusarvon muutosta, kun luokkamuuttujan z arvo vaihtuu z_1:stä z_2:hteen.

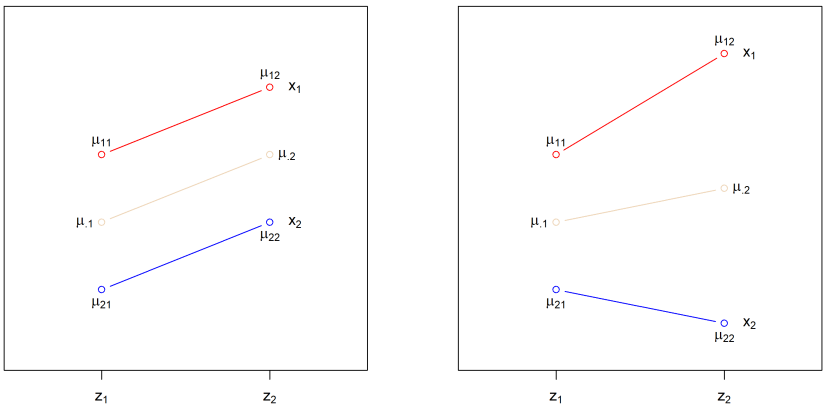
Tilanne $I = 2$ ja $J = 3$ eli

		z			
		z_1	z_2	z_3	
x	x_1	μ_{11}	μ_{12}	μ_{13}	$\mu_{1\cdot}$
	x_2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\cdot}$
		$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot 3}$	

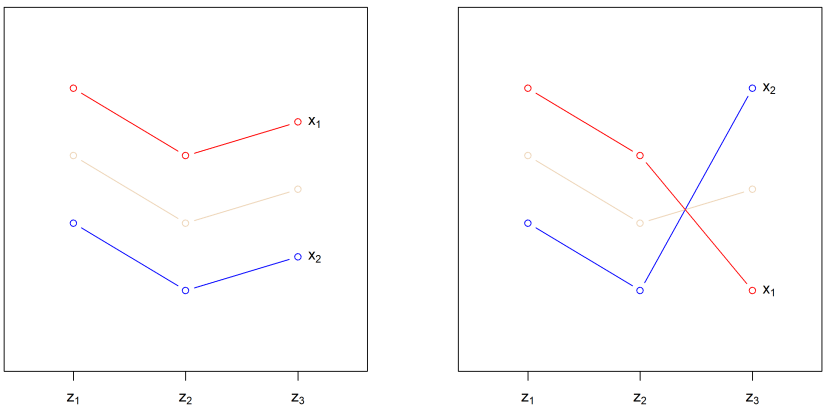
on kuvan (13.12) aihe. Ruudussa vasemmalla satunnaismuuttujan Y_{ij} odotusarvo muuttuu luokkamuuttujan z arvojen mukaan riippumatta luokkamuuttujan x arvosta. Keskimmäinen käyrä ($\mu_{\cdot 1} \to \mu_{\cdot 2} \to \mu_{\cdot 3}$) polveilee erotusten eli luokkamuuttujan z päävaikutusten mukaisesti. Ruudussa oikealla on yhdysvaikutus luokkamuuttujan z arvolla z_3. Erotus $\mu_{\cdot 3} - \mu_{\cdot 2}$ ei enää kuvaa satunnaismuuttujan Y_{ij} odotusarvon muutosta siirtymässä $z_2 \to z_3$ kummallakaan luokkamuuttujan x arvolla x_1 tai x_2.[160]

Jatketaan tarkastelua regressiokehikossa. Vakiotermi toimii vertailuluokkana. Pää- ja yhdysvaikutukset voidaan pyydystää osoitinmuuttujien avulla. Luokkamuuttujia x ja z varten luodaan $I - 1$ ja $J - 1$ ja yhdysvaikutuksia varten $(I - 1) \times (J - 1)$ osoitinmuuttujaa. Regressiomallissa on parametreja $1 + (I - 1) + (J - 1) + (I - 1) \times (J - 1) = I \times J$ kappaletta eli yhtä monta kuin luokkien yhdistelmiä. $(I \times J)$:llä parametrilla voi kuvata minkälaisen tahansa odotusarvorakenteen $(I \times J)$:ssä solussa.

Taulukko



Kuva 13.11: Luokkamuuttujat ja odotusarvot (2×2 -taulukko). Vasen: Ei yhdysvaikutusta. Oikea: Yhdysvaikutus.



Kuva 13.12: Luokkamuuttujat ja odotusarvot (2×3 -taulukko). Vasen: Ei yhdysvaikutusta. Oikea: Yhdysvaikutus.

x	z	x_2	z_2	z_3	$x_2 \times z_2$	$x_2 \times z_3$	μ_{ij}
1	1	0	0	0	0	0	$\mu_{11} = \beta_0$
1	2	0	1	0	0	0	$\mu_{12} = \beta_0 + \gamma_2$
1	3	0	0	1	0	0	$\mu_{13} = \beta_0 + \gamma_3$
2	1	1	0	0	0	0	$\mu_{21} = \beta_0 + \beta_2$
2	2	1	1	0	1	0	$\mu_{22} = \beta_0 + \beta_2 + \gamma_2 + \delta_{22}$
2	3	1	0	1	0	1	$\mu_{23} = \beta_0 + \beta_2 + \gamma_3 + \delta_{23}$

konkretisoi tilanteen $I = 2$ ja $J = 3$. Parametreja tarvitaan tällöin $2 \times 3 = 6$: vakiotermille 1, x:lle 1, z:lle 2 ja 2 poimimaan yhdysvaikutukset. Kaksi ensimmäistä saraketta erittelevät luokkien kaikki mahdolliset yhdistelmät. Taulukkoa vastaava malli on

$$y = \beta_0 + \beta_2 x_2 + \gamma_2 z_2 + \gamma_3 z_3 + \delta_{22} x_2 z_2 + \delta_{23} x_2 z_3 + \varepsilon. \tag{13.21}$$

Sillä voidaan mallittaa vasteen Y_{ij} odotusarvo kaikissa luokkien yhdistelmissä. Odotusarvon yhteys parametreihin on kuvattu taulukon viimeisessä sarakkeessa. Jos yhdysvaikutuksia ei ole, osoittimien x_2, z_2 ja z_3 kertoimet β_2, γ_2 ja γ_3 ovat niiden päävaikutukset. Yhdysvaikutustilanteessa ristitulotermien $x_2 z_2$ ja $x_2 z_3$ kertoimet muokkaavat odotusarvot μ_{22} ja μ_{23} yhdysvaikutusten mukaisiksi.

Kuvan (13.12) oikean ruudun poukkoilevat odotusarvot voi ajatella poimittaviksi mallilla (13.21) näin: Suureet β_0, $\beta_0 + \gamma_2$ ja $\beta_0 + \gamma_3$ viitoittavat punaisen murtoviivan ($\mu_{11} \to \mu_{12} \to \mu_{13}$). Summa $\beta_0 + \beta_2$ merkitsee sinisen murtoviivan ($\mu_{21} \to \mu_{22} \to \mu_{23}$) alkupisteen. Jäljellä olevat kaksi parametria δ_{22} ja δ_{23} seuraavat muista sopeutumalla yhtälöihin $\mu_{22} = \beta_0 + \beta_2 + \gamma_2 + \delta_{22}$ ja $\mu_{23} = \beta_0 + \beta_2 + \gamma_3 + \delta_{23}$.

Esimerkki. Nuoren tytön itsevarmuus (jatkoa). Mallitetaan nuoren tytön itsevarmuutta sekä äidin (x) että isän (z) kasvatustyyleillä (laiminlyövä = 1, autoritaarinen = 2, salliva = 3 ja auktoritatiivinen = 4). Kuudesta havainnosta puuttuu tieto isän kasvatustyylistä. Jatketaan mallittamista $183 - 6 = 177$:llä havainnolla, joissa on tieto sekä äidin että isän kasvatustyylistä.

Estimoidaan PNS-menetelmällä pelkkiä päävaikutuksia sisältävä malli

$$\begin{aligned} y = \; & \underset{(0.158)}{2.820} + \underset{(0.258)}{0.138 x_2} + \underset{(0.226)}{0.407 x_3} + \underset{(0.205)}{0.348 x_4} \\ & - \underset{(0.193)}{0.002 z_2} + \underset{(0.208)}{0.089 z_3} + \underset{(0.184)}{0.571 z_4} + \hat{\varepsilon}. \end{aligned}$$

$$s = 0.804,\ \mathsf{R}^2 = 0.1545,\ F = 5.178,\ n = 177.$$

F-testisuureen p-arvo on alle 0.0001 (`1-pf(5.178,6,170)`), joten ainakin yksi selittäjistä on yhteydessä itsevarmuuteen. Äidin sallivalla tai auktoritatiivisella kasvatustyyleillä on kohtuullisen kokoiset kertoimet, mutta niiden t-arvojen p-arvot ovat suurehkoja (0.074 ja 0.091). Isän auktoritatiivisen kasvatustyylin kerroin on suurin ja yksin sen t-arvo $0.571/0.184 \approx 3.095$ vaikuttaa huomionarvoiselta ($p = 0.002$; `2*pt(-3.095, 170)`). Selitysaste 0.155 on lähes kaksi kertaa 0.084 eli selitysaste selitettäessä itsevarmuutta yksin äidin kasvatustyyleillä. Vertailua vaikeuttaa, että malleissa ei ole selitetty aivan samoja havaintoja ($n = 183$ tai $n = 177$).

Estimoidaan PNS-menetelmällä yhdysvaikutusmalli, jossa itsevarmuutta selitetään äidin ja isän auktoritatiivisella kasvatustyylillä ja niiden yhdysvaikutuksella:

$$\begin{array}{rcccccccc} y & = & 3.090 & + & 0.027x_4 & + & 0.256z_4 & + & 0.424x_4z_4 \quad + \quad \hat{\varepsilon}. \\ & & (0.097) & & (0.159) & & (0.261) & & (0.309) \end{array}$$

$$s = 0.805,\ \mathsf{R}^2 = 0.1386,\ F = 9.277,\ n = 177.$$

Vakiotermi nappaa nyt itsevarmuuden ryhmässä, jossa vanhempien kasvatustyyli on ollut laiminlyövä, autoritatiivinen tai salliva. Estimaatti 3.09 tällaisissa kodeissa kasvaneiden tyttöjen itsevarmuusindeksille on suurempi kuin edellisen mallin 2.820. Se oli estimaatti itsevarmuusindeksin arvolle laiminlyövässä kasvatustyylissä kasvaneiden tyttöjen luokassa. Vakiotermin estimaatin suureneminen on luontevaa.

Selittäjät ovat mielekkäitä; F-testisuureen p-arvo on alle 0.001:n (`1-pf(9.277,3,173)`). Selittäjien kertoimien t-testisuureet eivät osoita yhtään selittäjistä yksinään erityisen tärkeäksi. Selitysaste 0.139 on huomattavasti suurempi kuin äidin kasvatustyylien toimiessa yksin selittäjinä (vertailua ongelmoi eroavat aineistot) ja lähellä juuri estimoidun paljon useampia selittäjiä sisältävän mallin selitysastetta. Estimoitujen kertoimien mukaan nuoren tytön suuri itsevarmuus liittyy varsinkin isän auktoritatiiviseen kasvatustyyliin ja erityisesti olosuhteeseen, jossa sekä äiti että isä ovat molemmat kasvatustyyliltään auktoritatiivisia. Viimeksi mainitussa tilanteessa estimoitu odotusarvo tytön itsevarmuusindeksille on $3.090 + 0.027 + 0.256 + 0.424 \approx 3.80$. Se on suurempi kuin estimoitu itsevarmuusindeksin maksimaalinen odotusarvo 3.50 selittäjien ollessa yksin äitien kasvatustyylejä. Jaksossa 13.11.2 pohditaan, milloin estimoitujen regressiokertoimien suuruudesta voidaan tehdä päätelmiä selittäjien tärkeydestä. □

13.7 PNS-estimaattorin optimaalisuus ja tarkentuvuus

Estimoidaan yhtälön (13.10) parametrit PNS:llä. Oletetaan, että

1. selittäjät ovat kiinteitä ja sellaisia, että parametrit ovat estimoitavissa,
2. satunnaistermit eivät korreloi keskenään eivätkä selittäjien kanssa,
3. satunnaistermien odotusarvo on nolla ja niiden varianssi on vakio ja että
4. satunnaistermit ovat normaalijakautuneita.

Voidaan osoittaa, että kolmen ensimmäisen oletuksen pätiessä PNS-estimaattori on paras harhaton lineaarinen estimaattori.[161] Mikäli estimoitavia parametreja on vain yksi, "paras" on pienimmän keskineliövirheen — estimaattorin ollessa harhaton pienimmän keskivirheen — omaava. Mikäli estimoitavia parametreja on useita, "paras" tarkoittaa, että mikä tahansa estimoitujen parametrien lineaarikombinaatio $\sum_{i=1}^{k} a_i \hat{\beta}_i$ on keskivirheeltään pienin mahdollinen lineaarikombinaation $\sum_{i=1}^{k} a_i \beta_i$ estimaattori (a_i:t ovat mielivaltaisia vakioita). Erityisesti kunkin parametrin β_i estimaattorin $\hat{\beta}_i$ keskivirhe on pienin mahdollinen (tulos seuraa asettamalla $a_i = 1$ ja $a_j = 0$, $j \neq i$). Tämä Gaussin–Markovin lause on tilastotieteen tunnetuimpia tuloksia. Mikäli oletetaan lisäksi satunnaistermien normaalisuus, PNS-estimaattori on keskineliövirheellä mitattuna paras harhaton estimaattori (ei vain lineaaristen estimaattorien joukossa).

PNS-estimaattori on tarkentuva, elleivät selittäjät käyttäydy hyvin erikoisesti. Tarkentuvuus edellyttää, että havainnot kustakin x_j-selittäjästä vaihtelevat riittävästi ($\sum_{j=1}^{n} x_{ji}^2/n$ ei mene nollaan n suuretessa) ja ettei selittäjien välille muodostu deterministisiä lineaarisia kytköksiä havaintomäärän kasvaessa.

PNS-estimaattorin ominaisuudet ovat vaikuttavia. Silti kannattaa pitää jalat maassa: Tulokset olettavat, että malli (13.10) on oikein määritelty ja että oletukset pätevät. Niin ei välttämättä ole empiirisessä tutkimuksessa. Selittäjien vaikutus saattaa välittyä epälineaarisesti selitettävään eikä satunnaistermi ole välttämättä normaalijakautunut tai muuten oletusten mukainen.

13.8 Ennustaminen

Ihmismieltä kiehtoo ajatus, että kykenesimme ennustamaan. Regressiomallilla kykenemme. Monen regressiomallin pääkäyttötarkoitus on ennustaminen. Mallin selittäjiä — tai muuttujia, joista selittäjät on funktiomuunnoksin luotu —

saatetaankin kutsua ennustimiksi (*predictor*; Faraway 2015, 7, Fox ja Weisberg 2019, 174, Weisberg 2014, 56). Ennustaminen ei tässä yhteydessä lähtökohtaisesti tarkoita tulevaisuuteen kurkistamista, sillä regressiomalli ei ylipäänsä konkreettisesti kiinnity aikaan. Edellisen jakson ehtojen 1–4 oletetaan pätevän alla.

Regressiomallilla voidaan ennustaa vasteen odotusarvoa $\mathsf{E}(Y_u) = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ui}$ tai havaintoa vasteesta $Y_u = \mathsf{E}(Y_u) + \varepsilon_u$, kun selittäjäyhdistelmä on $x_{u1}, \ldots, x_{uk}$. Alaindeksi u viittaa uuteen havaintovektoriin $[x_{u1} \ \ldots \ x_{uk} \ y_u]$, joka ei kuulu aineistoon. Vasteen odotusarvon ennuste on

$$\hat{\mathsf{E}}(Y_u) \equiv \hat{\beta}_0 + \hat{\beta}_1 x_{u1} + \cdots + \hat{\beta}_k x_{uk}.$$

$\hat{\mathsf{E}}(Y_u)$ on merkintä summalle $\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_{ui}$. Ennuste on harhaton eli

$$\mathsf{E}[\hat{\mathsf{E}}(Y_u)] = \beta_0 + \beta_1 x_{u1} + \cdots + \beta_k x_{uk}, \tag{13.22}$$

koska PNS-estimaattorit $\hat{\beta}_i$ ovat harhattomia. Sama ennuste on myös uuden havainnon Y_u harhaton ennuste, sillä $\mathsf{E}(\varepsilon_u) = 0$.

Ennusteen tarkkuuden arviointi on oleellista. Vasteen odotusarvon ennuste on selittäjäyhdistelmän mukainen lineaarikombinaatio PNS-estimaattoreista $\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_{ui}$. Odotusarvon ennusteen varianssi riippuu selittäjäyhdistelmästä $x_{u1}, \ldots, x_{uk}$, aineiston selittäjävektoreista $[x_{i1} \ \ldots \ x_{ik}]$, $i = 1, \ldots, n$ sekä satunnaistermin varianssista σ^2. Odotusarvon ennusteen varianssi on esitettävissä muodossa

$$\mathsf{V}[\hat{\mathsf{E}}(Y_u)] = \mathsf{V}[\hat{\mathsf{E}}(Y_u) - \mathsf{E}(Y_u)] = \sigma^2 v_u$$

(vakion vähentäminen satunnaismuuttujasta ei muuta varianssia). Varianssin komponentti v_u määräytyy yksinomaan selittäjien arvoista, havaintojen lukumäärästä ja selittäjäyhdistelmästä $x_{u1}, \ldots, x_{uk}$, eli v_u on laskettavissa ja tunnettu. Sen kaava yhden selittäjän tilanteessa löytyy alta.

$100 \times (1 - \alpha)$ %:n luottamusväli odotusarvolle $\mathsf{E}(Y_u)$ on

$$\hat{\mathsf{E}}(Y_u) \pm \mathsf{t}_{1-\alpha/2}(n - k - 1) s \sqrt{v_u} \tag{13.23}$$

(Seber ja Lee 2003, 129). Yllä s on regressiomallin satunnaistermin keskihajonnan estimaatti kaavasta (13.12). Tyypillisemmin pyritään ennustamaan uutta havaintoa. Sen ennustevirheen varianssi on

$$\begin{aligned} \mathsf{V}[\hat{\mathsf{E}}(Y_u) - Y_u] &= \mathsf{V}\{\hat{\mathsf{E}}(Y_u) - [\mathsf{E}(Y_u) + \varepsilon_u]\} = \mathsf{V}\{[\hat{\mathsf{E}}(Y_u) - \mathsf{E}(Y_u)] - \varepsilon_u\} \\ &= \sigma^2 v_u + \sigma^2 = \sigma^2 (v_u + 1). \end{aligned}$$

Kolmas yhtäsuuruus seuraa satunnaistermin riippumattomuudesta erotuksesta $[\hat{\mathsf{E}}(Y_u) - \mathsf{E}(Y_u)]$. Uuden havainnon $100 \times (1 - \alpha)$ %:n *ennusteväli* (*prediction interval*) on

$$\hat{\mathsf{E}}(Y_u) \pm \mathsf{t}_{1-\alpha/2}(n-k-1)s\sqrt{v_u + 1} \tag{13.24}$$

(Seber ja Lee 2003, 132). Uuden havainnon ennusteväli (13.24) on yleensä huomattavasti leveämpi kuin sen odotusarvon ennusteen luottamusväli (13.23).

Luottamusväli peittää parametrin arvon tietyllä todennäköisyydellä riippumattomissa toistokokeissa. Ennusteväli peittää uuden havainnon vastaavasti.

Yhden selittäjän tilanteessa odotusarvon ennusteen ja uuden havainnon ennusteen varianssit ovat

$$\sigma^2 v_u = \sigma^2 \left(\frac{1}{n} + \frac{(x_{u1} - \bar{x}_1)^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} \right)$$

ja

$$\sigma^2 v_u + \sigma^2 = \sigma^2 \left(\frac{1}{n} + \frac{(x_{u1} - \bar{x}_1)^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} \right) + \sigma^2$$

(Weisberg 2014, 294–295). Edellä todettu v_u:n eksklusiivinen riippuvuus selittäjien arvoista typistyy yllä x_{i1}:n $(i = 1, \ldots, n)$ ja x_{u1}:n arvoihin. Uuden havainnon ennusteen varianssin kaavassa viimeinen termi σ^2 tulee satunnaistermistä ε_u. Termi σ^2 on edellistä termiä suurempi, ellei x_{u1} poikkea hyvin paljon $\bar{x}_1$:stä. Muulloin yksittäisen havainnon sattumanvaraisuus dominoi uuden havainnon ennusteen varianssia.

Vasteen odotusarvon ja uuden havainnon $100 \times (1 - \alpha)$ %:n luottamus- ja ennustevälien ala- ja ylärajat ovat

$$\begin{aligned} &\hat{\mathsf{E}}(Y_u) \pm \mathsf{t}_{1-\alpha/2}(n-k-1)s\sqrt{v_u} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{u1} \pm \mathsf{t}_{1-\alpha/2}(n-k-1)s\sqrt{\frac{1}{n} + \frac{(x_{u1} - \bar{x}_1)^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}} \end{aligned} \tag{13.25}$$

ja

$$\begin{aligned} &\hat{\mathsf{E}}(Y_u) \pm \mathsf{t}_{1-\alpha/2}(n-k-1)s\sqrt{v_u + 1} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{u1} \pm \mathsf{t}_{1-\alpha/2}(n-k-1)s\sqrt{\frac{1}{n} + \frac{(x_{u1} - \bar{x}_1)^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} + 1}. \end{aligned} \tag{13.26}$$

Luottamus- ja ennusteväli ovat sitä leveämmät, mitä enemmän x_{u1} poikkeaa $\bar{x}_1$:stä. Selitys on, että pienikin ero $\hat{\beta}_1$:n ja β_1:n välillä johtaa suureen eroon

odotusarvon ennusteen $\hat{\mathsf{E}}(Y_u)$ ja odotusarvon $\mathsf{E}(Y_u)$ välille, jos x_{u1} eroaa paljon $\bar{x}_1$:stä (vrt. kuvat 13.7 ja 13.13).

Ennusteen harhattomuus (13.22) sekä luottamus- ja ennustevälit (13.23) ja (13.24) pätevät, vaikka selittäjäyhdistelmä $x_{u1}, \ldots, x_{uk}$ (yhden selittäjän tilanteessa x_{u1}:n arvo) poikkeaisi mielivaltaisen paljon aineiston selittäjäyhdistelmistä. Nämä tulokset edellyttävät tietenkin, että estimoitu malli on oikein määritelty. Mikäli ennuste lasketaan samantapaisella selittäjäyhdistelmällä kuin jolla malli on estimoitu, on luontevaa olettaa malli päteväksi ja ennustamisessa hyödynnettäväksi. Mikäli selittäjäyhdistelmä poikkeaa selvästi aineistossa olleista, ennuste ei ole välttämättä luotettava eikä sen tarkkuuden arviointi kaavoilla edellä osuvaa. Vaikka ennusteiden luottamus- ja ennustevälit ovat leveämmät tällaisissa tilanteissa, ne mittaavat epävarmuuden lisääntymistä mallin puitteissa. Todellinen epävarmuus voi olla suurempaa, kun mallin toimivuutta venytetään havaintojen, joista mallin parametrit on estimoitu, ulkopuolelle. Teoreettinen havainnollistus on kuvassa 13.7. Esimerkki alla havainnollistaa välien leventymistä mallin ennustamalla tavalla.

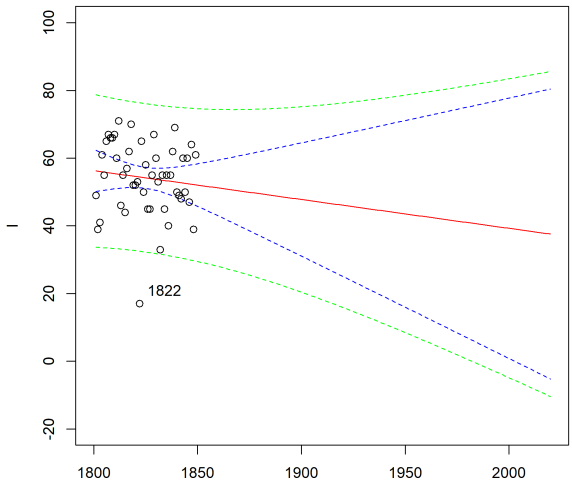
Esimerkki. Jäidenlähtö Kokemäenjoesta. Eklöf (1850) julkaisi ensimmäisen suomenkielisen tilastotieteellisen tutkimuksen. Se oli myös lähes ensimmäinen suomenkielinen tieteellinen tutkimus (Alho ym. 2021, luku 3). Eklöf estimoi PNS-menetelmällä jäidenlähdön päivää Kokemäenjoesta 1801–1849 selittävän yhtälön

$$l = 56.311 - 0.0851 \times (v - 1800) + \hat{\varepsilon}.$$

Siinä l on päivien lukumäärä maaliskuun 1. päivästä alkaen jäidenlähtöön (esim. 2.4. $\rightarrow$ 33) ja v on vuosi ($v = 1801, \ldots, 1849$). Eklöfin mukaan jäidenlähtö on aikaistunut 4.3 päivää 1800-luvun alkupuoliskolla. Malli ennustaa 18.7 päivän aikaistumista 2020. (Harjoitustehtävä.)

Kuvassa 13.13 on sirontakuvio havainnoista ja jäidenlähtöpäivän (maaliskuun 1:stä) odotusarvon ja uuden havainnon ennusteen yhtälöiden (13.25) ja (13.26) mukaiset 95 %:n luottamus- ja ennustevälit ajanjaksolle 1801–2020. Luottamus- ja ennusteväli levenevät etäännyttäessä selittäjän keskiarvosta 1825. Välit suurentuvat nopeasti ekstrapoloitaessa jäidenlähtöpäivän odotusarvoa ja ennustetta. Aineiston vaihteluvälin sisällä epävarmuus juontaa paljolti satunnaistermin vaihtelusta; selittäjän ääriarvoilla yhä enemmän ja ekstrapoloitaessa ennen kaikkea regressiosuoran sijoittumisen epävarmuudesta. Kuvassa on osoitettu oudokki 1822.[162] □

Ennustustarkkuutta kuvaavat kaavat edellyttävät, että malli on määritelty oikein. Monesti malliin liittyy epävarmuutta, jota kaavat eivät huomioi. Empiiri-



Kuva 13.13: Jäidenlähtöpäivän Kokemäenjoessa odotusarvon 95 %:n luottamusväli ja uuden havainnon ennusteen 95 %:n ennusteväli.

sessä ennustamisessa ennustevirheen varianssi voi olla siksi kaavojen ilmaisemaa suurempi.

13.9 Mallin valinta

Sovellusalan teorian tai aiempien empiiristen tutkimusten perusteella voi olla muodostunut vahva näkemys, että tietyt selittäjät kuuluvat malliin. Yleensä kaikki selittäjät eivät kuitenkaan ole tutkimusta aloitettaessa täysin tiedossa — ainakaan monilla tieteenaloilla kuten yhteiskunta- ja käyttäytymistieteissä. Aiemmin tutkimatonta ilmiötä mallitettaessa voi olla hyvin epäselvää, mitä selittäjiä tulisi malliin sisällyttää.

Perinteinen näkemys tieteessä on, että yksinkertaisin selitys on paras. Tilastollisen mallin yhteydessä puhutaan vastaavasti säästäväisyysperiaatteesta (*principle of parsimony*): sitä parempi, mitä vähemmän parametreja. Regressio-

malliin tulisi siis valita mieluummin vähemmän kuin enemmän selittäjiä. Tällöin havaintoja on enemmän kutakin parametria kohti, ja ne tullevat estimoitua tarkemmin.

Tutkimusten tulisi olla toistettavissa, jotta tulosten luotettavuus ja yleistettävyys voitaisiin varmistaa. Sitä edesauttaa, että selittäjien ja mallin valinta on toteutettu ymmärrettävällä tavalla, jonka muut tutkijat voivat toisintaa.

Yleisimmät mallin valinta -menetelmät perustuvat testaukseen. Tässä esitetään yksi sellainen.

Poistovalinta (*backward elimination*) on paljon käytetty tilastotieteellinen menettely selittäjien valitsemiseksi. Malliin pyritään sisällyttämään kaikki vasteeseen mahdollisesti yhteydessä olevat selittäjät. Mallista poistetaan yksi kerrallaan tarpeettomalta vaikuttava selittäjä, kunnes kaikki vaikuttavat tarpeellisilta. Tyypillinen poistokriteeri on, että selittäjän kertoimen t-arvon p-arvo ylittää etukäteen asetetun rajan. Selittäjä, johon liittyvä p-arvo ylittää rajan eniten, poistetaan mallista, ja malli estimoidaan uudellaan. Mallista poistetaan jälleen vähiten merkitykselliseltä vaikuttava selittäjä. Näin jatketaan, kunnes suurin p-arvo alittaa poistorajan. Näin löydetty malli valitaan raportoitavaksi ja sovellettavaksi.

Poistovalinta-algoritmia voi sopeuttaa tarpeen mukaan. Osa selittäjistä saatetaan arvioida välttämättömiksi, ja ne voidaan sisällyttää malliin, vaikka poistovalintakriteeri edellyttäisi selittäjän poistamista. Tällöin poistetaan selittäjä, jonka t-arvon p-arvo ylittää poistokriteerin ja on suurin muiden selittäjien joukossa. Mallin sisältäessä luokkamuuttujia ja niiden yhdysvaikutuksia, päävaikutusosoittimet tavataan poistaa mallista vasta niiden yhdysvaikutusosoittimen poistamisen jälkeen. Poistovalinta-algoritmia voidaan soveltaa tämäkin huomioiden.

Poistovalinnan yksi etu on, että mallin valinta on periaatteessa toistettavissa ja tarkistettavissa. Haitta on, että selittäjien valinta p-arvojen perusteella tapaa johtaa liioiteltuun kuvaan selittäjien merkityksestä (esim. Weisberg 2014, 245). Vaikka selittäjistä yksikään ei olisi yhteydessä vasteeseen, mutta testejä tehdään useita, jotkin p-arvot voivat alittaa poistorajan.

Mallia ei ole välttämätöntä muodostaa algoritmisesti. Mallin valinnassa voi olla luova, ja uutta voi löytää poikkeamalla vanhoista käytännöistä. Luova voi olla myös selittäjien muodostamisessa (jakso 13.11.4).

Valitun mallin kelpoisuutta tulisi koeponnistaa jäännöstarkasteluilla ja muulla regressiodiagnostiikalla. Mallin valintaa, muuttujien muuntamista ja mallien diagnostikkaa opastetaan Foxin (2016), Weisbergin (2014) sekä Foxin ja Weisbergin (2019) kirjoissa.

13.10 Satunnaismuuttujaselittäjät

Edellä oletettiin, että regression selittäjät ovat kiinteitä. Tarkkaavainen lukija saattoi huomata, että kuvatuissa empiirisissä esimerkeissä voi olla luontevampaa ajatella selittäjät satunnaisiksi. Useimmissa yhteiskunta- ja käyttäytymistieteellisissä ja monissa muissa yhteyksissä kiinteiden selittäjien oletus ei päde. Punottu teoria pätee tietyin tarkennuksin, vaikka selittäjät olisivat satunnaismuuttujia. Keskeistä on, että satunnaistermi ja selittäjät ovat riippumattomia. Lisäksi selittäjien satunnaisuuden tulee olla sopivanlaista.

Jaksossa laajennetaan ensin teorian soveltuvuutta satunnaisten selittäjien tilanteeseen. Seuraavaksi varoitetaan kolmesta tilanteesta, jossa satunnainen selittäjä vääristää PNS-estimoinnin. Asiat esitetään yhden selittäjän kehikossa, mutta ne yleistyvät monen selittäjän tilanteeseen.

Varoitusjaksoista voi siitä kysymys, kannattaako regressioanalyysia käyttää, jos tulokset voivat olla epäluotettavia tai jopa järjettömiä. Vastaus on kyllä. Regressioanalyysia — kuten ylipäänsäkään tilastotiedettä tai mitään muuta tiedettä — ei vain tule soveltaa huolimattomasti ja ajattelemattomasti.

13.10.1 Satunnaismuuttujaselittäjä ja regression satunnaistermi

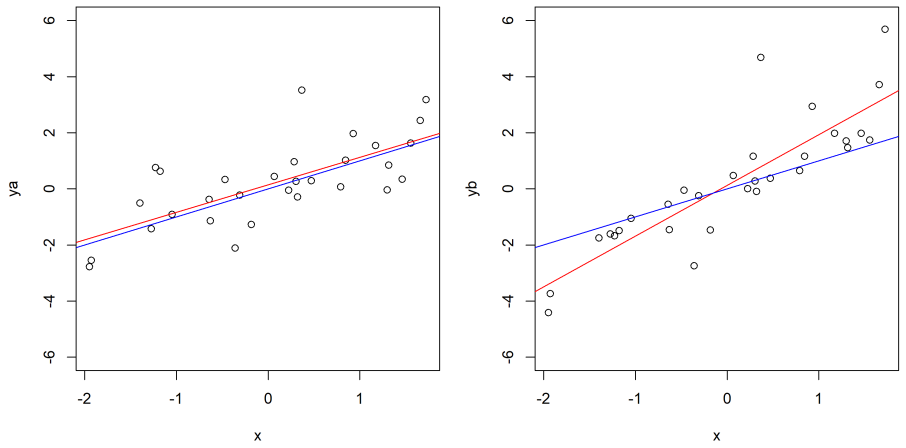
Opittua PNS-estimointi- ja jakaumateoriaa voidaan soveltaa, vaikka selittäjä olisi satunnaismuuttuja. Edellytys on, että regression satunnaistermi $\varepsilon_1, \ldots, \varepsilon_n$ ($\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$, $i = 1, \ldots, n$) on riippumaton selittäjästä $X_1, \ldots, X_n$. Analyysien pätevyys seuraa niiden ehdollistamisesta toteumille $x_1, \ldots, x_n$. Ehdollistamisen jälkeen toteumat rinnastuvat havaintoihin ei-satunnaisesta selittäjästä. Yksityiskohdat sivuutetaan. Jos satunnaistermi ei ole riippumaton selittäjästä, analyysit vinoutuvat.

PNS-estimoinnin ja jakaumateorian toimivuus satunnaismuuttujaselittäjien tilanteessa perustellaan tarkemmin Baltagin (2011, 53–54 ja 96–97), Hamiltonin (1994, 207–208) ja Johnstonin (1984, 281–285) oppikirjoissa.

Esimerkki. Käskyllä `rnorm(30)` arvottiin 30 (x) + 30 (ε) riippumatonta havaintoa standardinormaalijakaumasta. Niistä muodostettiin vasteet $ya = x + \varepsilon$ ja $yb = x + (\varepsilon + x \times |\varepsilon|)$ (notaatio ei erottele satunnaistermiä ja sen toteumaa).

Kuvassa 13.14 on vasemmalla x:n ja ya:n sirontakuvio, aineistoon PNS-menetelmällä sovitettu punainen regressiossuora ($\hat{\beta}_0 + \hat{\beta}_1 x$) ja todellisen yhteyden piirtävä sininen suora. Satunnaistermin toteumat siroutuvat tasaisesti suorien ylä- ja alapuolelle, koska satunnaistermi on riippumaton selittäjästä. Regressiosuora on estimoitu harvinaisen onnistuneesti. Se, että selittäjä on satunnaismuuttuja, ei ole este regressiomallin soveltamiselle.

Oikeassa puoliskossa kuvaa 13.14 sirontakuvio on muodostettu samoista satunnaismuuttujien toteumista, mutta vaste on yb. Nyt regression satunnaistermi $\varepsilon + x \times |\varepsilon|$ korreloi selittäjän x kanssa: Selittäjä ja satunnaistermi pyrkivät saamaan yhtä aikaa joko negatiivisen tai positiivisen arvon. Kuviossa se näkyy satunnaistermien siroutumisena sinisen suoran ala/-yläpuolelle negatiivisilla/positiivisilla x:n arvoilla. Regressiosuora estimoituu virheellisesti. Satunnaistermin ja selittäjän korrelaatio vääristää estimoinnin. □



Kuva 13.14: Vasen: Regressio, jossa selittäjä ja satunnaistermi ovat riippumattomia. Oikea: Regressio, jossa selittäjä ja satunnaistermi eivät ole riippumattomia. Molemmat: Sininen suora on todellinen yhteys.

13.10.2 Aikasarjamallit

Lukuisissa sovelluksissa tutkitaan *aikasarjoja* (*time series*) regressiomallilla. Aikasarja koostuu ajankohdan mukaan järjestetyistä havainnoista. Esimerkkejä

ovat syntyvyys, osakekurssit ja monet talouden ja hyvinvoinnin indikaattorit. Niiden arvo tiettynä ajankohtana riippuu läheisten ajankohtien arvoista: Jos väestön lukumäärä on suuri, on se suuri viereisinäkin ajankohtina. Aikasarjat eivät ole täysin ennustettavia, eli ne ovat satunnaisia. Jakson 13.10.1 ehto regression satunnaistermin ja selittäjien riippumattomuudesta ei ylipäänsä päde aikasarjoja analysoitaessa. Regressioanalyysia voidaan soveltaa aikasarjojen tutkimiseen mutta lisähuomioin ja -ehdoin.

Yksinkertaisin aikasarjamalli on 1. asteen autoregressio:

$$Y_t = \alpha Y_{t-1} + \varepsilon_t. \tag{13.27}$$

Siinä α on regressiokerroin. Autoregressio viittaa regressioon itsensä kanssa. Mallin mukaan satunnaismuuttujan Y arvo ajankohdalla t (vaikkapa 2021), määräytyy sen arvon edellisellä ajankohdalla (2020) perustella. Satunnaistermi $\varepsilon_t \sim \mathsf{N}(0, \sigma^2)$, $t = 1, \ldots, n$ ja on riippumaton satunnaistermeistä toisina ajankohtina sekä Y_{t-1}:stä. Sijoittamalla $Y_i = \alpha Y_{i-1} + \varepsilon_i$ toistuvasti kaavaan (13.27) huomataan, että vaste Y_t on painotettu summa menneistä satunnaistermeistä sekä alkuarvosta Y_0:

$$Y_t = \alpha Y_{t-1} + \varepsilon_t = \alpha^2 Y_{t-2} + \varepsilon_t + \alpha\varepsilon_{t-1} = \ldots = \sum_{i=0}^{t-1} \alpha^i \varepsilon_{t-i} + \alpha^t Y_0.$$

Selittäjä $Y_1, \ldots, Y_{t-1}$ riippuu vahvasti — suorastaan koostuu — satunnaistermeistä menneinä ajankohtina $\varepsilon_1, \ldots, \varepsilon_{t-1}$!

Voiko mallin (13.27) estimoida PNS:llä ja soveltaa jakaumateoriaa edellä opittuun tapaan? Kyllä ja ei. Selittäjän Y_{t-1} korrelaation satunnaistermin menneiden arvojen kanssa takia PNS-estimaattori $\hat{\alpha}$ on harhainen. Tilanteen pelastaa se, että satunnaistermi ε_t ei korreloi selittäjän Y_{t-1} kanssa. Voidaan osoittaa, että PNS-estimaattori on tarkentuva ja että mikäli $|\alpha| < 1$, niin suurilla havaintomäärillä tavanomainen jakaumateoria pätee. Mallia voidaan käyttää pitäen mielessä estimaattorin harhaisuus ja jakaumateorian toimimattomuus pienillä havaintomäärillä.

Dynaamisessa regressiomallissa on sekä vasteen että muun selittäjän viipeitä selittäjinä. Yksinkertainen dynaaminen regressiomalli on

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$$

ilmeisin merkinnöin ($|\alpha_1| < 1$). Mallilla voidaan selittää vaikkapa asuntojen hintaa niiden hinnalla edellisellä ajanjaksolla sekä muulla asuntojen hintaan

vaikuttavalla tekijällä. Mallin (13.27) tapaan selittäjä Y_{t-1} riippuu aiemmista satunnaistermeistä muttei samanhetkisestä satunnaistermistä. Jälleen PNS-estimaattorit ovat harhaisia mutta tarkentuvia ja opittu jakaumateoria pätee suurilla havaintomäärillä.

Aikasarjamallit laajentavat PNS-estimaattorien jakaumateorian soveltamismahdollisuuksia tavattomasti. Luotettava estimointi ja tilastollinen päättely edellyttävät kuitenkin enemmän havaintoja kuin silloin, kun selittäjät ovat kiinteitä tai muuten satunnaistermistä täysin riippumattomia.

13.10.3 Varoitus I: Trendimäiset aikasarjat

Jo Hooker (1901) ja Yule (1926) varoittivat, että trendimäisten aikasarjojen väliset korrelaatiot voivat olla suuria vaikkei niillä ole todellista yhteyttä. Yule todensi simuloimalla(!), että kahden toisistaan riippumattoman trendimäisen aikasarjan otoskorrelaation jakauma voi olla jopa U:n muotoinen, jolloin korrelaatio on todennäköisemmin lähellä ± 1:htä kuin 0:aa. Tulos selittyy sillä, että yhtälössä (13.1) satunnaistermi ei toteuta PNS-estimoinnissa tehtyjä oletuksia, jos $\beta_1 = 0$ ja selitettävä ja selittävä muuttuja ovat trendimäisiä. Tällöin satunnaistermi muodostuu itse trendimäiseksi ja korreloituneeksi vierekkäisinä ajankohtina. Tällaisiin *hölynpölyregressioihin* (*nonsense regression*; aikasarjaekonometriassa myös *spurious regression*) liittyviä korrelaatioita kutsutaan *hölynpölykorrelaatioiksi* (*nonsense correlation*; aikasarjaekonometriassa myös *spurious correlation*), koska ne eivät heijasta todellista yhteyttä muuttujien välillä. Yleisemmin puhutaan *näennäiskorrelaatioista* (*spurious correlation*), kun muuttujat korreloivat otoksessa mutta niillä ei ole mielekästä todellista yhteyttä.

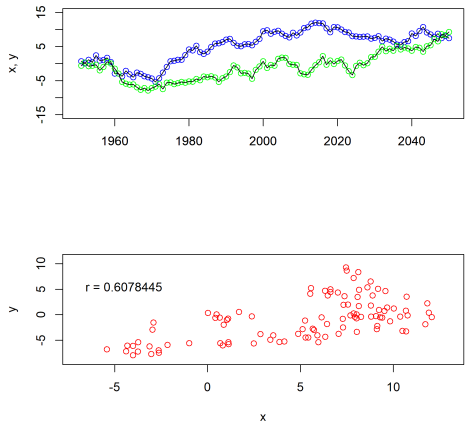
Esimerkki. Tuotetaan R:llä kaksi toisistaan riippumatonta sadan havainnon otosta standardinormaalijakaumasta. Merkitään näitä satunnaismuuttujia ε_{yt}:llä ja ε_{xt}:llä, joissa $t = 1950, \ldots, 2050$ (kuvitteellinen vuosisadan ajanjakso). Muodostetaan niistä kaksi aikasarjaa

$$y_t = y_{t-1} + \varepsilon_{yt} \quad \text{ja} \quad x_t = x_{t-1} + \varepsilon_{xt}$$

($y_{1949} = x_{1949} = 0$). Kumpikin aikasarja noudattaa mallia (13.27), jossa $\alpha = 1$. Aikasarjat on piirretty kuvassa 13.15. Kuvan yläosassa aikasarjat on piirretty aikaa vastaan; alaosassa toisiaan vastaan. Muuttujien otoskorrelaatio on 0.608, vaikka ne ovat todellisuudessa riippumattomia.

Granger ja Newbold (1974) tekivät simulointikokeita tällaisista muuttujista 50 havainnon otoksilla. He havaitsivat, että niitä regressoitaessa yhtälöstä (13.1)

PNS:llä estimoidun β_1-kertoimen t-arvon (13.9) jakauman 0.025. tai 0.975. kvantiilit ovat (mainitulla otoskoolla) noin ± 11.2 tavallisen jakaumateorian mukaisen noin ± 2:n sijaan. Jos käytettäisiin jälkimmäisiä kriittisiä arvoja, testin koko olisi noin 75 % tarkoitetun 5 %:n sijaan. Hendry (1995) raportoi vastaavia simulointeja. Phillips (1986) osoitti, että havaintojen lukumäärän kasvaessa PNS-estimaattorin t-arvon itseisarvo kasvaa niin, että nollahypoteesi tulee varmasti hylätyksi eikä regression selitysaste R^2 mene nollaan. □



Kuva 13.15: Kaksi riippumatonta satunnaiskulkua. $x =$ sininen, $y =$ vihreä.

Esimerkki. 30 000 näennäiskorrelaatiota. Tyler Vigen dokumentoi nettisivullaan `http://www.tylervigen.com/` (haettu 9.3.2020) ja kirjassaan[163] 30 000 järjetöntä empiiristä korrelaatiota ja regressiota. Vigen löysi ne laatimallaan algoritmilla, joka haki suurista tietokannoista aineistoja ja laski niistä itseisarvoltaan mahdollisimman suuria korrelaatioita.

Kuvissa 13.16 ja 13.17 esitetään kaksi Vigenin esimerkkiä Yhdysvalloista: Tutkimusmenojen luonnontieteisiin ja tekniikkaan ja itsemurhien korrelaatio on 0.99 ajanjaksolla 1999–2009. Hunajaa tuottavien mehiläisyhdyskuntien ja nuorten pidätysten marihuanan hallussapidosta korrelaatio on -0.93 ajanjaksolla 1990–2009.

Vigenin löytämät näennäiskorrelaatiot selittyvät kahdella seikalla: 1) Monet

hänen louhimistaan aikasarjoista ovat trendimäisiä. Trendimäisyydellä tarkoitetaan tässä karkeasti sitä, että tietyllä aikavälillä (mahdollisesti koko tarkasteluajanjaksolla) aikasarja vaikuttaa kehittyvän poispäin alkuarvostaan. Hölynpölykorrelaation riski on tällöin suuri. 2) Mikä tahansa yksin merkitsevyystestaukseen perustuva tilastollinen menettely tuottaa hylkäysvirheitä valitun merkitsevyystason mukaisesti. Korrelaatioita — suuriakin — väistämättä löytyy, jos testaa tarpeeksi. □

Hölynpölyregression voi paljastaa data-analyysilla tutkimalla mallin jäännöstä. Mikäli se korreloi vierekkäisten havaintojensa kanssa, todetaan, että regressiomallin soveltamisen edellytykset eivät täyty ja hylätään estimaatit, testit ja malli pätemättöminä. Mahdollinen seuraava askel on perehtyä kehittyneempiin aikasarja-analyysin menetelmiin ja oppikirjoihin. Yksinkertainen ratkaisu voi olla poistaa trendit laskemalla aikasarjojen muutokset $(y_t - y_{t-1})$:t ja $(x_t - x_{t-1})$:t ja mallittaa muutoksia.

13.10.4 Varoitus II: Mittausvirheet

Toinen ongelmatilanne syntyy, kun selittäjien satunnaisuus johtuu mittausvirheistä. Tällöin regressiokertoimen PNS-estimaattori ei ole harhaton eikä tarkentuva.

Määräytyköön satunnaismuuttuja Y yhtälöstä

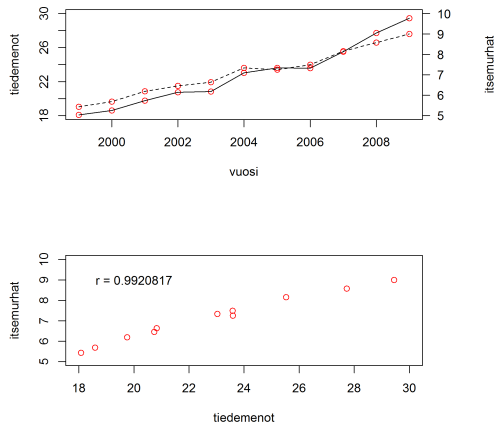
$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

Selittäjä x ei ole havaittavissa mutta mittausvirheellä δ suolattu $X^* = x + \delta$ on. Havainnoittaiset mittausvirheet δ_i ovat riippumattomia toisistaan, selittäjistä x_i ja satunnaistermeistä ε_i. Mittausvirheiden odotusarvo ja varianssi ovat $\mathsf{E}(\delta) = 0$ ja $\mathsf{V}(\delta) = \sigma_\delta^2 > 0$. Mittausvirheiden myötä mallin yllä estimointi muuntuu mallin

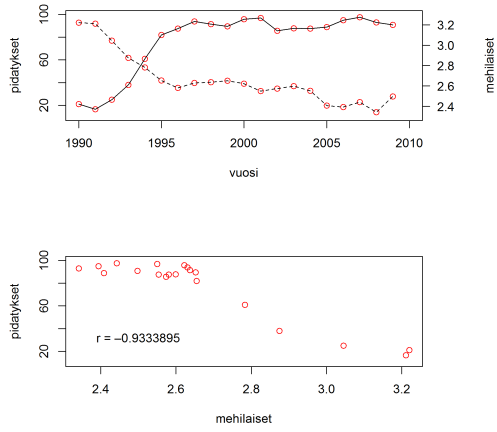
$$Y = \beta_0 + \beta_1(X^* - \delta) + \varepsilon = \beta_0 + \beta_1 X^* + (\varepsilon - \beta_1\delta)$$

estimoinniksi. Selittäjä $X^* = x + \delta$ ja satunnaistermi $\varepsilon - \beta_1\delta$ korreloivat. Jos mittausvirhe δ ja regressiokerroin β_1 ovat positiivisia, niin δ suurentaa selittäjää ja pienentää satunnaistermiä. Selittäjän X^* suureneminen näyttää kasvattavan vastetta vähemmän kuin x:n suureneminen todellisuudessa kasvattaa vastetta. Regressiokertoimen $\hat{\beta}_1$ PNS-estimaattori on tällöin harhainen kohti nollaa $(\mathsf{E}(|\hat{\beta}_1|) < |\beta_1|)$. Korrelaatio säilyy havaintojen lukumäärän kasvaessa, joten PNS-estimaattori on tarkentumaton. Havaintomäärän kasvaessa kohti ääretöntä $\hat{\beta}_1$ on noin

$$\frac{\beta_1}{1 + \sigma_\delta^2/\sigma_x^2} \tag{13.28}$$



Kuva 13.16: Tiedemenot = — (miljardia inflaatiosta puhdistettua dollaria), itsemurhat = - - (tuhatta itsemurhaa).



Kuva 13.17: Pidätykset = — (tuhatta pidätystä), mehiläiset = - - (tuhatta yhdyskuntaa).

(esim. Johnston 1984, 430). Yllä σ_x^2 on luku, johon x_1:n otosvarianssin $\sum_{i=1}^{n}(x-\bar{x})^2$:n oletetaan suppenevan suurilla havaintomäärillä. Kaavasta nähdään intuitiivinen tulos, että suurilla havaintomäärillä $\hat{\beta}_1$ on altis eroamaan oikeasta arvosta β_1 sitä enemmän, mitä suurempi on mittausvirheen varianssi σ_δ^2 suhteessa x:n vaihteluun σ_x^2.

Esimerkki. Olkoon havaintoja paljon. Toivemaassa $\sigma_\delta^2/\sigma_x^2 = 0$, mittausvirhettä ei ole ja $\hat{\beta}_1$ on noin β_1. Jos $\sigma_\delta^2/\sigma_x^2 = a$, mittausvirheen varianssi σ_δ^2 on $100 \times a$ % σ_x^2:sta ja $\hat{\beta}_1$ on noin $\beta_1/(1+a)$. Kauhuelokuvassa $\sigma_\delta^2/\sigma_x^2 = 1$, jolloin $\hat{\beta}_1$ on noin $\beta_1/2$. □

Monen selittäjän regressiossa selittäjien mittausvirheistä ei seuraa, että kaikkien regressiokertoimien PNS-estimaatit olisivat harhaisia nollaa kohti. Samantapainen tulos pätee silti monen selittäjän tilanteessa (Ruud 2000, 530, vrt. Fox 2016, 120–123, Johnston 1984, 428–430).

Mittausvirhe vasteessa ei ole samallalailla ongelma:

$$Y + \delta = \beta_0 + \beta_1 x + \varepsilon \quad \Leftrightarrow \quad Y = \beta_0 + \beta_1 x + (\varepsilon - \delta).$$

Mittausvirhe δ on havainnoittain riippumaton, riippumaton selittäjistä ja vasteesta, $\mathsf{E}(\delta) = 0$ ja $\mathsf{V}(\delta) = \sigma_\delta^2 > 0$. Vasteen mittausvirhe voidaan tulkita osaksi regressiomallin satunnaistermiä. Uuden satunnaistermin varianssi on $\mathsf{V}(\varepsilon - \delta) = \sigma^2 + \sigma_\delta^2 > \sigma^2$. Kaikki opittu PNS-estimaattorin harhattomuudesta, tarkentuvuudesta jne. pätee. Ilman vasteen mittausvirhettä satunnaistermin varianssi olisi toki pienempi ja estimointi tarkempaa ja tilastollinen päättely napakampaa.

Esimerkki. Käsien pituus. Selitetään mallilla

$$y = \beta_0 + \beta_1 x + \varepsilon$$

vasemman käden pituutta oikean käden pituudella ja oikean käden pituutta vasemman käden pituudella.[164] Aineisto on simuloitu `rnorm`-käskyllä. Aineiston lähtökohta on peruspituus $\mathsf{N}(45, 4)$-jakaumasta. Kuhunkin peruspituuteen on lisätty havainto $\mathsf{N}(0, 0.25)$-jakaumasta. Saatu luku on havainto vasemman käden pituudesta. Samaan peruspituuteen on lisätty seuraavaksi toinen riippumattomasti arvottu havainto $\mathsf{N}(0, 0.25)$-jakaumasta. Näin on saatu havainto oikean käden pituudesta. Toistamalla edellä kuvatut kolme vaihetta riippumattomasti 200 kertaa on tuotettu aineisto käsiparin pituuksista ($n = 200$). Kummankin käden mitattu pituus noudattaa $\mathsf{N}(45, 4.25)$-jakaumaa. Idea on, että käsiparien peruspituus noudattaa $\mathsf{N}(45, 4)$-jakaumaa mutta kunkin käsiparin käsien

pituuksissa on toisistaan riippumatonta $\mathsf{N}(0, 0.25)$-jakautunutta lisäsatunnaisvaihtelua kuten mittausvirhettä.

Selitetään ensin käden pituutta toisen käden peruspituudella. Selitettävän käden pituudessa on mittausvirhettä, mutta selittäjäkäden peruspituudessa ei ole. Kuvan 13.18 yläkerrassa on sirontakuviot tällaisista regressioista. Punainen regressiosuora on saatu selittämällä joko vasemman käden pituutta oikean käden peruspituudella tai oikean käden pituutta vasemman käden peruspituudella. Molemmissa kuvioissa sininen suora piirtää yhteyden "kädet ovat yhtäpitkiä". Regressiosuorat melkein yhtyvät sinisiin suoriin, ja regressiokertoimen PNS-estimaatit 1.021 ja 1.027 ovat lähes oikeat. Regressioiden satunnaistermit eivät ole minkäänlaisessa yhteydessä selittäjään, joten PNS-estimointi toimii oivasti, vaikka selittäjä on satunnaismuuttuja.

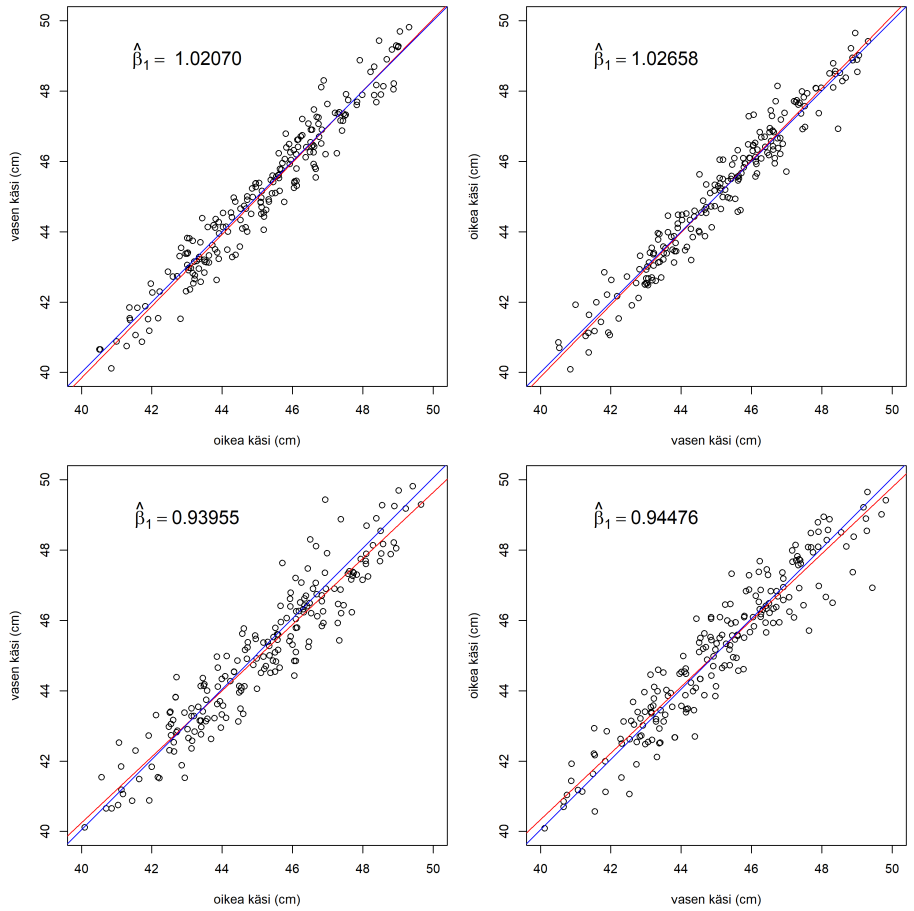
Kuvan 13.18 alakerrassa on sirontakuviot, joissa selitetään vasemman käden pituutta oikean käden pituudella ja päinvastoin. Myös selittäjässä on nyt mittausvirhettä. Kaavan (13.28) mukaan PNS-estimaatin tulisi olla suurinpiirtein $\beta_1/(1 + \sigma_\delta^2/\sigma_x^2) = 1/(1 + 0.25/4) \approx 0.941$. PNS-estimaatit ovat teorianmukaiset noin 0.940 ja 0.945. Asettamalla toistojen määräksi 100 000 (tuottamalla 100 000 käsiparin pituusmittausta), R laskee teorian ennustamat PNS-estimaatit noin 0.941 ja 0.941.

Alakerran kuviot voidaan tulkita myös esimerkkeinä regressiosta odotusarvoa kohti. Peruspituuden päälle tuleva $\mathsf{N}(0, 0.25)$-satunnaisvaihtelu voi olla mittausvirhettä, käsien pituuteen liittyvää muuta satunnaisvaihtelua tai yhdistelmää niistä. Regressiokertoimen β_1 PNS-estimaatti on

$$\hat{\beta}_1 = \hat{\rho}\frac{s_y}{s_x}$$

(kaava (13.5)). Nyt selitettävän ja selittäjän keskihajonnat ovat samat, joten suurilla havaintomäärillä $s_y \approx s_x$. Käsiparien käsien pituuksissa on satunnaista eroa, jolloin $\hat{\rho} < 1$. Näin ollen suurilla havaintomäärillä $\hat{\beta}_1 < 1$. Ennustettaessa toisen käden pituutta toisen käden pituudella, regressiokertoimen tulee olla yhtä pienempi! □

Oletus, että mittausvirhe ei korreloi selittäjän kanssa voi tuntua luontevalta muttei päde aina. Jos selittäjä on vastaus kyselytutkimuksen kysymykseen, johon ihmiset tietoisesti tai tiedostamattaan arvioivat tietynlaisen vastauksen positiiviseksi tai negatiiviseksi, selittäjä ja mittausvirhe ovat alttiita korreloimaan. Muulloinkin selittäjän ja mittausvirheen korrelaatio on mahdollinen. Mikäli mahdollisuus on ilmeinen, ratkaisua voi hakea kirjallisuudesta mittausvirhemalleista (*measurement error model*) tai arkaluonteisista kyselyistä (*sensitive survey*).



Kuva 13.18: Vasen: Vasemman käden pituuden selittäminen oikean käden pituudella. Oikea: Oikean käden pituuden selittäminen vasemman käden pituudella. Yläkerta: Mittausvirhettä vain selitettävässä. Alakerta: Mittausvirhettä sekä selitettävässä että selittäjässä.

13.10.5 Varoitus III: Endogeenisuus

On oleellista, että selitettävä ei vaikuta yhteenkään selittäjistä. Yhteiskunta- ja käyttäytymistieteissä on usein — mahdollisesti hyvin usein (Hamilton 1994,

235) — syytä ajatella, että muuttuja, jota haluttaisiin selittää, vaikuttaa itse regression selittäviin muuttujiin. Tällöin estimoidut regressiokertoimet eivät ole harhattomia eivätkä tarkentuvia. Intuitio on selvä: Selittäjän vaikutuksen selitettävään arviointi vaikeutuu, jos selitettävä muuttaa edelleen selittäjää, joka taas muuttaa selitettävää jne. Selittäjän sanotaan tällöin olevan *endogeeninen*.

Esimerkki. Endogeenisuus (teoreettinen pohdinta). Oletetaan, että selitettävä muuttuja Y vaikuttaa välittömästi selittävään muuttujaan regressiomallissa

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Selittävä muuttuja X on endogeeninen. Oletetaan lisäksi, että Y:n suureneminen pyrkii seuraamaan X:n suureneminen ja että $\beta_1 > 0$. Jos satunnaistermi ε on positiivinen, se suurentaa selitettävää muuttujaa Y, joka edelleen suurentaa selittävää muuttujaa X. Satunnaistermin ja selittävän muuttujan muutokset sekoittuvat. Tällöin β_1:n PNS-estimaattori ei ole tarkentuva eikä harhaton. Satunnaistermi ei saa korreloida selitettävän muuttujan kanssa estimoitaessa regressiomallia PNS:llä. □

Esimerkki. Endogeenisuus (empiirinen pohdinta). Tutkija arvioi, että yksilöiden tulot (X) selittävät heidän onnellisuuttaan (Y) ja estimoi PNS:llä yhtälön (13.1). Onnellisuus saattaa kuitenkin tehdä ihmisistä tuottavampia (masentuneet ovat enemmän poissa töistä). Tällöin X vaikuttaisi Y:hyn ja päinvastoin. PNS-estimaattorit eivät olisi harhattomia eivätkä tarkentuvia. □

Ennen regressioanalyysia selittäjän endogeenisuuden mahdollisuus tulee sulkea pois tai mahdollinen endogeenisuus arvioida mitättömäksi. Vaihtoehtoisesti regressioanalyysin tulokset tulee tulkita tilastollisina yhteyksinä liittämättä niihin kausaalisuutta ja ajatusta vaikutuksen suuruudesta, jota ei voi selvittää yhtälön (13.1) PNS-estimoinnilla selittäjän ollessa endogeeninen. (Havaitaan esimerkiksi, että tulot ja onnellisuus korreloivat muttei tehdä PNS-estimaatista johtopäätöstä tulojen vaikutuksen suuruudesta onnellisuuteen.) Mikäli ollaan kiinnostuneita endogeenisten selittävien muuttujien vaikutuksesta selitettävään muuttujaan, edetään opiskelemaan kehittyneempiä estimointitekniikoita esimerkiksi ekonometrian oppikirjoista.

13.11 Erityiskysymyksiä

13.11.1 Regressio origon kautta

Kuvassa 13.19 vasemmalla ylhäällä on sirontakuvio köyhyyden muutoksesta ja talouskasvusta 31 kehitysmaassa vuosina 1987–1999.[165] Köyhyydenmuutosta mitataan %-yksikkömuutoksella väestönosuudessa, joka elää dollarilla tai alle päivässä. Talouskasvun mittari on bruttokansantuotteen henkeä kohti keskimääräinen kasvuvauhti (%). Punainen suora on regressiosuora regressiosta (13.1) (vakiotermi ja selittäjä). Kuvassa on raportoitu talouskasvun regressiokertoimen PNS-estimaatti -3.703. Talouden kasvaessa %:n köyhyys vähenee 3.7 %-yksikköä.

Joskus on luontevaa ajatella, että regressiosuoran tulisi kulkea origon kautta eli että regressiomalli olisi

$$Y = \beta x + \varepsilon. \tag{13.29}$$

Köyhyyden muutosta mallitettaessa vakiotermitön malli voi tuntua teoreettisesti sopivalta: Jollei ole talouskasvua, ei ole köyhyyden vähenemistäkään. Myös aineisto viestii, että vakiotermi ei ehkä ole tarpeen. Regressiosuora kulkee lähes origon kautta.

R-käsky `lm(y~0+x)` estimoi vakiotermittömän mallin (13.29). Kuvan 13.19 oikeaan yläkulmaan on piirretty näin saatu regressiosuora ja kirjattu edellisestä regressiokertoimen PNS-estimaatista vain vähän eroava PNS-estimaatti -3.733. Regressiosuora lävistää nyt origon. Estimaatti, suora ja sovitteet ovat hyvin samanlaiset kuin kuviossa ylävasemmalla. Näin saatuaa mallia voi pitää teoreettisesti tyydyttävämpänä kuin vakiotermin sisältävää mallia. Lisäetuna voi ajatella, että tilastotieteessä pääsääntöisesti estimointi tarkentuu, kun malliin tuodaan lisäinformaatiota. Tässä malliin ujutettiin näkemys, että vakiotermiä ei tarvita. Voi perustellusti ajatella, että regressiokerroin on lisätiedon avulla saatu estimoitua hieman tarkemmin.

Jatketaan saman aineiston analyysia, mutta ajatellaan, että vaste on lämpötila celsiusasteissa (c) ja selittäjä lämpötilaan vaikuttava tekijä (x). Tutkitaan vaikkapa kuinka pakastimen lämpötila riippuu sen säätimen asennosta.

Estimoidaan malli vakiotermin kanssa, ja tuotetaan (akselimerkintöjä lukuunottamatta) kuvan 13.19 vasemman yläkulman kuvio uudestaan. Pakastimen amerikkalaista valmistajaa varten muutetaan lämpötilahavainnot fahrenheitasteiksi ($f = 32 + (9/5)c$), ja toistetaan PNS-estimointi. Saadaan sirontakuvio kuvan 13.19 keskivälissä vasemmalla ja regressiokertoimen PNS-estimaatti -6.665. Muuttunut kerroin johtuu lämpötila-asteikon skaalaamisesta.

Regressiot kuvan 13.19 vasemmassa yläkulmassa ja sen alla näyttävät mittayksikön muutosta vaille yhteneviltä ja ovat sitä. Uuden mallin mukaan säätimen muutos yksiköllä laskee pakastimen lämpötilaa 6.665:llä fahrenheitasteella eli $(5/9)\times 6.665 = 3.703$ celsiusasteella alkuperäisen estimoinnin mukaisesti. (Myöskään selitysaste tai regressiokertoimen t-arvo ei muutu.)

Kuvio keskivälissä oikealla (kuva 13.19) havainnollistaa regressiota, kun vaste on fahrenheitasteita ja estimoidaan malli ilman vakiotermiä. Regressiosuora on pakotettu menemään origon kautta, jolloin se ei voi mitenkään kuvata fahrenheitlämpötilan keskimääräistä käyttäytymistä. Regressiokertoimen PNS-estimaatti -7.801 eroaa sekä numeerisesti että sisällöllisesti aiemmista. Regressiosuora on livennyt sivuun havaintopilvestä ja antaa systemaattisesti väärän kuvan lämpötilan ja sen säätimen asennon yhteydestä.

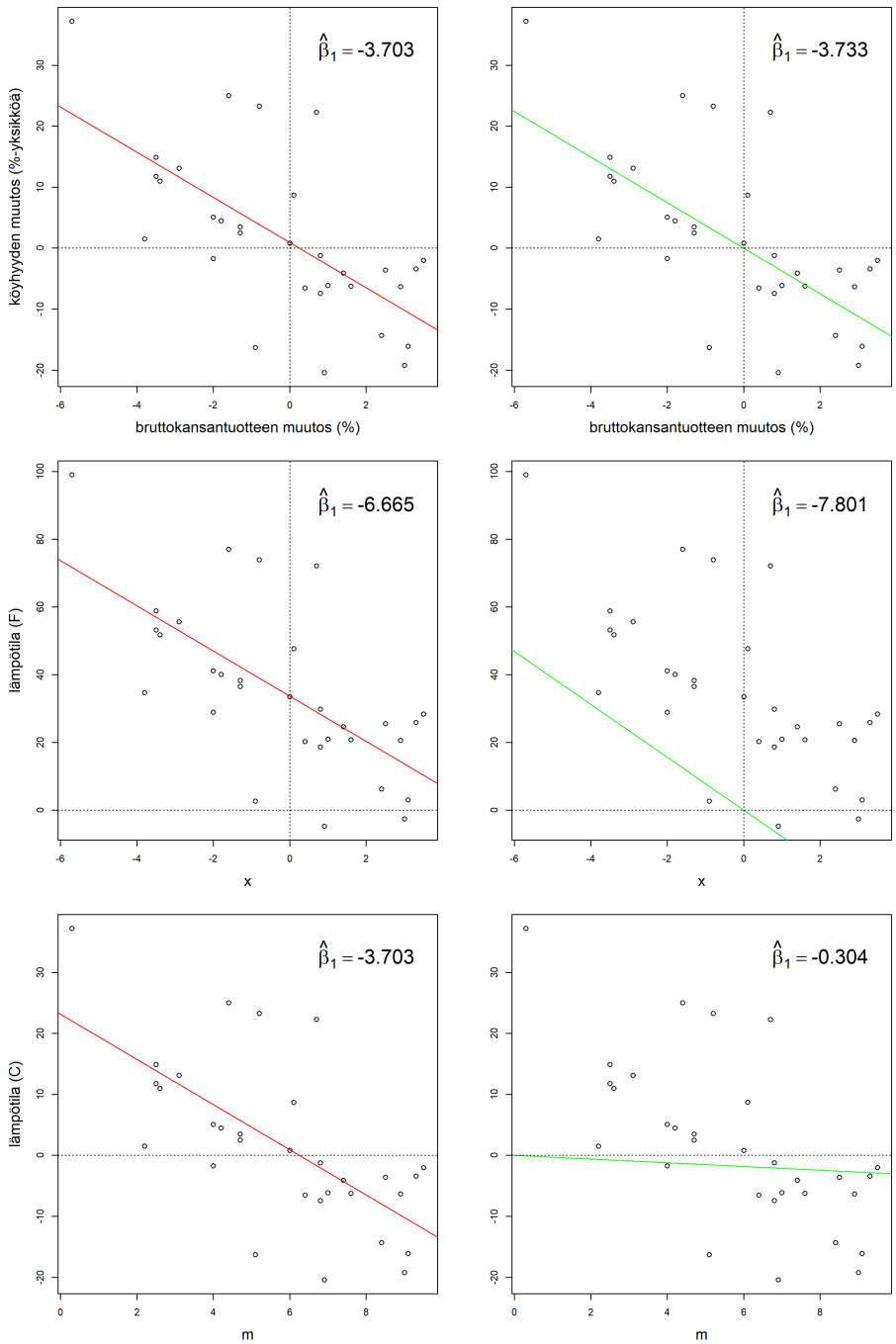
Mitä tapahtuu, jos pakastintutkija päättääkin vasteen sijaan näpelöidä selittäjää? Hän pitää $0-10$:ntä luontevampana lämpötilasäätimen asteikkona kuin alkuperäistä kuvioiden $-6-4$:ää. Hän summaa selittäjähavaintoihin 6, ja selittää celsiuslämpötiloja uusilla muunnetuilla säädinarvoilla (m). Kuvan 13.19 vasemman alakulman mukaan ei tapahdu juuri mitään, jos mallissa on vakiotermi. Havainnot ja regressiosuora vain siirtyvät oikealle 6:n verran. Jos mallissa ei ole vakiotermiä, regressiosuora hapuilee kummallisesti kuvan 13.19 oikeassa alakulmassa ja regressiokertoimen PNS-estimaatti on -0.304. Sen mukaan lämpötila ei juurikaan riippuisi lämpötilasäätimen asennosta, mikä ei selvästikään pidä paikkaansa.

Vakiotermi takaa, että PNS-estimoinnin tulos on yhtäpitävä riippumatta siitä, onko vaste y vai sen lineaarimuunnos $a+by$ tai selittäjä x vai sen lineaarimuunnos $a+bx$. Edellä $a \neq 0$ ja $b \neq 0$ ovat kiinteitä lukuja. Harjoitustehtävässä ratkotaan, mitä tapahtuu, jos vaste kerrotaan vakiolla, kun mallissa ei ole vakiotermiä.

Vakiotermiä ei pidä poistaa mallista vain siksi, että se vaikuttaa empiirisesti merkityksettömältä. Vakiotermin voi poistaa, jos se on teoreettisesti vahvasti perusteltua ja on selvää, että selitettävällä muuttujalla ja selittäjällä on yksikäsitteinen nollapiste, jota ei voi mielekkäästi siirtää. Yleensä vakiotermi kannattaa sisällyttää malliin.

13.11.2 Selittäjien tärkeyden vertaaminen

Mikä on tärkein selittäjä? Kuinka paljon selittäjien itsenäisten muutosten vaikutukset selitettävään muuttujaan eroavat? Tällaiset kysymykset kiinnostavat. Regressioanalyysistä saattaa saada vastauksen — ja mahdollisesti hyvin selkeän, jos nollahypoteesia selittäjän tarpeettomuudesta ei hylätä. Vastausta haetaan



Kuva 13.19: Vasen: Regressio vakiotermin kanssa. Oikea: Regressio ilman vakiotermiä.

joskus vertaamalla estimoitujen kertoimien suuruutta. Se saattaa erehdyttää ja oivaltavinta voi olla pidättäytyä lopulliseen malliin sisällytettyjen selittäjien tärkeyden vertaamisesta. Estimoitujen kertoimien vertailu voi olla järkevää, jos selittäjät on mitattu samoissa mittayksiköissä.

Esimerkki. Siivoojien tuntipalkat II. Siivoojan palkkaa selitettiin yhtälössä (13.16) iällä ja työsuhteen kestolla. Molemmat on mitattu vuosissa. Ikämuuttujan estimoidun kertoimen -0.001 itseisarvo on alle kolmassadasosa työsuhteen pituus -muuttujan estimoidusta kertoimesta 0.339. Työsuhteen pituuden kasvu vuodella kasvattaa tavattomasti enemmän palkkaa kuin iän kasvu vuodella sitä vähentää. Vaikka työntekijöiden ikä vaihtelisi paljon enemmän kuin työsuhteen kesto, kertoimien ero on niin suuri, että on selvää, että työsuhteen kesto on tärkeämpi siivoojan palkan selittäjä kuin siivoojan ikä. □

Jos selittäjät on mitattu eri mittayksiköissä, selittäjät ja estimoidut kertoimet saatetaan voida muuntaa vertailukelpoisiksi. Yksinkertaisimmillaan selittäjän x_j mittayksikkö yhdenmukaistetaan toisen selittäjän kanssa jakamalla havainnot x_j:stä vakiolla a. Tällöin x_j:n PNS-estimaatti muuttuu $a\hat{\beta}_j$:ksi ($\hat{\beta}_j x_j = a\hat{\beta}_j(x_j/a)$). Jos mittayksiköt eivät ole yhdenmukaistettavissa, selittäjien kertoimet pyritään joskus saamaan vertailukelpoisiksi jakamalla kukin selittäjä otoskeskihajonnallaan (s_{x_j}). Standardoitu kerroin $s_{x_j}\hat{\beta}_j$ kuvaa tällöin muutosta selitettävässä, kun x_j/s_{x_j}-selittäjä muuttuu yksiköllä eli x_j:n keskihajonnan verran. Standardoitujen kertoimien vertailu saattaa olla mielekästä muttei ole välttämättä.

Esimerkki. Mittayksikön vaikutus regressiokertoimeen. Jakson 13.4.1 isä-poika-regressiossa isien pituudet oli kirjattu senttimetreissä. Isän pituus -muuttujan kertoimen PNS-estimaatti oli 0.531. Jaetaan isien pituudet 100:lla eli ilmaistaan pituudet metreissä. (Poikien pituudet ilmaistaan edelleen senttimetreissä.) PNS-estimointi tuottaa uudeksi estimaatiksi $100 \times 0.531 = 53.1$. Sekä alkuperäisen että uuden kerroinestimaatin mukaan isän pituuden kasvu metrillä ennustaa pojan venymistä 53.1 senttimetrillä. Jos isien pituudet jaetaan otoskeskihajonnalla (6.173), standardoitu kerroin on $6.173 \times 0.531 = 3.277$. Isän pituuden suuretessa otoskeskihajontansa verran, regressio ennustaa pojan pituuden kurottumista 3.277 senttimetrillä. Sovite on yhtäpitävä alkuperäisestä mallista lasketun kanssa: $0.531 \times 6.173 = 3.278$ (ero kolmannessa desimaalissa johtuu laskutarkkuudesta). □

Esimerkki. Keskihajontojen vaikutus regressiokertoimeen I. Isä-poika-aineistossa standardoitu kerroin on 3.277, kun isien pituudet jaetaan otoskeskihajonnalla. Lisätään alkuperäiseen aineistoon havaintoja $(140, 161.110)$ ja $(230, 208.885)$

molempia kymmenen. Ne toteuttavat regressioyhtälön täydellisesti (y-koordinaatit ovat mallin ennusteet 140 tai 230 senttimetriä pitkien isien pojille). Koska jäännökset ovat nollia näille havainnoille, täydennetty aineisto tuottaa täsmälleen saman PNS-estimaatin 0.531 kuin alkuperäinen aineisto. Isien pituuksien uudella otoskeskihajonnalla (30.715) standardoitu regressiokerroin on $30.715 \times 0.531 = 16.304$. Vaikka aineistoissa pätee täsmälleen sama regressioyhtälö $86.79 + 0.531x$, standardoidut kertoimet 3.277 ja 16.304 eroavat suuresti. Silti ei pidä päätellä, että jälkimmäisessä aineistossa isien pituus on moninkertaisesti tärkeämpi poikien pituuden määrääjä kuin ensimmäisessä, vaikka standardoitu kerroin on siinä moninkertainen. □

Osoitinmuuttujaa ei tule standardoida. (Mitä tarkoittaisi osoitinmuuttujan kasvu yhdellä keskihajonnallaan?) Osoitinmuuttujien estimoituja kertoimia voi vertailla samoin kuin muitakin estimoituja kertoimia. Vertaillessa on syytä pohtia, mitä tarkoittaa tärkeydellä. Yksin sitä, että estimoitu kerroin on itseisarvoltaan suuri? Vai pitäisikö huomioida myös, kuinka paljon osoitinmuuttujan arvo vaihtelee aineistossa tai populaatiossa? Molemmat näkökulmat voivat olla relevantteja. Pienelläkin estimoidulla kertoimella varustettu selittäjä saattaa olla tärkeämpi kuin suurella estimoidulla kertoimella ryyditetty osoitinmuuttuja, jos selittäjä vaihtelee paljon.

Selittäjän tärkeyttä arvioitaessa on syytä huomioida sekä estimoidut kertoimet että selittäjien vaihtelu ja päättää, pohtiiko tärkeyttä vain otoksessa vai yleisemmin.[166] Jos selittäjän vaihteluväli on rajattu, se rajaa myös selittäjän potentiaalista merkitystä vasteeseen. Jos aineistot kerätään (riittävän suurella) satunnaisotannalla, niiden otoshajonnat heijastavat selittäjien todellisia keskihajontoja. Tällöin standardoituja kertoimia (välimatka-asteikollisista selittäjistä) saatetaan voida tulkita kattavammin kuin vain kyseistä aineistoa koskien. Tällöin on luovuttu oletuksesta kiinteistä selittäjistä.

Esimerkki. Keskihajontojen vaikutus regressiokertoimeen II. Kahden selittäjän regressiossa havainnot on kerätty satunnaisotannalla, selittäjät on mitattu samassa yksikössä ja niiden estimoidut kertoimet ovat lähes samat. Selittäjän x_1 otoskeskihajonta (s) on kymmenesosa selittäjän x_2 otoskeskihajonnasta. Standardoidut kertoimet ovat $s\hat{\beta}_1$ ja $10s\hat{\beta}_2 \approx 10s\hat{\beta}_1$. Standardoidut kertoimet viestivät selittäjän x_2 suuremmasta merkityksestä selitettävän määräytymisessä. Samaan päätelmään olisi päästy vertaamalla alkuperäisiä PNS-kertoimia ja selittäjien vaihtelua. □

Regressiomalli (13.10) voidaan esittää yhtäpitävästi siten, että osa selittäjistä korvataan niiden lineaarikombinaatiolla. Syy korvaamiselle voi olla luontevampi tulkinta. Alkuperäisen ja

yhtäpitävän muokatun mallin estimointi PNS:llä tuottaa samat sovitteet. Selittäjän kerroin saattaa estimoitua malleissa erisuureksi ja -merkkiseksi. Se, että selittäjän PNS-estimaatti saattaa olla aivan erilainen yhtäpitävissä malleissa, vahvistaa, että selittäjien tärkeyden vertailu niiden estimoitujen kertoimien perusteella voi olla harhaanjohtavaa. Weisberg (2016, jakso 4.1.3) kuvaa esimerkin.

Jaksossa pohdittiin selittäjien tärkeyttä tilastotieteelliseltä kannalta. Vaikutukseltaan ja selityskyvyltään heikko muuttuja saattaa olla jossain muussa mielessä tärkeä tai tärkein.

13.11.3 Painotettu PNS-menetelmä

Esimerkki. Raiskaustuomiot. Kuvaan 13.11.3 on piirretty käräjä- ja hovioikeuksien määräämiä korvauksia seksuaalisesta väkivallasta.[167] Vaaka-akselilla on seksuaalisen väkivallan uhrin rikoksentekijältä vaatima korvaus (x) ja pystyakselilla oikeuden määräämä korvaus (y). Aineistossa on paljon havaintopareja, joissa vaadittu ja tuomittu korvaus ovat samoja. Havaintoja on sen takia alemmassa kuviossa *ravisteltu* (*jittered*) eli niihin on lisätty satunnaisia lukuja, jotta yksittäiset havainnot erottuvat. Molempiin kuvioihin on piirretty (ravistelemattomasta aineistosta) PNS-menetelmällä laskettu regressiosuora. Mallin satunnaistermin varianssi vaikuttaa suurenevan vaaditun korvaustason kasvaessa. □

Mitä tapahtuu, jos lineaarisen regressiomallin (13.10) satunnaistermin varianssi ei ole vakio? Parametrien PNS-estimaatit voivat olla edelleen järkeviä mutteivät ole enää optimaalisia. Kuvattu vakioiseen varianssiin nojaava testisuureiden jakaumateoria ei myöskään päde.

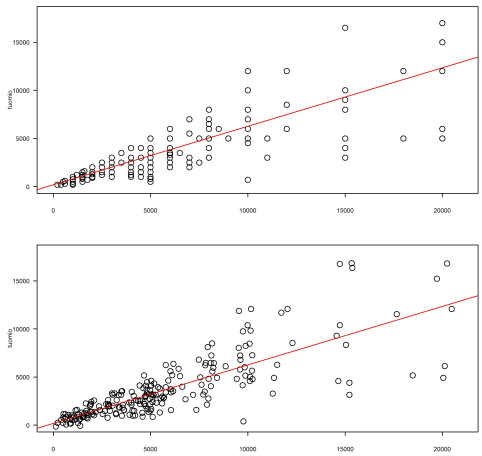
Jos varianssi on joidenkin havaintojen kohdalla hyvin pieni, näissä havainnoissa on erityisen paljon informaatiota mallin parametreista. Vastaavasti jos varianssi on suuri, havainto on vähemmän informatiivinen. Tällainen tieto voidaan hyödyntää estimoinnissa.

Oletetaan, että satunnaisvirheiden varianssi on

$$\mathsf{V}(\varepsilon_i) = \frac{\sigma^2}{w_i}.$$

Tässä w_i on havaintokohtainen paino, joka on kiinteä tunnettu luku. Kun w_i on suuri, satunnaistermin varianssi σ^2/w_i on pieni. Parametrit kannattaa tällöin estimoida minimointitehtävästä

$$\min_{\beta_0,\ldots,\beta_k} \sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2.$$



Kuva 13.20: Seksuaalisen väkivallan uhrin rikoksentekijältä vaatima korvaus (x) ja oikeuden määräämä korvaus (y). Alemman kuvion aineistoa on ravisteltu.

Siinä painotetaan niitä havaintoja, joissa satunnaistermin varianssi on pientä eli niitä havaintoja, joissa on eniten informaatiota parametreista. Parametrien estimaatit pyritään tällöin valitsemaan niin, että satunnaistermit muodostuisivat pieniksi erityisesti havainnoissa, joita kerrotaan suurilla w_i-painoilla. Näin saatuja estimaattoreita kutsutaan painotetuiksi PNS-estimaattoreiksi (PPNS; *weighted least squares*).

PPNS-estimoinnin jälkeen analyysi etenee jo kuvattuun tapaan. Tilasto-ohjelmistot tyypillisesti raportoivat — nyt hieman eri tavalla lasketut — estimoidut keskivirheet estimaattoreille. PPNS-estimaattorit voidaan jakaa estimoiduilla keskivirheillään, ja testaus sujuu sen jälkeen kuten edellä. Myös F-testit toimivat vastaavaan tapaan, ja PPNS-estimaattorit ovat samallalailla optimaalisia kuin PNS-estimaattorit satunnaistermin vakioisen varianssin tilanteessa. Intuitiivisesti nämä tulokset seuraavat siitä, että PPNS-minimointitehtävässä

yllä satunnaistermit kerrotaan painoilla $\sqrt{w_i}$

$$\begin{aligned} & \textstyle\sum_{i=1}^n w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\ = \; & \textstyle\sum_{i=1}^n w_i(\varepsilon_i)^2 \\ = \; & \textstyle\sum_{i=1}^n (\sqrt{w_i}\,\varepsilon_i)^2. \end{aligned}$$

Summalausekkeen uuden satunnaistermin $\sqrt{w_i}\,\varepsilon_i$ varianssi on vakio:

$$\mathsf{V}(\sqrt{w_i}\,\varepsilon_i) = w_i \mathsf{V}(\varepsilon_i) = w_i \frac{\sigma^2}{w_i} = \sigma^2$$

(jakso 6.3). Satunnaistermien kertominen painoilla $\sqrt{w_i}$ ikään kuin palauttaa mallin sellaiseksi, että satunnaistermin varianssi on vakio.

Miten havaitsemattomat satunnaistermit voidaan kertoa painoilla $\sqrt{w_i}$? Käytännössä siihen päästään kertomalla mallin kaikkien muiden muuttujien *i.* havainnot $\sqrt{w_i}$:llä:

$$\begin{aligned} \sqrt{w_i}y_i = \beta_0\sqrt{w_i} + \beta_1\sqrt{w_i}x_{1i} + \cdots + \beta_k\sqrt{w_i}x_{ki} + \sqrt{w_i}\varepsilon_i \quad \Leftrightarrow \\ \sqrt{w_i}y_i - \beta_0\sqrt{w_i} - \beta_1\sqrt{w_i}x_{1i} - \cdots - \beta_k\sqrt{w_i}x_{ki} = \sqrt{w_i}\varepsilon_i. \end{aligned}$$

Näin muunnettu malli estimoidaan PNS:llä.

Esimerkki. Oppimistulokset. Selitetään koulujen oppimistuloksia koulun oppilaaksiottoalueen piirteillä kuten kouluttamattoman, korkeakoulutetun ja vieraskielisen väestön osuuksilla ja keskimääräisellä tulotasolla.[168] Oppimistulos on keskiarvo koulun oppilaiden oppimistuloksista. Merkitään yksittäisen oppilaan oppimistuloksen varianssia σ^2:lla. Koulun i oppimistuloksen varianssi on tällöin σ^2/n_i (jakso 9.6), jossa n_i on oppilaiden lukumäärä *i.* koulussa ($i = 1, \ldots, n$, jossa n on koulujen lukumäärä aineistossa). Oppimistulosten varianssi on pienempi oppilaslukumäärältään suuressa kuin pienessä koulussa. Regressiomalli koulujen oppimistuloksille kannattaa estimoida PPNS-menetelmällä painottamalla havaintoja oppilaslukumäärillä $w_i = n_i$. □

Esimerkki.[169] Painotuksen suuri vaikutus. Reinhart ja Rogoff (2010) argumentoivat kuuluisassa artikkelissaan, että talouden kasvu hidastuu suuresti, jos valtion velan suhde bruttokansantuotteeseen ylittää 90 %. Moni tiukan talouspolitiikan kannattaja vetosi tutkimukseen. Taloustieteen opiskelija Thomas Herndon huomasi, että Reinhartin ja Rogoffin tulokset ovat virheellisiä. Hän kirjoitti ohjaajiensa kanssa artikkelin, jossa kritisoi muun muassa tapaa, jolla Reinhart ja Rogoff painottivat havaintoja. Seurasi vyöry kritiikkiä politiikasta, jota oltiin

perusteltu Reinhartin ja Rogoffin tutkimuksella. Maziarz (2017) arvioi, että erot Reinhardtin ja Rogoffin ja Herndonin ym:iden tulosten välillä johtuvat ennen kaikkea erilaisesta tavasta painottaa havainnot. □

Satunnaistermin varianssi voi vaihdella havainnoittain eli olla *heteroskedastinen* niin, että varianssin suuruus kullakin havainnolla ei ole tarkkaan tiedossa. Raiskaustuomiot-esimerkissä voitaisiin w_i:den riippuvuus vaaditusta korvauksesta ehkä estimoida. Toisissa epäselvemmissä tilanteissa, joissa varianssi vaihtelee havainnoittain mutta tiedossa ei ole miten riippuvuus liittyy esimerkiksi selittäjiin, voidaan käyttää tällaisiin tilanteisiin kehitettyjä estimointi- ja testausmenetelmiä. Niitä selitetään esimerkiksi Baltagin (2011), Verbeekin (2017), Wilcoxin (2012) ja Weisbergin (2014) oppikirjoissa. Fox (2016, 307) esittää (hyvin karkean) ohjenuoran, että heteroskedastisuus on estimoinnin tehokkuuden ja testien validiuden kannalta vakava ongelma, jos satunnaistermin varianssi on suurimmillaan yli nelinkertainen sen pienimpään arvoon verrattuna.

13.11.4 Muuttujien logaritmointi

Logaritmifunktio (luonnollinen logaritmi) kaartuu kuvassa 13.21. Logaritmifunktio saa arvoja välillä $(-\infty, \infty)$ (miinus ja plus ääretön) x:n kasvaessa välillä $(0, \infty)$ (nollasta äärettömään).

Avataan ensin logaritmin muutoksen ja suhteellisen muutoksen yhteys. Lähtökohta on likiarvoistus

$$\log(1 + \delta) \approx \delta,$$

jossa $|\delta|$ on "pieni" ja $1 + \delta > 0$.

Esimerkki. Olkoon $\delta = 0.1$. Tällöin $\log(1 + 0.1) = \log(1.1) = 0.0953 \approx 0.1$. □

Likiarvoistuksesta ja laskusäännöstä $\log(xy) = \log(x) + \log(y)$ $(x > 0$ ja $y > 0)$ seuraa, että

$$\log[x(1 + \delta)] = \log(x) + \log(1 + \delta) \approx \log(x) + \delta$$

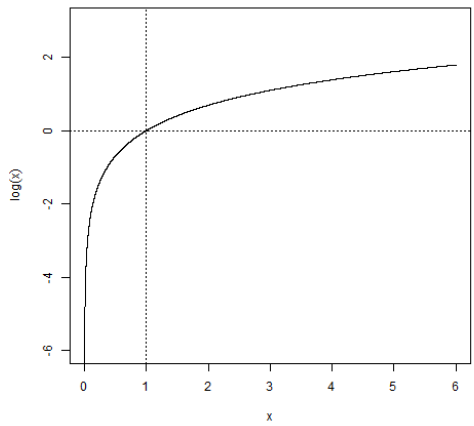
ja

$$\log[x(1 + \delta)] - \log(x) \approx \delta.$$

Logaritmin argumentin muuttuessa $100 \times \delta$ prosentilla muuttuu logaritmi noin δ:lla. Logaritmin muutokset kerrottuna sadalla ovat likimääräisiä prosenttimuutoksia.

Esimerkki. Olkoon $x = 100$ ja $\delta = 0.05$:

- $100(1 + 0.05) = 105$ — logaritmin argumentti suurenee 5 prosenttia.



Kuva 13.21: Logaritmifunktio $\log(x)$.

- $\log[100(1 + 0.05)] - \log(100) = \log(105) - \log(100) = 0.0488$ — logaritmi suurenee noin 0.05:llä.[170] □

Muuttujien logaritmointi on yleistä empiirisessä regressioanalyysissä. Pohditaan yhden selittäjän lineaarisen mallin tulkintaa, kun jompikumpi tai molemmat muuttujista ($x > 0$ ja $y > 0$) on logaritmoitu. Analyysi yleistyy ilmeisellä tavalla monen selittäjän lineaarisen mallin tilanteeseen.

1. Logaritmoidaan selittäjä:

$$y = \beta_0 + \beta_1 \log(x) + \varepsilon.$$

   Tulkinta: x:n muuttuessa 1 prosentin, y muuttuu $\beta_1 \times 0.01$:n verran. (Perustelu: Logaritmin argumentin muuttuessa $100 \times \delta$ prosentilla muuttuu logaritmi noin δ:lla. Tässä $\delta = 0.01$.) Tällöin x:n (absoluuttisen) muutoksen vaikutus y:hyn pienenee x:n kasvaessa.

2. Logaritmoidaan vaste:

$$\log(y) = \beta_0 + \beta_1 x + \varepsilon.$$

Tulkinta: x:n muuttuessa yksikön verran y muuttuu $100 \times \beta_1$ prosenttia. (Perustelu yo. tapaan.) Tällöin x:n muutoksen vaikutus y:hyn (absoluuttisesti) suurenee x:n kasvaessa.

3. Logaritmoidaan sekä selittäjä että vaste:

$$\log(y) = \beta_0 + \beta_1 \log(x) + \varepsilon.$$

Tulkinta: x:n muuttuessa 1 prosentin, y muuttuu β_1 prosenttia.[171]

Huom! Selitysastetta R^2 tilanteissa 2 ja 3 vertailemalla voidaan pohtia, tulisiko selittäjä logaritmoida vai ei. Vertailua ei voi laajentaa kattamaan tilanteen 1 mallia, sillä siinä vaste on eri.[172] Vasteen logaritmointipäätös täytyy tehdä muilla perusteilla kuin selitysastetta (esim. tilanteissa 1 ja 3) vertailemalla.

Luku 14

Kaksiarvoinen vastemuuttuja ja regressio

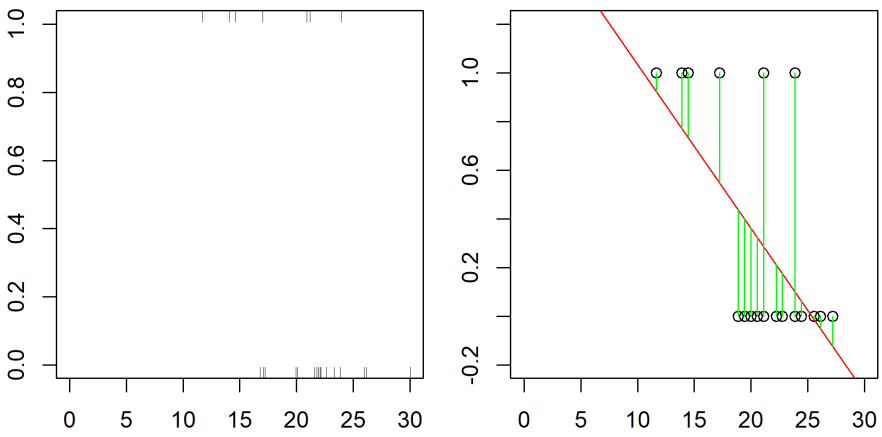
> Todelliseksi tieteeksi ei voida kutsua mitään inhimillistä tutkimusta, joka ei ole matemaattisesti todennettavissa. Jos pidät erehtymättöminä todellisia tieteitä, jotka alkavat ja päättyvät ajatuksissa, niin olen eri mieltä ja kiistän näkemyksesi monien seikkojen perusteella. Tärkein niistä on, että mielen harjoituksista puuttuu kokemuksen testi, ja ilman sitä ei ole vakuutta varmuudesta.[173]
>
> Leonardo da Vinci (1452–1519)

Tapahtuuko? Parantuuko? Ostaako? Vai ei? Monet ilmiöt ovat kaksiarvoisia Bernoulli-satunnaismuuttujan tapaan (jakso 7.1.1). Ostopäätös voidaan kuvata numeroilla 1 (ostaa) tai 0 (ei osta). Ilmeistä on, että päätös riippuu monista seikoista kuten tuotteen hinnasta ja ostoa harkitsevan tuloista. Luonnollinen ajatus on selittää päätöstä luvussa 13 kuvattuun tapaan regressiolla. Koska selitettävä muuttuja on *kaksiarvoinen* (*binary*), luvun 13 regressiomallit ja PNS-menetelmä niiden estimoimiseksi eivät sovi tarkoitukseen hyvin. Luvussa esitetään kaksi regressiomallia, joita voidaa käyttää kaksiarvoisen vastemuuttujan tilanteessa.

14.1 Lineaarinen todennäköisyysmalli

Kuvan 14.1 vasemmassa puoliskossa on *hapsukuvio* (*rug plot*), jossa kutakin havaintoa kaksiarvoisesta (1/0) satunnaismuuttujasta — tässä lämpörasitusvauriosta *td* (*thermal distress*) — on merkitty hapsulla. (Vaurio = 1; ei vauriota =

0.) Aineisto kuvataan tarkemmin jaksossa 14.2.2. Havaintoja on ravisteltu, jotta päällekkäiset havainnot erottuisivat. Vaaka-akselilla on lämpötila c (celsiusta). Alhaisilla lämpötiloilla esiintyy yksinomaan vaurioita; korkeilla ei lainkaan. Vaurioita voisi ajatella selitettävän lämpötilalla — mutta miten?



Kuva 14.1: Havaintoja kaksiarvoisesta satunnaismuuttujasta sekä regressiosuora ja jäännökset.

Regressiomallin (13.1) mukaan vaste kasvaa tai vähenee selittäjän ja jäännöksen arvojen mukaisesti. Kaksiarvoinen vaste ei voi määräytyä näin. Jäännöksen pitäisi jakautua regressiosuoran ympärille, mikä ei ole mahdollista kaksiarvoisen vasteen tilanteessa. Kuvan 14.1 oikeaan puoliskoon on piirretty lämpörasitusvaurioaineistoon PNS-menetelmällä istutetun mallin (13.1) regressiosuora ja jäännökset (janat palluraisten ja regressiosuoran välillä). Regressiosuora ennustaa yhtä suurempaa todennäköisyyttä lämpötilan alittaessa noin 10.5 astetta ja nollaa pienempää todennäköisyyttä lämpötilan ylittäessä 25.4 astetta. Korkein lämpötila aineistossa on 27.2 astetta. Regressiosuora ei ole järkevä. Jäännökset viestivät luvussa 13 tehtyjen oletusten rikkoontumisesta. Satunnaistermin jakauma muuttuu havainnoittain, satunnaistermin odotusarvo ja varianssi vaihtelevat eivätkä satunnaistermit ole riippumattomia.

Jos lämpötilavaurioaineistossa olisi paljon havaintoja, ne voitaisiin ryhmitellä lämpötilaluokkiin ja laskea keskiarvo 1/0-havainnoista kussakin luokassa. Näin saadut luokkakeskiarvot voisivat toimia regressiomallin (13.1) (keskeisen

raja-arvolauseen perusteella) likimäärin normaalijakautuneina vasteina, ja lämpötilavaurion todennäköisyyttä voitaisiin yrittää selittää lämpötilalla regressiomallin avulla. Tähän tapaan toimittiin uraseurantakyselyesimerkissä jaksossa 13.6.1.

Lineaarisella todennäköisyysmallilla (*linear probability model*) selitetään tapahtuman todennäköisyyttä selittäjien lineaarikombinaatiolla:

$$\mathsf{E}(Y) = \pi(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{14.1}$$

Tässä $\pi(x_1, \ldots, x_k)$ on tapahtuman $Y = 1$ todennäköisyys. Suluissa osoitetaan todennäköisyyden riippuvuus k:sta selittäjästä. Todennäköisyys määräytyy selittäjien lineaarikombinaatiosta $\sum_{i=0}^{k} \beta_i x_i$. Mallia estimoitaessa vasteet voivat olla kaksiarvoisia havaintoja tai niistä laskettuja luokkakeskiarvoja. Yksinkertaisuuden vuoksi selittäjät oletetaan luvussa kiinteiksi.

PNS on helppo ja varsin luonteva estimointimenetelmä, kun mallia sovelletaan luokkakeskiarvoihin. PNS ei ole tällöin silti paras estimointimenetelmä. Luokkakeskiarvot on laskettu Bernoulli-jakautuneista satunnaismuuttujista, joiden varianssi on yhden selittäjän tilanteessa $\pi(x_1)[(1 - \pi(x_1)]$ (jakso 7.1.1). Jos todennäköisyys $\pi(x_1)$ riippuu selittäjästä x_1, *i.* luokkakeskiarvon varianssi on $\pi(x_1)[1 - \pi(x_1)]/n_i$, jossa n_i on havaintojen lukumäärä *i.* luokassa. Tällöin vasteen varianssi vaihtelee selittäjän arvon ja n_i:n mukaan eikä PNS-menetelmä ole optimaalinen (jaksot 13.7 ja 13.11.3). Silti PNS-menetelmä voi olla varsin toimiva. Suositeltavinta on estimoida malli suurimman uskottavuuden menetelmällä (jakso 9.2).

Vasteen erivarianssisuus ei huononna PNS-menetelmän tarkkuutta paljoa, jos estimoidut todennäköisyydet $\hat{\pi}(x_{i1}, \ldots, x_{ik})$ ovat välillä $(0.3, 0.7)$ (Agresti 2013, 118, vrt. Cox ja Snell 1989, 18, Collett 2003, 53). Cox ja Snell (1989, 19) sekä Collett (2003, 54) selittävät, miten lineaarinen todennäköisyysmalli estimoidaan painotetulla PNS-menetelmällä, joka huomioi havaintojen erivarianssisuuden.

Malli voidaan estimoida suurimman uskottavuuden (SU) -menetelmällä (jakso 9.2). Oletetaan, että on tehty n riippumatonta Bernoulli-koetta, joissa tapahtuman todennäköisyys on $\pi(x_{i1}, \ldots, x_{ik})$. Kuhunkin havaintoon liittyvän satunnaismuuttujan odotusarvo on $\pi(x_{i1}, \ldots, x_{ik})$ ja varianssi on $\pi(x_{i1}, \ldots, x_{ik})[1 - \pi(x_{i1}, \ldots, x_{ik})]$ (jakso 7.1.1). Merkitään $y_i = 1$ tai $y_i = 0$ sen mukaan, toteutuiko tapahtuma *i.* havainnossa vai ei. Kunkin havainnon todennäköisyys on $[\pi(x_{i1}, \ldots, x_{ik})]^{y_i}[1-\pi(x_{i1}, \ldots, x_{ik})]^{1-y_i}$ eli $\pi(x_{i1}, \ldots, x_{ik})$ tai $1-\pi(x_{i1}, \ldots, x_{ik})$, jos $y_i = 1$ tai $y_i = 0$. Kaikkien havaintojen todennäköisyys on riippumattomuuden perusteella havaintojen todennäköisyyksien tulo

$$\prod_{i=1}^{n} [\pi(x_{i1}, \ldots, x_{ik})]^{y_i} [1 - \pi(x_{i1}, \ldots, x_{ik})]^{1-y_i}. \tag{$*$}$$

Kukin todennäköisyys riippuu β_j-parametreista kaavan (14.1) mukaisesti. SU-menetelmässä havaintoja $x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$, pidetään kiinteinä, lauseke yllä maksimoidaan parametrien β_j suhteen ja estimaateiksi valitaan kaavan maksimoivat arvot $\hat{\beta}_j$, $j = 0, \ldots, k$. Ylipäänsä SU-estimaateille ei ole analyyttista ratkaisua (kaavaa joka tuottaisi estimaatit sijoittamalla havaintojen arvot siihen). Tilasto-ohjelmisto etsii maksimiarvon numeerisesti kokeilemalla erilaisia arvoja parametreille β_j. Maksimoinnin yksityiskohdat sivuutetaan.

Poikkeavat havainnot voivat vaikuttaa suuresti lineaarisen todennäköisyysmallin regressiokertoimien SU-estimaatteihin. PNS voi olla tällöin sopiva menetelmä mallin estimoimiseksi. Estimaattoreiden keskivirheet tulee tällöin estimoida tavanomaisesta poikkeavalla kaavalla jäännösten heteroskedastisuuden takia. (Battey ym. 2019.)

R-koodi lineaarisen todennäköisyysmallin estimoimiseksi SU-menetelmällä löytyy Laura Thompsonin ja Alan Agrestin oppaista `http://users.stat.ufl.edu/~aa/cda/Thompson_manual.pdf` ja `http://users.stat.ufl.edu/~aa/cda/R_web.pdf` (haettu 19.4.2020) sekä Agrestin (2019, 71) kirjasta.

Lineaarisen todennäköisyysmallin valtti on sen yksinkertaisuus. Mallin rajoite on, että sen mukaan todennäköisyys saa väistämättä 1:stä suurempia tai 0:aa pienempiä arvoja, kun selittäjä suurenee tai pienenee riittävästi. Tällaiset sovitteet tai ennusteet eivät ole järkeviä. Lineaarinen todennäköisyysmalli voi olla käyttökelpoinen vain rajatulla selittäjien arvojen vaihteluvälillä.

14.2 Logistinen regressiomalli

14.2.1 Logistisen regressiomallin teoriaa

Logistisen regressiomallin idea on, ettei selitetä todennäköisyyttä vaan sen *logit-muunnosta* $\text{logit}\,[\pi(x_1)]$:

$$\text{logit}\,[\pi(x_1)] \equiv \log \frac{\pi(x_1)}{1 - \pi(x_1)} = \beta_0 + \beta_1 x_1. \tag{14.2}$$

Yllä $\pi(x_1) > 0$, jotta logaritmi on määritelty. Tarkasteltavana on yhden selittäjän (x_1) versio logistisesta regressiomallista. Osamäärä $\pi(x_1)/[1 - \pi(x_1)]$ on *vastasuhde* (*odds*), johon tutustuttiin jaksossa 4.2.3 (kaava (4.3)) ja joka piirrettiin jaksossa 4.6 (kuva 4.15).

Todennäköisyys sijoittuu välille $(0, 1)$, joten vastasuhde voi saada arvoja välillä $(0, \infty)$ (nollasta äärettömään). Vastasuhteen logaritmi eli todennäköisyyden logit-muunnos saattaa siten vaihdella välillä $(-\infty, \infty)$ (kuva 13.21). Logit-muunnoksen avulla vältetään yhtälön (14.1) periaatteellinen ongelma, että yhtälön vasen puoli on rajoitettu välille $(0, 1)$ mutta oikea puoli ei.

Yhtälöstä (14.2) voidaan ratkaista tapahtuman todennäköisyys:

$$\pi(x_1) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)} \tag{14.3}$$

(harjoitustehtävä). Tässä $\exp(\beta_0 + \beta_1 x_1)$ on Neperin luku e korotettuna potenssiin $\beta_0 + \beta_1 x_1$. Todennäköisyys $\pi(x_1)$ määräytyy selittävän muuttujan x_1 arvosta mutta on rajattu välille $(0, 1)$.

Kuvat 14.2–14.4 havainnollistavat.[174] Selittäjän vaikutus todennäköisyyteen riippuu selittäjän arvon suuruudesta, paitsi jos $\beta_1 = 0$, jolloin β_0 määrää todennäköisyyden. Se on tilanne kuvassa 14.2. Todennäköisyys on sitä suurempi, mitä suurempi β_0 on. Todennäköisyys on 0.5, jos $\beta_0 = 0$. Kuvassa 14.3 maalaillaan toista ääritilannetta $\beta_0 = 0$, jolloin todennäköisyys riippuu yksin β_1:stä. Vaakasuorat viivat pätevät, jos myös $\beta_1 = 0$, jolloin $\pi(x_1) = 0.5$ kaikilla x_1:n arvoilla. Muulloin todennäköisyys $\pi(x_1)$ riippuu x_1:stä sitä jyrkemmin, mitä suurempi β_1:n itseisarvo on. Idea on intuitiivinen: Vaikkapa ostopäätöstä tehdessä tulojen kasvaessa oston todennäköisyys voi kasvaa suuresti alkuun, mutta tavattoman suurilla tuloilla tulojen kasvu ei enää juurikaan suurenna oston todennäköisyyttä. Kuva 14.4 visualisoi todennäköisyyden määräytymistä β_0:n ja β_1:n eri yhdistelmillä. Vaakasuorien viivojen tilanteissa $\beta_1 = 0$. Ylipäänsä selittäjän saadessa yhä suurempia arvoja todennäköisyys kasvaa tai pienenee sen mukaan, onko β_1 positiivinen vai negatiivinen, mutta vähenevässä määrin. Pisteessä $x_1 = 0$ todennäköisyys riippuu yksin vakion β_0 suuruudesta.

Tärkeä ero lineaariseen todennäköisyysmalliin (14.1) on, että selittäjän x_1 yksikön suuruisen muutoksen vaikutus todennäköisyyteen ei ole vakio. Vaikutus vaihtelee sen mukaan, mistä lukuarvosta x_1:htä muutetaan (kuvat 14.3–14.4).[175]

Todennäköisyyskäyrän (14.3) kulmakerroin on

$$\beta_1 \pi(x_1)[1 - \pi(x_1)]. \tag{14.4}$$

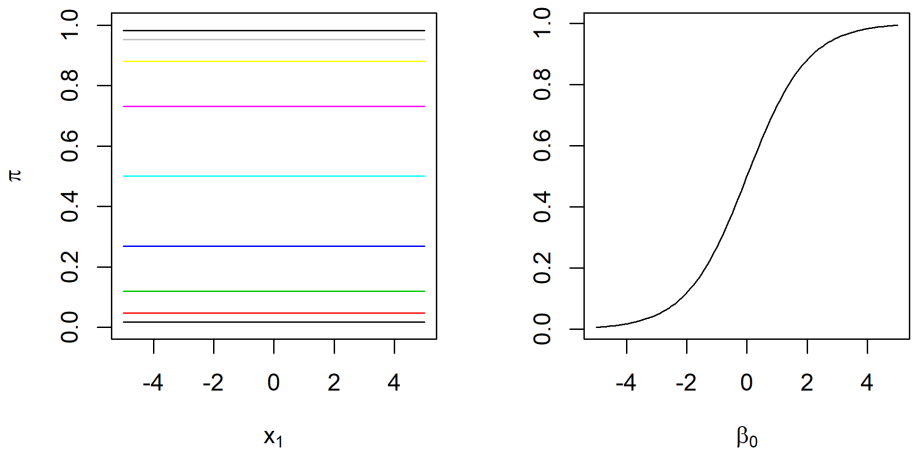
Sen itseisarvo on suurimmillaan $|0.25\beta_1|$, kun $\pi(x_1) = 1/2$ ja $x_1 = -\beta_0/\beta_1$ (harjoitustehtävä).

Korotetaan Neperin luku e yhtälön (14.2) vasemman ja oikean puolen mukaisiin potensseihin:

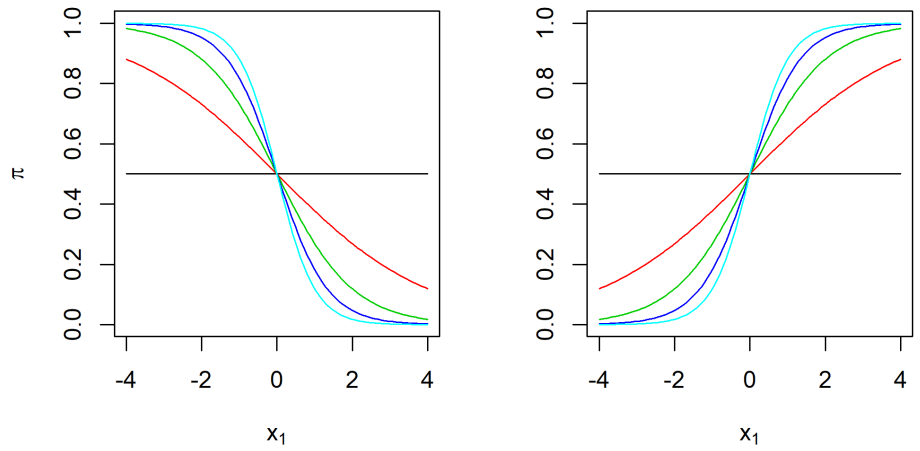
$$\exp\left[\log \frac{\pi(x_1)}{1 - \pi(x_1)}\right] = \frac{\pi(x_1)}{1 - \pi(x_1)} = \exp(\beta_0 + \beta_1 x_1). \tag{14.5}$$

Ensimmäinen yhtäsuuruus seuraa siitä, että $\exp[\log(x_1)] = x_1$ $(x_1 > 0)$. Jos x_1 suurenee yksiköllä, vastasuhde kertautuu vakiolla $\exp(\beta_1)$:

$$\exp[\beta_0 + \beta_1(x_1 + 1)] = \exp(\beta_0 + \beta_1 x_1) \exp(\beta_1). \tag{14.6}$$



Kuva 14.2: Todennäköisyys π yhden selittäjän logistisessa regressiossa, jos $\beta_1 = 0$. Vasemmalla: π, jos $\beta_0 = -4, -3, \ldots, 3, 4$ (suorat alhaalta ylös). Oikealla: π funktiona β_0:sta.



Kuva 14.3: Todennäköisyys π yhden selittäjän logistisessa regressiossa, jos $\beta_0 = 0$. Vasemmalla: $\beta_1 = 0, -0.5, -1, -1.5, -2$. Oikealla: $\beta_1 = 0, 0.5, 1, 1.5, 2$.

Logistisella regressiolla selviää siis vastasuhteiden $\pi(x_1+1)/[1-\pi(x_1+1)]$ ja $\pi(x_1)/[1-\pi(x_1)]$ suhde *ristisuhde* (*odds ratio*, OR) $\exp(\beta_1)$:

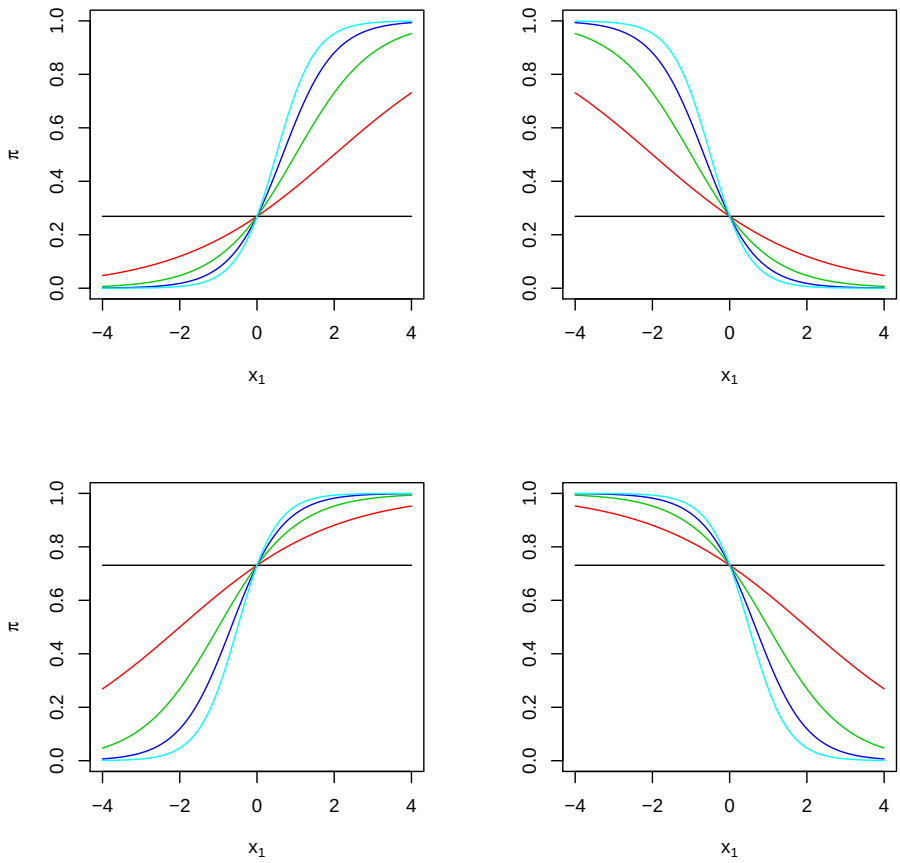
$$\theta = \frac{\dfrac{\pi(x_1+1)}{1-\pi(x_1+1)}}{\dfrac{\pi(x_1)}{1-\pi(x_1)}} = \frac{\exp[\beta_0+\beta_1(x_1+1)]}{\exp(\beta_0+\beta_1 x_1)} = \exp(\beta_1). \tag{14.7}$$

Toinen yhtäsuuruus tulee yhtälöstä (14.5). Ristisuhde ei riipu selittäjän x_1 arvosta. Selittäjä x_1 voi olla osoitinmuuttuja. Tällöin $\exp(\beta_1)$ kertoo vastasuhteiden suhteen osoitetussa luokassa ja vertailuluokassa. Jos todennäköisyys ja selittäjä eivät ole yhteydessä, $\beta_1 = 0$ ja ristisuhde $\theta = \exp(0) = 1$. Jos $\beta_1 > 0$, niin $\theta > 1$; jos $\beta_1 < 0$, niin $\theta < 1$. Mitä enemmän β_1 poikkeaa 0:sta, sitä enemmän ristisuhde θ eroaa 1:stä. Ristisuhde saa arvoja välillä $(0, \infty)$.

Jos selittäjiä on k kappaletta, logistinen regressiomalli on

$$\text{logit}\,[\pi(x_1,\ldots,x_k)] = \log\frac{\pi(x_1,\ldots,x_k)}{1-\pi(x_1,\ldots,x_k)} = \beta_0+\beta_1 x_1+\cdots+\beta_k x_k. \tag{14.8}$$

Yhtälöstä (14.8) seuraa yhtälöitä (14.3) ja (14.6) vastaavat todennäköisyyden



Kuva 14.4: Todennäköisyys π yhden selittäjän logistisessa regressiossa. Vasemmalla: $\beta_1 = 0$, 0.5, 1, 1.5, 2. Oikealla: $\beta_1 = 0, -0.5, -1, -1.5, -2$. Ylhäällä: $\beta_0 = -1$. Alhaalla: $\beta_0 = 1$.

ja vastasuhteen kaavat

$$\pi(x_1, \ldots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \tag{14.9}$$

ja

$$\frac{\pi(x_1, \ldots, x_k)}{1 - \pi(x_1, \ldots, x_k)} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k). \tag{14.10}$$

Tapahtuman todennäköisyydet eri selittäjien arvoilla voidaan laskea yhtälöstä (14.9). Selittäjän x_j $(j = 1, \ldots, k)$ yksikön suuruisen muutoksen vaikutus, kun muut selittäjät ovat kiinnitettyjä, ei riipu minkään selittäjän arvosta. Näin syntyvän ristisuhteen suuruuden $\exp(\beta_j)$ määrää yksin β_j:

$$\begin{aligned}\theta_j &= \frac{\dfrac{\pi(x_1, \ldots, x_j + 1, \ldots, x_k)}{1 - \pi(x_1, \ldots, x_j + 1, \ldots, x_k)}}{\dfrac{\pi(x_1, \ldots, x_k)}{1 - \pi(x_1, \ldots, x_k)}} \\ &= \frac{\exp[\beta_0 + \beta_1 x_1 + \ldots + \beta_j(x_j + 1) + \cdots + \beta_k x_k]}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} = \exp(\beta_j).\end{aligned}$$

Toinen yhtäsuuruus johtuu yhtälöstä (14.10). Selittäjät voivat olla jatkuva-arvoisia, luokka-, osoitin- tai yhdysvaikutusmuuttujia luvun 13 regressiomallien tapaan.

14.2.2 Logistisen regressiomallin estimointi, testaus ja selityskyky

Aineisto koostuu n:n riippumattoman Bernoulli-kokeen tuloksena syntyneistä havaintovektoreista $[x_{11} \ \ldots \ x_{1k} \ y_1], \ldots, [x_{n1} \ \ldots \ x_{nk} \ y_n]$ $(n \geq k)$. Vasteet y_i saavat arvon 0 tai 1. Selittäjät ovat kiinteitä. Logistinen regressiomalli estimoidaan yleensä SU-menetelmällä (jakso 9.2). Estimoinnin yksityiskohdat sivuutetaan. Vakiotermin ja regressiokertoimien estimaattorit ovat (käyvin oletuksin) tarkentuvia mutteivät ylipäänsä harhattomia. Estimointi onnistuu R:ssä versioimalla koodia

```
malli <- glm(y~x1+x2+x3, family=binomial(link=logit), data=aineisto)
summary(malli)
```

sopivasti (kolme selittäjää; "aineisto" muuttujista y, x_1, x_2 ja x_3 valmiina R:ssä).

Uskottavuusfunktion ydin on yhtälö (∗) sivulla 323. Todennäköisyydet $[\pi(x_{i1}, \ldots, x_{ik})]^{y_i}$ ja $[1 - \pi(x_{i1}, \ldots, x_{ik})]^{1-y_i}$ riippuvat nyt β_j-parametreista kaavan (14.9) mukaisesti. SU-menetelmässä havaintoja $x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$, pidetään kiinteinä, kaava (∗) maksimoidaan numeerisilla menetelmillä parametrien β_j suhteen ja estimaateiksi valitaan kaavan maksimoivat arvot $\hat{\beta}_j$, $j = 0, \ldots, k$.

SU-menetelmällä voidaan laskea myös SU-estimaattorien estimoidut keskivirheet. Jakamalla SU-estimaatti $\hat{\beta}_j$ estimoidulla keskivirheellään saadaan t-arvo nollahypoteesille $\beta_j = 0$. R raportoi ne edellä osoitetun `summary(malli)`-käskyn palautteessa. Siinä ne on nimetty z-arvoiksi, koska nollahypoteesin pätiessä niiden pienotosjakaumaa ei tunneta mutta jakauma on standardinormaali suurilla havaintomäärillä. Varovainen konservatiivinen testaaja — joka ei halua altistua testin nimellistä kokoa suuremmalla riskille tehdä hylkäysvirhe — voi verrata t-testisuuretta standardinormaalijakaumaa paksuhäntäisempään `t`-jakaumaan $n - k - 1$:llä vapausasteella (vrt. jakso 11.3).

Logistisen regressiomallin selityskykyä voidaan mitata jaksosta 4.6 tutuilla herkkyydellä ja tarkkuudella sekä erityisesti todennäköisyydellä luokitella havainto oikein. Tulkitaan, että malli sovittaa vastemuuttujan arvon oikein, jos $\hat{\pi}_i > \pi_0$, jos $y = 1$ tai $\hat{\pi}_i < \pi_0$, jos $y = 0$. Todennäköisyys π_0 on itse asetettu raja. Yksi luonteva sellainen on $\pi_0 = 0.5$. Tällöin mallin katsotaan ennustavan havainnon i kohdalla tapahtumaa $y = 1$, jos $\hat{\pi}_i > 0.5$ ja vastatapahtumaa $y = 0$, jos $\hat{\pi}_i < 0.5$. Toinen luonteva π_0-raja on tapahtumien eli $y = 1$ -havaintojen osuus aineistossa $\bar{y}$. Muitakin rajoja voidaan käyttää.

Merkitään $\hat{y}_i = 1$ tai $\hat{y}_i = 0$, jos $\hat{\pi}_i > \pi_0$ tai $\hat{\pi}_i < \pi_0$. Herkkyys ja tarkkuus määritellään tässä yhteydessä todennäköisyyksiksi, että malli luokittelee tapahtumat $y = 1$ ja $y = 0$ oikein:

$$\text{herkkyys} = \mathsf{P}(\hat{y} = 1 \mid y = 1) \quad \text{ja} \quad \text{tarkkuus} = \mathsf{P}(\hat{y} = 0 \mid y = 0).$$

Oikean luokittelun todennäköisyys on painotettu keskiarvo herkkyydestä ja tarkkuudesta:

$$\begin{aligned}
&\mathsf{P}[(\hat{y} = 1 \cap y = 1) \cup (\hat{y} = 0 \cap y = 0)] \\
&\quad = \mathsf{P}(\hat{y} = 1 \cap y = 1) + \mathsf{P}(\hat{y} = 0 \cap y = 0) \\
&\quad = \mathsf{P}(\hat{y} = 1 \mid y = 1)\mathsf{P}(y = 1) + \mathsf{P}(\hat{y} = 0 \mid y = 0)\mathsf{P}(y = 0) \\
&\quad = \text{herkkyys} \times \mathsf{P}(y = 1) + \text{tarkkuus} \times \mathsf{P}(y = 0).
\end{aligned}$$

Ensimmäinen yhtäsuuruus seuraa tapahtumien $\hat{y} = 1 \cap y = 1$ ja $\hat{y} = 0 \cap y = 0$ erillisyydestä sekä erillisten tapahtumien todennäköisyyksien yhteenlaskusäännöstä (4.4). Toisessa yhtäsuuruudessa on hyödynnetty todennäköisyyslaskennan

tulosääntöä (4.9). Oikean luokittelun todennäköisyys riippuu suuresti valitusta rajasta π_0.

Erottelukykykäyrä (*receiver operating characteristic curve*, *ROC curve*) piirtää (1−tarkkuus, herkkyys)-pisteet kaikille π_0:n arvoille (1−tarkkuus, herkkyys)-koordinaatistoon (kuva 14.8). Käyrä alkaa $(0, 0)$-pisteestä: Jos $\pi_0 = 1$, niin kaikki havainnot tulevat luokitelluiksi vastatapahtumiksi. Tällöin sekä 1−tarkkuus että herkkyys ovat 0. Käyrä päättyy pisteeseen $(1, 1)$: Kun $\pi_0 = 0$, kaikki havainnot luokitellaan tapahtumiksi ja 1−tarkkuus ja herkkyys ovat 1. Pisteiden välillä erottelukykykäyrä kulkee yleensä pisteet yhdistävän suoran yläpuolella. Pisteet suoralla ovat tilanteita, joissa mallilla ei ole selityskykyä (harjoitustehtävä). Malli toimii sitä paremmin, mitä suurempi on herkkyys kullakin 1−tarkkuus -arvolla eli mitä ylempänä käyrä lentelee. Ideaalitilanteessa erottelukykykäyrä singahtaa $(0, 0)$-pisteestä lähes pisteeseen $(0, 1)$ ja kulkee sen jälkeen likipitäen vaakasuorasti pisteeseen $(1, 1)$. Tällöin malli tunnistaa tapahtumat oikein aina, kun π_0 poikkeaa nollasta.

Erottelukykykäyrän alainen pinta-ala (KAP, *area under curve*, *AUC*) mittaa mallin selityskykyä. Mitä suurempi pinta-ala on, sitä paremmin malli selittää y-havaintoja. KAP saa arvoja välillä $[0.5, 1]$.

Erottelukykykäyrän piirtämiseen ja KAPin laskemiseen on monia R-paketteja. Kuva 14.8 on piirretty paketin pROC komentojen `roc`, `plot.roc` ja `auc` avulla. Erottelukykykäyrä on kuvassa porrasmainen, koska havaintoja on vähän. Suuremmilla havaintomäärillä käyrä on yleensä tasaisemmin kaartuva.

Mallin selityskykyä voidaan arvioida myös kaksiarvoisen vastemuuttujan havaintojen ja sovitteiden $\hat{\pi}_i$ korrelaatiokertoimella eli yhteiskorrelaatiokertoimella tai sen neliöllä jakson 13.5.1 tapaan (vrt. kaavat (13.7) ja (13.8)). Yhteiskorrelaatiokertoimen neliötä kutsuttiin selitysasteeksi PNS-menetelmällä estimoidun lineaarisen regressiomallin yhteydessä. Suuretta ei voi tulkita samoin SU-menetelmällä estimoidun logistisen regressiomallin yhteydessä. Vastemuuttujan kaksiarvoisuus voi rajoittaa suuresti yhteiskorrelaatiokertoimen mahdollisia arvoja. Logististen regressiomallien selityskyvyn vertailuun yhteiskorrelaatiokerroin, tai sen neliö, sopii silti.

Esimerkki. Avaruussukkula. Challenger-avaruussukkulan räjähdys 1986 järkytti monia (kuva 14.5). Wikipedia[176]:

> Challenger – – oli yhdysvaltalainen avaruussukkula, joka tuhoutui 73 sekuntia laukaisunsa jälkeen 28. tammikuuta 1986. – – Avaruussukkulan onnettomuuden syyksi paljastui – – apuraketin O-tiivisteen pettäminen. Laukaisua edeltäneen yön pakkasesta johtuen tiiviste oli menettänyt kimmoisuuttaan ja raketin polttoaine pääsi virtaamaan tiivisteen välistä aiheuttaen onnettomuuden. – – Challengerin viimeisellä lennolla oli mukana opettaja Christa McAuliffe, josta piti

> tulla ensimmäinen siviili avaruudessa. – – kaikki seitsemän astronauttia kuolivat.

Englanninkielisen Wikipedian mukaan lämpötila oli onnettomuuspäivänä noin −2 astetta. Avaruussukkuloilla oli tehty 23 onnistunutta lentoa ennen avaruus-

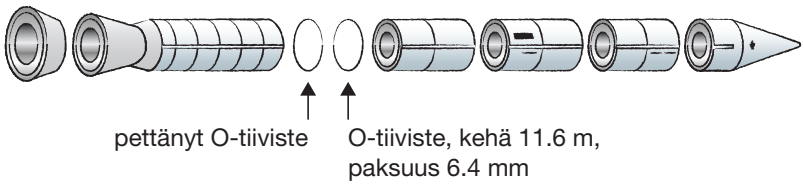
Kuva 14.5: Challenger-sukkulan lähtö 28.1.1986.[1]

sukkulan nousussa hajoamiseen päättynyttä 24. lentoa. Yhdysvaltain ilmailu- ja avaruushallinto NASA oli aiemmin seurannut laukaisuhetken lämpötilan ja avaruussukkulan pää-O-tiivisteiden lämpörasitusta. Kussakin sukkulassa oli kuusi pää-O-tiivisterengasta (jatkossa O-tiivistettä, kuva 14.2.2). Lämpötilat olivat vaihdelleet 12:sta 27 asteeseen.

Tutkitaan logistisella regressiolla O-tiivisteiden lämpörasituksen riippuvuutta lämpötilasta. Määritellään muuttuja, joka saa arvon 1, jos avaruussukkulan

[1]Piirros: Mika Kettunen (2023). Piirros pohjaa NASAn valokuvaan.

[2]Piirros: Mika Kettunen (2023). Piirroksen tiedot ovat Massachusetts Institute of Technologyn professori Daniel Roosin havainnollistuksesta.



Kuva 14.6: Pettäneen O-tiivisteen sijainti sukkulan polttoainetankin viereisessä raketissa.[2]

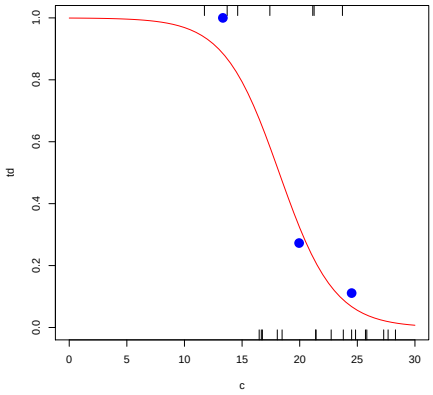
O-tiivisteistä ainakin yhdessä on lämpörasitusvaurio (*suffered thermal distress*) ja 0, jos yhdessäkään ei ole vauriota.[177] Lämpörasitusvaurio ennakoi O-tiivisteen rikkoutumista. SU-menetelmällä estimoitu logistinen regressio on

$$\log \frac{\hat{\pi}(c)}{1-\hat{\pi}(c)} \quad = \quad \underset{(3.933)}{7.614} \quad - \quad \underset{(0.195)}{0.418c}.$$

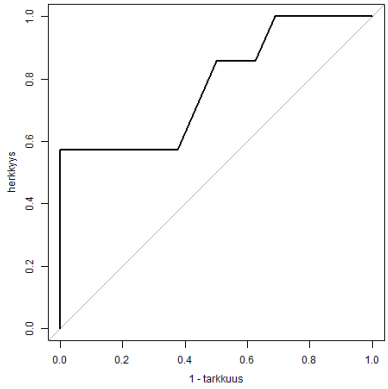
Yllä c on lämpötila, $\hat{\pi}(c)$ on lämpötilasta riippuva estimoitu todennäköisyys ja suluissa ovat estimaattien likimääräiset keskihajonnat. Havaintopareja on 23. Lämpötilamuuttujan kertoimen estimaatin t-arvo on $-0.418/0.195 \approx -2.145$. Vertaaminen normaalijakaumaan tuottaa p-arvon 0.032 (`2*pnorm(-2.145)`). Jos verrataan `t`-jakaumaan, p-arvo on hieman suurempi 0.044 (`2*pt(-2.145,21)`). Pienehkö p-arvo viittaa siihen, että lämpötila vaikuttaa O-tiivisterenkaan lämpörasitusvaurion todennäköisyyteen. Pieni aineisto heikentää testiä.

Kuva 14.7 havainnollistaa O-tiivisteen lämpörasitusvaurion todennäköisyyden riippuvuutta lämpötilasta.[178] Logistinen muunnos takaa, että estimoitu todennäköisyys rajoittuu välille (0,1). Havaintoja on ravistettu, jotta ne erottuisivat toisistaan. Estimointi on tehty alkuperäisillä havainnoilla. Lisäksi aineisto on jaettu kolmeen luokkaan. Siniset täplät ovat luokkakeskiarvoja, jotka ovat itsessään estimaatteja lämpörasitusvaurion todennäköisyydelle. Ne osuvat melko lähelle mallin sovitetta. Sovite kuvaa luokkakeskiarvoja kattavammin todennäköisyyden käyttäytymistä ja on mielekkäämpi kuin lineaarisen todennäköisyysmallin sovite kuvassa 14.7.

Logistisen regression tuloksia on hyvä kuvata laskemalla estimoidusta mallis-



Kuva 14.7: Avaruussukkulan O-tiivisteen lämpörasitusvaurion todennäköisyys.



Kuva 14.8: Avaruussukkulamallin erottelukykykäyrä (KAP = 0.781).

ta todennäköisyyksiä. Lämpötilan estimoitu kerroin on negatiivinen, joten lämpötilan kylmetessä lämpörasitusvaurion todennäköisyys suurenee. Sovitteesta näkyy, että lämpötilavaurion todennäköisyys on lähes 1, jos lämpötila on onnettomuuspäivän -2 astetta. Havaintoaineiston kylmimmällä lämpötilalla 12 astetta vaurion todennäköisyys on noin 0.93:

$$\hat{\pi}(12) = \frac{\exp(7.614 - 0.418 \times 12)}{1 + \exp(7.614 - 0.418 \times 12)} \approx 0.931.$$

Avaruussukkuloiden laukaisujen aikaan lämpötila on ollut keskimäärin noin 20.9 astetta. Estimoitu lämpörasitusvaurion todennäköisyys 20.9 asteessa on noin 0.25:

$$\hat{\pi}(20.9) = \frac{\exp(7.614 - 0.418 \times 20.9)}{1 + \exp(7.614 - 0.418 \times 20.9)} \approx 0.246.$$

Tässä lämpötilassa sovitteen kulmakerroin on noin -0.08 (yhtälöstä (14.4)):

$$\hat{\beta}_1 \hat{\pi}(c)[1 - \hat{\pi}(c)] = -0.418 \times 0.246(1 - 0.246) \approx -0.077.$$

Lämpötilasta riippumatta lämpötilan kasvaessa asteella vastasuhde pienenee

0.66-kertaiseksi, eli ristisuhde on 0.66:

$$\frac{\dfrac{\hat{\pi}(c+1)}{1-\hat{\pi}(c+1)}}{\dfrac{\hat{\pi}(c)}{1-\hat{\pi}(c)}} = \exp(\hat{\beta}_1) = \exp(-0.418) \approx 0.66.$$

Taulukko havainnollistaa, kuinka hyvin malli luokittelee havainnot oikein ($\hat{y} = 1$, jos $y = 1$ tai $\hat{y} = 0$, jos $y = 0$), kun $\pi_0 = 1/5$, $\pi_0 = 7/23 = 0.304$ tai $\pi_0 = 1/2$. Osamäärä $7/23$ on lämpörasitusvauriotapahtumien osuus aineistossa.

	$\pi_0 = 1/5$		$\pi_0 = 7/23$		$\pi_0 = 1/2$		
	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Σ
$y = 1$	6	1	4	3	4	3	7
$y = 0$	8	8	5	11	0	16	16

Herkkyys ja tarkkuus ovat $6/7 = 0.857$ ja $8/16 = 0.500$, jos $\pi_0 = 1/5$, $4/7 = 0.571$ ja $11/16 = 0.688$, jos $\pi_0 = 7/23$ tai $4/7 = 0.571$ ja $16/16 = 1$, jos $\pi_0 = 1/2$. Vastaavat pisteet (1−tarkkuus, herkkyys) ovat $(0.5, 0.857)$, $(0.313, 0.571)$ ja $(0, 0.571)$. Ne osuvat kuvan 14.8 oikean ruudun erottelukykykäyrälle. Mallin KAP on 0.781.

Miten NASA saattoi laukaista sukkulan avaruuteen kylmänä aamuna 28. 1.1986, vaikka O-tiivisteen lämpörasitusvaurion riski kasvaa nopeasti lämpötilan kylmetessä? Edellisenä iltana NASAn asiantuntijat pohtivat, voiko kylmä aamu olla ongelma ja arvioivat, että selvää lämpötilavaikutusta O-tiivisteisiin ei ole. He olivat oleellisesti tutkineet kuvan 14.7 yläosan havaintoja yksin, eivätkä havainneet niissä selvää lämpötilavaikutusta. Uskomaton virhe oli katsoa osa-aineistoa, josta puuttui lennot, joissa ei ollut tapahtunut lämpötilavaurioita. Niiden sisällyttäminen analyysiin olisi paljastanut lämpötilavaikutuksen. Valikoitu otos (jakso 8.4) johti seitsemän ihmisen kuolemaan. Tilastotieteilijät ovat tehneet esimerkin analyysit jälkikäteen. □

Esimerkki. Päätökset lasten asumisesta ja vanhempien ulkomaalaistausta. Äidin ja isän väliset oikeusriidat lapsen asumisesta ovat keskimääräistä yleisempiä avioeroissa, joissa ainakin toinen vanhemmista on ulkomaalaistaustainen. Pere ym. (2017) selittävät äidin voittotodennäköisyyttä riidassa logistisella regressiomallilla.[179] Tutkimuksen pontimena oli ulkomaalaistaustaisten miesten kyselytutkimuksella selvitetty näkemys, että suomalainen oikeuslaitos suosii naisia. Aineisto koostui 846:sta eteläsuomalaisen käräjäoikeuden päätöksestä 2004–2015.

Selitettävä on äidin voittotodennäköisyyden logit-muunnos (voitto = 1; häviö = 0). Selittäjinä ovat osoittimet vanhempien ulkomaalaistaustalle: $x_1 = 1$, jos vain äiti on, $x_2 = 1$, jos molemmat ovat ja $x_3 = 1$, jos vain isä on. Muutoin osoittimet saavat arvon 0. SU-estimointi tuottaa mallin

$$\begin{array}{rcccccccc} \log \dfrac{\hat{\pi}(x_1,\ x_2,\ x_3)}{1-\hat{\pi}(x_1,\ x_2,\ x_3)} & = & 0.272 & + & 0.257x_1 & + & 0.701x_2 & + & 0.771x_3. \\ & & (0.088) & & (0.237) & & (0.229) & & (0.227) \end{array}$$

Estimoitujen kertoimien estimoidut keskivirheet ovat suluissa. Kertoimien β_1, β_2 ja β_3 nolluutta testaavien t-arvojen p-arvot ovat 0.278, 0.002 ja alle 0.001. Isän ulkomaalaistaustalla vaikuttaa olevan merkitystä huoltoriitoja ratkottaessa.

Äidin estimoidut voittotodennäköisyydet eri tilanteissa saadaan sijoittamalla estimoidut kertoimet kaavaan

$$\hat{\pi}(x_1,\ x_2,\ x_3) = \frac{\exp(\hat{\beta}_0 + \sum_{i=1}^{3} \hat{\beta}_i x_i)}{1 + \exp(\hat{\beta}_0 + \sum_{i=1}^{3} \hat{\beta}_i x_i)}$$

ja asettamalla osoittimille kaikille arvo 0 tai yksi kerrallaan arvo 1. Vanhempien ollessa kantasuomalaisia osoitinmuuttujat saavat kaikki arvon 0 ja äidin voittotodennäköisyys määräytyy yksin vakiotermin estimaatista 0.272. Äidin voittotodennäköisyys on tällöin noin 0.57:

$$\hat{\pi}(0,0,0) = \frac{\exp(0.272)}{1+\exp(0.272)} \approx 0.568.$$

Jos vain äiti on ulkomaalaistaustainen, äidin voittotodennäköisyys on noin 0.63:

$$\hat{\pi}(1,0,0) = \frac{\exp(0.272+0.257)}{1+\exp(0.272+0.257)} \approx 0.629.$$

Ero kantasuomalaisen äidin voittotodennäköisyyteen ei ole tilastollisesti merkitsevä p-arvon edellä mukaan. Jos molemmat vanhemmat ovat ulkomaalaistaustaisia, äidin voittotodennäköisyys nousee noin 0.73:een:

$$\hat{\pi}(0,1,0) = \frac{\exp(0.272+0.701)}{1+\exp(0.272+0.701)} \approx 0.726.$$

Jos yksin isä on ulkomaalaistaustainen, äidin voittotodennäköisyys on noin 0.74:

$$\hat{\pi}(0,0,1) = \frac{\exp(0.272+0.771)}{1+\exp(0.272+0.771)} \approx 0.739.$$

Todennäköisyydet kasvavat 0.57:stä 0.74:ään. Mallin mukaan ulkomaalaistaustaiset isät pärjäävät riidoissa lasten asumisesta huonosti. Malli tarjoaa yhden selityksen ulkomaalaistaustaisten miesten näkemykselle naisten suosimisesta oikeuslaitoksessa.

Mallista voi laskea ristisuhteita. Esimerkkinä muodostetaan ristisuhde vastasuhteista ulkomaalaistaustaisten ja kantasuomalaisten vanhempien tilanteissa:

$$\frac{\frac{\hat{\pi}(0,1,0)}{1-\hat{\pi}(0,1,0)}}{\frac{\hat{\pi}(0,0,0)}{1-\hat{\pi}(0,0,0)}} = \frac{\frac{0.726}{1-0.726}}{\frac{0.568}{1-0.568}} \approx 2.016.$$

Ristisuhde on noin 2. □

14.2.3 Risti- ja riskisuhteen ero

Ristisuhde θ ei ole kovin intuitiivinen suure eikä aivan helppo tulkittava. Ei ole harvinaista, että ristisuhde sekoitetaan *riskisuhteeseen* (*risk ratio*). Verrattaessa kahta todennäköisyyttä $\pi_1 \in (0,1)$ ja $\pi_0 \in (0,1)$ ristisuhde on

$$\theta = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}.$$

Helpommin hahmotettava riskisuhde on

$$\lambda = \frac{\pi_1}{\pi_0}.$$

Sellaisen voi estimoida logistisen regression estimointitulosten avulla, mutta regressiokertoimien estimaatteihin riskisuhde ei ylipäänsä liity suoraan.

Jos todennäköisyydet ovat yhtäsuuria, niin risti- ja riskisuhde yhtyvät: $\theta = \lambda = 1$. Jos todennäköisyydet ovat hyvin pieniä, niin osamäärä $(1-\pi_0)/(1-\pi_1)$ on noin 1. Tällöin risti- ja riskisuhde ovat likimain samoja:

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{\pi_1}{\pi_0} \times \frac{1-\pi_0}{1-\pi_1} \approx \frac{\pi_1}{\pi_0} = \lambda.$$

Likiarvoistuksen $\theta \approx \lambda$ toimivuuden peukalosääntö on, että todennäköisyydet ovat pienempiä kuin 0.1. Muulloin voi olla vaarallista sekoittaa risti- ja riskisuhde. Kaavasta yllä nähdään myös, että ristisuhde on suurempi kuin riskisuhde, jos $\pi_1 > \pi_0$.

Olkoon $\pi_1 > \pi_0 > 0.2$. Jos lähteessä raportoidaan risti- muttei riskisuhde eikä todennäköisyyksiä, niin riskisuhteesta voi tällöin rykäistä ensiapuarvion VanderWeelen (2017) kaavalla:

$$\lambda \approx \sqrt{\theta}.$$

Jos aineisto on käytettävissä, todennäköisyydet ja riskisuhde tulee estimoida aineistosta eikä käyttää VanderWeelen likiarvoa.

Esimerkki. Risti- ja riskisuhteen erot ja riskisuhteen approksimointi. Olkoon $(\pi_1, \pi_0) = (0.03, 0.01)$, $(\pi_1, \pi_0) = (0.3, 0.1)$ tai $(\pi_1, \pi_0) = (0.69, 0.23)$. Kaikissa tilanteissa riskisuhde $\lambda = 3$ on vakio mutta ristisuhde θ vaihtelee välillä $(3, 7.5)$: $\theta = 3.062$, $\theta = 3.857$ tai $\theta = 7.452$. Ensimmäisessä tilanteessa todennäköisyydet ovat hyvin pieniä, jolloin $\lambda \approx \theta$:

$$3 \approx 3.062.$$

Toisen todennäköisyysparin kohdalla ei päde kumpikaan likiarvoistus, koska todennäköisyydet eivät ole riittävän pieniä tai suuria ($3 \neq 3.857$ eli $\lambda \neq \theta$ ja $3 \neq \sqrt{3.857} \approx 1.964$ eli $\lambda \neq \sqrt{\theta}$.) Kolmannessa tilanteessa ehto $\pi_1 > 0.2$ ja $\pi_2 > 0.2$ pätee ja VanderWeelen likiarvo $\lambda \approx \sqrt{\theta}$ antaa karkean käsityksen riskisuhteesta:

$$3 \approx \sqrt{7.452} = 2.730.$$

□

Luku 15

Parametrittomia menetelmiä

> Tilastotiede puhaltaa matematiikkaan elämän.[180]
> Eugene Demidenko (1948–)

Eniten käytetyt tilastolliset menetelmät ovat *parametrisia*. Niissä jakaumaoletusten ja otantamenetelmän (tilastollisen mallin) perusteella estimoidaan populaatiota kuvaavia parametreja (jakso 6.2) ja tehdään aineistosta tilastollisia päätelmiä populaatiosta. *Parametrittomia menetelmiä* (*nonparametric*) saattaisi kuvitella vastakohdaksi parametrittomille menetelmille, mutta käsitteet ovat limittäiset.

Parametrittomiksi menetelmiksi voisi ajatella menetelmiä, jotka eivät milläänlailla tukeudu populaatiota kuvaaviin parametreihin, ja niin joskus määritellään. Tällainen menetelmä on esimerkiksi kahden empiirisen jakauman vertailu χ^2-testillä (jakso 12.2.2). Testi tavataan silti esittää oppikirjoissa parametristen testien yhteydessä. Toisaalta parametrittomana menetelmänä järjestään pidetään testiä kahden odotusarvon yhtäsuuruudelle, kun ei tehdä oletusta tietystä otokset tuottaneesta jakaumasta. Kahta parametria ollaan silti tutkimassa.

Parametriton menetelmä ei ole välttämättä tarkasti rajattu käsite, mutta sen ydinajatus on selkeä: tilastollinen menetelmä lievemmillä oletuksilla. Vähäoletuksisempi (*assumption freer*) voisi olla osuvampi käsite kuin parametriton (Sheskin 2014), mutta jälkimmäinen termi dominoi. Parametritonta menetelmää käytetään, jos tarkasti parametroitu malli, tietty jakaumaoletus, välimatka- tai jopa järjestysasteikko tuntuvat liian rajoittavilta oletuksilta, ylipäänsä halutaan

varautua moninaisiin tilanteisiin tai jos parametrinen tilastollinen päättely vaatisi hyvin monimutkaista (mahdollisesti vielä selvittämätöntä) matematiikkaa.

Esimerkki. Jaksossa 8.2 sivuttiin pistetodennäköisyysfunktion, tiheysfunktion ja kertymäfunktion estimointia otospistetodennäköisyysfunktiolla, histogrammilla ja otoskertymäfunktiolla. Estimoitavaa jakaumaa ei parametroitu mitenkään, eikä parametreja estimoitu. Mainitut estimointimenetelmät ovat parametrittomia. □.

Parametrittomalle menetelmälle läheinen käsite on *jakaumasta riippumaton* (*distribution free*) menetelmä. Sellaisen tilastolliset ominaisuudet eivät riipu jakaumasta, josta aineisto on muodostunut.

Parametrittoman ja jakaumasta riippumattoman menetelmän ero ei ole aina kirkas. Parametrittomat ja jakaumasta riippumattomat menetelmät saatetaan käytännössä samaistaa (Ugarte ym. 2016, 587). Jakaumasta riippumattomat menetelmät voidaan hahmottaa parametrittomien menetelmien osajoukoksi (Meeker ym. 2017, 186). Ne voidaan silti kuvata hengeltään aivan erilaisiksi (Dickinson Gibbons ja Chakraborti 2021, 3).

Esimerkki. Parametrinen, parametriton vai jakaumasta riippumaton. Onko tilastollinen päättely suurilla havaintomäärillä keskeiseen raja-arvolauseen perustuen parametrinen, parametriton vai jakaumasta riippumaton menetelmä vai niitä kaikkia? Keskeisen raja-arvolauseen mukaan riippumattomien satunnaismuuttujien keskiarvo noudattaa suurilla havaintomäärillä normaalijakaumaa $\mathsf{N}(\mu, \sigma^2/n)$ satunnaismuuttujan jakaumasta riippumatta, kunhan satunnaismuuttujalla on odotusarvo (μ) ja varianssi (σ^2) (jakso 7.3). Menetelmä ei oleta tiettyä jakaumaa mutta olettaa kaksi parametria, ja päättely perustuu niihin tai niiden estimaatteihin. Menetelmän voi hahmottaa parametriseksi vaikkakin pitkälti jakaumasta riippumattomaksi. □.

Parametrittomien menetelmien etu on niiden yleispätevyys. Monet vaativat vain järjestysasteikolliset havainnot, ne saattavat toimia satunnaismuuttujan jakaumasta riippumatta tai ne voivat tunnistaa lineaarisia monimuotoisempia yhteyksiä. Parametrittomat menetelmät ovat tyypillisesti myös vakaita erilaisten poikkeavuuksien suhteen. Jaksossa 9.1 selitettiin vakaa estimaattori. Tilastollinen menetelmä on vastaavasti vakaa tietyn poikkeavuuden suhteen, jos se ei tapaa muuttaa menetelmän tuloksia suuresti. Menetelmä voi olla vakaa esimerkiksi oudokkien suhteen. Myös parametrinen menetelmä voi olla vakaa.

Esimerkki. Studentin t-testi. Studentin t-testi on vakaa satunnaismuuttujan jakauman suhteen. Vaikka normaalijakaumaoletus ei pätisi, t-jakauma voi olla

hyvä likiarvoistus t-testisuureen nollajakaumalle. Erityisesti t-jakaumaan nojautuvat kaksisuuntaiset testit ja luottamusvälit voivat toimia tällöinkin hyvin. Agresti ja Finlay (2009, 122, vrt. 155) arvioivat, että jo 15 havaintoa voi riittää likiarvoistuksen toimivuudelle luottamusväliä laskettaessa.[181] Testi ei ole vakaa havaintojen riippumattomuuden suhteen. □.

Parametrittomien menetelmien yleispätevyyden hinta on, että ne tapaavat olla joissain suhteissa vastaavia parametrisia menetelmiä huonompia. Parametrittoman menetelmän toteuttaminen voi olla jopa mahdotonta, vaikka vastaava parametrinen menetelmä olisi mahdollinen (esim. luottamusväli tietyllä luottamustasolla pienellä havaintomäärällä). Toisaalta parametriton menetelmä voi toimia paremmin kuin parametrinen, jos tutkittava ilmiö ei ole parametrisen menetelmän olettaman kaltainen.

Yksi näkemys on, että parametriton-käsite lokeroi tarpeettomasti menetelmiä, ja siitä tulisi luopua (Noether 1984). Kyse on yksinkertaisesti yleispätevistä tilastotieteellisistä menetelmistä.

15.1 Parametrittomia testejä

Jaksossa osoitetaan testejä, jotka perustuvat lievemmille oletuksille kuin likeiset normaalijakauman olettavat testit (luku 12). Kaikilla testataan keskiluvun tai keskilukujen erotuksen suuruutta. Yksi jakson uutuuksista on epäsymmetrisen jakauman mediaanin testaus.

15.1.1 Merkkitesti

Tutkittava aineisto $x_1, \ldots, x_n$ on riippumattomia havaintoja jatkuva-arvoisista satunnaismuuttujista X_i, joiden mediaani on M. Havaintojen ei tarvitse olla samasta jakaumasta! Riittää, että pätee $\mathsf{P}(X_i > M) = \mathsf{P}(X_i < M) = 1/2, i = 1, \ldots, n$. Testin tarkentuvuus (jakso 11.1) saattaa kuitenkin vaatia, että havainnot ovat samasta jakaumasta (Hollander ym. 2014, 73). Jatkuva-arvoisuusoletuksesta luovutaan jakson lopussa.

Merkkitestillä (*sign test*) testataan nollahypoteesia, että mediaani on M_0. Testin nimi tulee siitä, että lasketaan plussia (ja miinuksia) siitä, kuinka moni havainnosta on suurempia (tai pienempiä) kuin M_0. Testisuure on plus-merkkien lukumäärä (s). Koska plussa on tapahtuma $X_i > M_0$, niin summaa s vastaavan satunnaismuuttujan S nollajakauma on binomijakauma: $S \sim \mathsf{Bin}(n, 0.5)$. Testi voitaisiin perustaa siihen. (Jakso 12.1.) Erityisen kätevää on tehdä testi

standardoidun testisuureen

$$\frac{S/n - 0.5}{\sqrt{0.5(1-0.5)/n}}$$

avulla. Nollahypoteesin pätiessä se noudattaa standardinormaalijakaumaa riittävän suurilla havaintomäärillä (jakso 9.3). Approksimaatio toimii yleensä hyvin, jos $n > 20$ (Agresti ja Finlay 2009, 156, 172, Ugarte ym. 2016, 589) ja mahdollisesti jopa jos $n \geq 12$ (Dickinson Gibbons ja Chakraborti 2020, 181). Testisuure hälyttää, jos s poikkeaa paljon binomijakauman mukaisesta odotusarvosta $0.5n$ eli s/n 0.5:stä. Testaus sujuu näppärästi BSDA-paketin `SIGN.test(x,md=m)`-komennolla, jossa `x` koostuu tutkittavan muuttujan havainnoista ja `m` on nollahypoteesin mukainen lukuarvo mediaanille.

Samaan tapaan voidaan testata kvantiilin q suuruutta (H_0: $q = q_0$). Oletus on tällöin, että kaikille havainnoille pätee $\mathsf{P}(X_i > q) = 1 - \mathsf{P}(X_i < q)$. Nyt s on havaintojen lukumäärä, jotka ovat suurempia kuin q_0, S:n nollajakauma on $\mathsf{Bin}(n, q_0)$ ja standardoitu testisuure on $(S/n - q_0)/[q_0(1-q_0)/n)]^{1/2}$. Normaalisuusapproksimaatio vaatii suuremman havaintomäärän kuin testattaessa nollahypoteesia mediaanista.

Tasahavainnot ovat mahdollisia pyöristysten takia, vaikka satunnaismuuttuja olisi jatkuva-arvoinen. Jos testattavalle parametriarvolle osuu havaintoja, niin tyypillisesti suositeltu menettely on poistaa ne aineistosta ennen testisuureen laskua. Havaintojen lukumäärä testissä on tällöin otoksen koko vähennettynä kyseisten tasahavaintojen lukumäärällä.

Esimerkki. Osakkeiden tuotto (jatkoa). Suomalaisen osakeindeksin tuoton 31. 10.1912–31.8.2022 histogrammi piirrettiin kuvassa 8.2. Usein tuottojen jakaumien hännät ovat paksumpia kuin normaalijakaumalla ja tuottojen vaihtelu on rypäyksittäistä. Osakkeiden tuottojen varianssi ei siten liene riippumaton edellisistä havainnoista. Toisaalta tavanomaisen talousteorian mukaan osakkeiden tuotot ovat riippumattomia aiemmista tuotoista. Merkkitesti luovii tilanteessa. Se peräti sallii jokaisen havainnon olevan peräisin eri jakaumasta — kunhan pätee, että $\mathsf{P}(X_i > M) = \mathsf{P}(X_i < M) = 1/2$, jossa X_i on tuotto.

Testataan nollahypoteesia, että mediaanituotto on 0 ($M_0 = 0$). Havaintoja on 1 318. Niistä 15 on arvoltaan tuoton tyyppiarvo 0 (harjoitustehtävä). Niiden poistamisen jälkeen havaintoja on 1 303. Testisuureen arvo on suuri 6.621:

$$\frac{771/1303 - 0.5}{\sqrt{0.5(1-0.5)/1303}} = 6.621.$$

Sen p-arvo poikkeaa nollasta 11. desimaalissa (`2*pnorm(-6.621032)`, kaksisuuntainen testi). Luku 771 on `SIGN.test`-komennon laskema satunnaismuuttujan

S toteuma aineistossa. (Komento palauttaa niin ikään 11. desimaalissa nollasta poikkeavan p-arvon.) On syytä hylätä nollahypoteesi 0-mediaanista. Otosmediaani on 0.008 (harjoitustehtävä), eli osakeindeksin mediaanituoton estimaatti on noin 0.8 % kuukaudessa.

Tuoton ja edeltävän ajankohdan tuoton korrelaatio on 0.20, joten riippumattomuusoletus ei pätene eikä aineisto ole talousteorian ennustamanlainen. Tämä seikka sivuutetaan esimerkissä. □

Merkkitestiä voi soveltaa myös parittaisen t-testin (jakso 12.4.7) parametrittomana vastineena mediaanin testaamiseen. Aineisto on nyt erotuksia $z_1 = x_1 - y_1, \ldots, z_n = x_n - y_n$, joissa x_i ja y_i ovat satunnaismuuttujien X_i ja Y_i toteumia. Satunnaismuuttujien ei tarvitse olla riippumattomia, mutta erotusten $Z_i = X_i - Y_i$, $i = 1, \ldots, n$, tulee olla. Havaintojen z_i ei tarvitse tulla samasta jakaumasta, mutta tulee päteä, että $\mathsf{P}(Z_i > M) = \mathsf{P}(Z_i < M) = 1/2, i = 1, \ldots, n$. Koeteltava nollahypoteesi on, että satunnaismuuttujien Z_i mediaani on M_0.

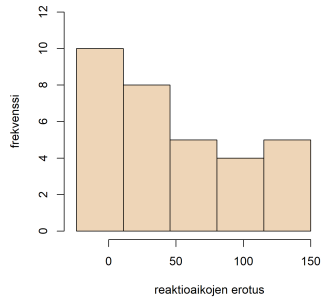
Testin kulku on sama kuin yhden otoksen tilanteessa. Lisäkommervenkki on, että mikäli nollahypoteesi erotuksen mediaanista M_0 hylätään, niin siitä ei voi yksioikoisesti päätellä, että satunnaismuuttujien X_i ja Y_i mediaanit eroavat. Niin voidaan päätellä, mikäli oletetaan, että satunnaismuuttujien X_i ja Y_i jakaumat ovat symmetrisiä ja samoja (Hutson ja Yu 2023). Tällöin testataan itse asiassa odotusarvojen eroa, koska symmetrisen jakauman tilanteessa mediaani on odotusarvo.

Esimerkki. Kännykkään puhuminen ja reaktioaika (jatkoa jaksoista 12.4.6–12.4.7). Koehenkilöiden reaktioaikaa verrattiin heidän puhuessaan puhelimeen tai kuunnellessaan radio-ohjelmaa tai äänikirjaa. Kunkin koehenkilön reaktioaika mitattiin molemmissa tilanteissa, eli havainnot ovat parittaisia ($n_1 = n_2 = 32$). Jaksossa 12.4.7 arvioitiin, että reaktioaikojen erotus ei ole ehkä normaalijakautunut. Kuvan 15.1 histogrammi tukee arviota. Merkkitestin käyttö voisi olla perusteltua. Merkkitestisuure on nyt

$$\frac{26/32 - 0.5}{\sqrt{0.5(1-0.5)/32}} = 3.536.$$

Sen p-arvo eroaa nollasta neljännessä desimaalissa (`2*pnorm(-3.535534)`). Luku 26 on `SIGN.test`-komennon palauttama satunnaismuuttujan S toteuma. (Komento laskee niinikään neljännessä desimaalissa nollasta poikkeavan p-arvon.) R palauttaa keskenään täsmälleen samat tulokset komennoilla `SIGN.test(x1,x2,md=0)` ja `SIGN.test(d,md=0)`. Muuttujat `x1` ja `x2` ovat jakson 12.4.7 merkinnät reaktioajoille, ja `d` on erotus `x1-x2`. Päätellään, että

reaktioaikojen erotuksen mediaani poikkeaa nollasta. Sen estimaatti on 35.5 (`SIGN.test`-komennon palautteesta). Kännykkään puhujien reaktioajat ovat reaktioaikojen erotuksen mediaanilla mitattuna pidempiä kuin radion tai äänikirjojen kuuntelijoiden. Johtopäätös on sopusoinnussa päätelmien jaksoissa 12.4.6–12.4.7 kanssa. □



Kuva 15.1: Reaktioaikojen erotuksen jakauma.

Merkkitestiä voi käyttää, vaikka satunnaismuuttuja X_i olisi järjestysasteikollinen. Tällöin tasahavainnot testattavalla parametriarvolla ovat erityisen mahdollisia. Niiden sivuuttamisen jälkeen oletukset $\mathsf{P}(X_i > M) = \mathsf{P}(X_i < M) = 1/2$ tai $\mathsf{P}(X_i > q) = 1 - \mathsf{P}(X_i < q)$ soveltuvat myös järjestysasteikolliseen tilanteeseen. Jos testattavalla parametriarvolla on suuri todennäköisyys, testin tulkintaan tulee kiinnittää erityistä huomiota. Suuren testisuureen arvon tilanteessa asianmukainen päätelmä voi olla pikemminkin, että $\mathsf{P}(X_i > M_0) > \mathsf{P}(X_i < M_0)$ tai $\mathsf{P}(X_i > q_0) > 1-\mathsf{P}(X_i < q_0)$ kuin että mediaani ei ole M_0 tai että q. kvantiili ei ole q_0. Vastaavia tarkennuksia tarvitaan, jos merkkitestiä sovelletaan järjestysasteikollisiin erotuksiin $Z_i = X_i - Y_i$.

15.2 Kertymäfunktioiden väliestimointi ja testaus

Otoskertymäfunktio $\hat{F}(x)$ (jakso 8.2) on yksi eniten käytetyistä ja joustavimmista parametrittomista menetelmistä. Otoskertymäfunktio visualisoi aineistoa ja muotoutuu lähes millaiseksi kertymäfunktioksi $F(x)$ (jakso 6.1) vain. Jaksossa huomio on kertymäfunktion väliestimoinnissa ja kahden kertymäfunktion yhtenevyyden testaamisessa.

15.2.1 Kertymäfunktion väliestimointi

Sallitaan aluksi, että satunnaismuuttuja X_i on diskreetti- tai jatkuva-arvoinen. X_i:n toteumat $x_1, \ldots, x_n$ ovat satunnaisotos.

Todennäköisyys, että satunnaismuuttuja X_i saa pienemmän tai yhtäsuuren arvon kuin x, on $F(x)$ (x on kiinteä reaaliluku). Satunnaismuuttuja X_i on siis Bernoulli-jakautunut parametrilla $\pi = F(x)$ (jakso 7.1.1). Merkitään $Y = n\hat{F}(x)$:llä lukumäärää havaintoja, jotka ovat pienempiä tai yhtäsuuria kuin x (jakso 8.2). Koska havainnot ovat riippumattomia, Y on binomijakautunut samalla parametrilla $\pi = F(x)$, eli $Y \sim \mathsf{Bin}(n, F(x))$ (jakso 7.1.3). Binomijakautuneisuudesta seuraa, että $\mathsf{E}(Y) = nF(x)$ ja $\mathsf{V}(Y) = nF(x)[1 - F(x)]$. Koska $\hat{F}(x) = Y/n$, niin $\mathsf{E}(\hat{F}(x)) = F(x)$ ja $\mathsf{V}(\hat{F}(x)) = F(x)[1 - F(x)]/n$ (jaksot 6.3 ja 7.1.3). Otoskertymäfunktio on siten kertymäfunktion harhaton estimaattori (jakso 9.1) pisteessä x ja niin, että sen varianssi menee nollaan otoskoon n kasvaessa kohti ääretöntä. Näin ollen otoskertymäfunktio on kertymäfunktion tarkentuva estimaattori pisteessä x (jakso 9.1). Samaa argumenttia voidaan soveltaa jokaisen pisteen x kohdalla. Otoskertymäfunktio $\hat{F}(x)$ on siten kertymäfunktion $F(x)$ tarkentuva estimaattori. Tulosta (sen matemaattisesti kehittyneempää versiota) on kutsuttu tilastotieteen peruslauseeksi (*Fundamental theorem of statistics*) ja kuvattu niin perustavanlaatuiseksi, että kaikki tilastollinen teoria perustuu siihen (Dudewicz ja Mishra 1988, 305, Stuart ja Ord 1991, 1188).

Otoskertymäfunktion arvo pisteessä x, $\hat{F}(x) = Y/n$, on keskiarvo riippumattomista samoin jakautuneista Bernoulli-satunnaismuuttujista (jakso 7.4.3). Keskeisestä raja-arvolauseesta seuraa, että $\hat{F}(x)$ on kunkin x-arvon kohdalla normaalijakautunut suurilla otoskoilla (jaksot 7.3 ja 7.4.3):

$$\hat{F}(x) \sim \mathsf{N}(F(x), F(x)[1 - F(x)]/n).$$

Tulos mahdollistaa kertymäfunktion luottamusvälin laskemisen mielivaltaisessa pisteessä x. Koska ollaan estimoimassa osuutta $0 \leq F(x) \leq 1$, voidaan soveltaa jakson 10.2.1 luottamusvälejä osuudelle. Waldin menetelmä (10.2) tuottaa $100 \times (1 - \alpha)$ %:n luottamusvälin rajoiksi

$$\hat{F}(x) \pm \mathsf{z}_{1-\alpha/2} \sqrt{\frac{\hat{F}(x)[1 - \hat{F}(x)]}{n}} \tag{15.1}$$

pisteessä x. Väli voidaan laskea eri x:n arvoille, jolloin saadaan käsitys kertymäfunktion sijainnista ja sen estimoinnin tarkkuudesta ylipäänsä. Näin saadaan *pisteittäinen luottamusvyöhyke* (*pointwise confidence band*, *pointwise confidence*

envelope). Jaksossa 10.2.1 kuvatut ongelmat Waldin luottamusvälin peittävyyden tarkkuudesta (pienillä tai suurilla todennäköisyyksillä pienillä havaintomäärillä) pätevät myös tässä yhteydessä. Luottamusväli ja pisteittäinen luottamusvyöhyke voi kannattaa laskea jakson 10.2.1 rukkaavilla menetelmillä.

Pisteittäiseen luottamusvyöhykkeeseen ei liity samaa todennäköisyyttä kuin yksittäiseen luottamusväliin. *Samanaikaiseen luottamusvyöhykkeeseen* (*simultaneous confidence band*) liittyy. Jälkimmäinen saatetaan siksi haluta laskea edellisen sijaan.

Esimerkki. Monen luottamusvälin ja otoskertymäfunktion pisteittäisen luottamusvyöhykkeen tulkinta. Lasketaan kymmenen riippumatonta 95 %:n luottamusväliä kymmenelle eri parametrille. Ne kaikki peittävät tutkittavat parametrit todennäköisyydellä $0.95^{10} = 0.599$. Luottamusvälien samanaikainen parametrien peittämistodennäköisyys on aivan eri kuin yksittäisten luottamusvälien luottamustaso.

Otoskertymäfunktion pisteittäinen luottamusvyöhyke koostuu tyypillisesti lukuisista luottamusväleistä. Ne eivät ole riippumattomia, koska ne on laskettu osin samoista havainnoista. Pisteittäiseen luottamusvyöhykkeeseen liittyvää samanaikaista kertymäfunktion arvojen peittämistodennäköisyyttä on siten ylipäänsä vaikea laskea. Oleellisinta on ymmärtää, ettei se ole yksittäisen luottamusvälin luottamustaso. □

Kertymäfunktiolle voidaan laskea samanaikainen luottamusvyöhyke. Idea on, että se peittää kaikki kertymäfunktion arvot osoitetulla todennäköisyydellä.

Oletetaan, että satunnaismuuttuja X on jatkuva-arvoinen. *Yhden otoksen Kolmogorovin–Smirnovin tunnusluku*

$$D_n = \sup_{-\infty<x<\infty} |\hat{F}(x) - F(x)|$$

on otoskertymäfunktion ja kertymäfunktion erotuksen (pystysuorassa suunnassa) suurin arvo reaalilukujen joukossa.[182] Sen hakemista osoitetaan yllä supmerkinnällä (*supremum*). Andrei Kolmogorov on osoittanut, että suurilla otoskoilla D_n:n jakauman $(1-\alpha)$. kvantiili $d_{1-\alpha}$ on noin

$$d_{1-\alpha} \approx \frac{z^*_{1-\alpha}}{\sqrt{n}}, \tag{15.2}$$

jossa $z^*_{1-\alpha}$ määräytyy taulukosta alla:

α	0.20	0.10	0.05	0.02	0.01
$z^*_{1-\alpha}$	1.073	1.224	1.358	1.517	1.628

Jakaumalikiarvoistuksen toimivuus edellyttää, että otoskoko n on (lähteestä riippuen) vähintään 35–40 (Hollander ym. 2014, 570, Dickinson Gibbons ja Chakraborti 2021, 123, 605). Kvantiili $d_{0.95}$ voidaan laskea kahden desimaalin tarkkuudella kaavalla

$$d_{0.95} = \frac{1.358}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}} \tag{15.3}$$

pienille otoskoille alkaen otoskoosta $n = 2$. Tulokset ovat mainioita: Tunnusluvun D_n jakauma tunnetaan, vaikka $F(x)$ olisi tuntematon!

Esimerkki. On laskettu otoskertymäfunktio satunnaisotoksesta ($n = 100$) jatkuva-arvoisesta satunnaismuuttujasta X. Todennäköisyys, että otoskertymäfunktio $\hat{F}(x)$ poikkeaa millään x-arvolla kertymäfunktion arvosta $F(x)$ enemmän kuin $1.358/\sqrt{100} \approx 0.136$:lla on kaavan (15.2) mukaan 0.05. Kaava (15.3) tuottaa 0.95. kvantiiliksi $1.358/(\sqrt{100}+0.12+0.11/\sqrt{100}) \approx 0.134$. Kvantiilien likiarvot ovat varsin samat $0.136 \approx 0.134$. Jälkimmäinen likiarvo on luotettavampi. □

Tunnusluvun ja sen jakauman avulla voidaan muodostaa samanaikainen luottamusvyöhyke kertymäfunktiolle $F(x)$, vaikka kertymäfunktiota $F(X)$ ja siten erotusta $\hat{F}(x) - F(x)$ ja tunnuslukua D_n ei tunneta. Kertymäfunktion $100 \times (1 - \alpha)$ %:n samanaikainen luottamusvyöhyke toteuttaa samanaikaisesti kaikilla x arvoilla yhtälöt (vrt. jakso 11.3)

$$\begin{aligned}
\mathsf{P}[\,|\hat{F}(x) - F(x)| \le d_{1-\alpha}] = 1 - \alpha \quad &\Leftrightarrow \\
\mathsf{P}[-d_{1-\alpha} \le \hat{F}(x) - F(x) \le d_{1-\alpha}] = 1 - \alpha \quad &\Leftrightarrow \\
\mathsf{P}[-\hat{F}(x) - d_{1-\alpha} \le -F(x) \le -\hat{F}(x) + d_{1-\alpha}] = 1 - \alpha \quad &\Leftrightarrow \\
\mathsf{P}[\hat{F}(x) - d_{1-\alpha} \le F(x) \le \hat{F}(x) + d_{1-\alpha}] = 1 - \alpha.&
\end{aligned}$$

Kertymäfunktion arvot ovat välillä $[0, 1]$, joten määritellään luottamusvyöhykkeen ala- ja ylärajoiksi (L ja U)

$$L = \begin{cases} \hat{F}(x) - d_{1-\alpha}, & \text{jos } \hat{F}(x) - d_{1-\alpha} \ge 0 \text{ ja} \\ 0, & \text{jos } \hat{F}(x) - d_{1-\alpha} < 0 \end{cases}$$

ja

$$U = \begin{cases} \hat{F}(x) + d_{1-\alpha}, & \text{jos } \hat{F}(x) - d_{1-\alpha} \le 1 \text{ ja} \\ 1, & \text{jos } \hat{F}(x) + d_{1-\alpha} > 1. \end{cases}$$

Näin alaraja ei voi olla 0:aa pienempi ja yläraja 1:htä suurempi.

Samanaikaisen luottamusvyöhykkeen lasku kertymäfunktiolle on jakaumasta riippumaton menetelmä, jos kertymäfunktio on jatkuva. Jos satunnaismuuttuja ei ole jatkuva, niin luottamusvyöhyke voi olla konservatiivinen eli liian leveä (Noether 1967, 17–18.)

Esimerkki. Vihkimiset 2013 II. Kuvassa 8.1 (jakso 8.2) on kirkollisten vihkimisten ($n = 12\,410$) ja siviilivihkimisten ($n = 8\,878$) päivittäiset lukumäärät ja otoskertymäfunktiot. Lasketaan pisteittäiset ja samanaikaiset luottamusvyöhykkeet kirkollisten vihkimisten ja siviilivihkimisten kertymäfunktioille. Vihkipäivä on satunnaismuuttuja.

Pisteittäiset luottamusvälit ovat kapeita, koska havaintoja on paljon. Esimerkiksi päivänä 60 (1.3.2013) kertymäfunktioiden 99 %:n Waldin luottamusvälin rajat ovat kirkollisille vihkimisille

$$0.0738114424 \pm 2.576 \times \sqrt{\frac{0.0738114424 \times (1 - 0.0738114424)}{12410}} \approx \begin{cases} 0.080 \\ 0.068 \end{cases}$$

ja siviilivihkimisille

$$0.133138094 \pm 2.576 \times \sqrt{\frac{0.133138094 \times (1 - 0.133138094)}{8878}} \approx \begin{cases} 0.144 \\ 0.124. \end{cases}$$

(Kaava (15.1).) Luottamusvälien leveydet ovat 0.012 ja 0.020. Leveimmillään luottamusvälit ovat 6205. ja 4439. havaintojen kohdilla, joilla otoskertymäfunktion arvo on 0.5 (jakso 10.2.1). Leveydet ovat tuolloin $2 \times 2.576 \times \sqrt{0.5^2/12410} \approx 0.023$ ja $2 \times 2.576 \times \sqrt{0.5^2/8878} \approx 0.027$.

Kirkollisille vihkimisille ja siviilivihkimisille 0.99. kvantiilit $d_{0.99}$ ovat

$$d_{0.99} \approx \frac{1.628}{\sqrt{12410}} = 0.01461398 \quad \text{ja} \quad d_{0.99} \approx \frac{1.628}{\sqrt{8878}} = 0.01727813.$$

Päivän 60 kohdalla 99 %:n samanaikaisten luottamusvyöhykkeiden rajat ovat

$$0.0738114424 \pm 0.01461398 \approx \begin{cases} 0.088 \\ 0.059 \end{cases}$$

(kirkolliset vihkimiset) ja

$$0.133138094 \pm 0.01727813 \approx \begin{cases} 0.150 \\ 0.116 \end{cases}$$

(siviilivihkimiset). Kertymäfunktioiden 99 %:n samanaikaisten luottamusvyöhykkeiden leveydet ovat $2 \times 0.01463193 \approx 0.029$ ja $2 \times 0.01729936 \approx 0.035$.

Samanaikaiset luottamusvyöhykkeet ovat leveämpiä kuin pisteittäiset luottamusvyöhykkeet leveimmillään (0.023 ja 0.027).

Luottamusvyöhykkeet ovat likimääräisiä suuresta havaintomäärästä huolimatta, sillä vihkimisten ajankohta on diskreetti satunnaismuuttuja — vaikkakin varsin tiheästi luokiteltu. Diskreettiyden vuoksi samanaikainen luottamusvyöhyke tapaa olla liian leveä. Suuren havaintomäärän takia liian leveäkin vyöhyke on kapea.

Pisteittäinen tai samanaikainen luottamusvyöhyke piirretään monesti havainnollisesti otoskertymäfunktion ympärille. Esimerkissä vyöhykkeet ovat niin kapeita, että ne eivät erottuisi selkeästi kuvassa 8.1. Harjoitustehtävässä pohditaan tilastollista päättelyä esimerkissä. □

Esimerkki. Tutkimusviittaukset I.[183] Havainnollistetaan pienellä aineistolla otoskertymäfunktion porrasmaisuutta ja kertymäfunktion samanaikaisen luottamusvyöhykkeen laskua.

Yksittäinen tutkimus ei useinkaan riitä vakuuttamaan tutkijoita. Tulos saa lisäpontta, jos se toistuu riippumattomassa tutkimuksessa. *Nature Human Behaviour* -lehdessä 2018 yritettiin toistaa 21 *Nature*- tai *Science*-lehdessä 2010–2015 julkaistua yhteiskuntatieteellistä tutkimustulosta (Serra-Garcia ja Gneezy 2021). Tuloksista 13 voitiin ja 8:aa ei voitu toistaa. Toistettujen/toistamattomien 13/8 tutkimuksen vuoden 2019 aikana saamien viittausten lukumäärän otoskertymäfunktio on piirretty vihreällä/sinisellä kuvassa 15.2. Otoskertymäfunktioiden porrasmaisuus näkyy kuvasta selkeästi. Kuvaajien yhteiset osuudet — molempien otoskertymäfunktioiden saadessa arvon 0 tai 1 — on merkitty vaaleanruskealla. Satunnaismuuttuja on viittausten lukumäärä (toistettuun tai toistamattomaan) tutkimukseen. (Luku 0.644 ja viereinen kaksipäinen nuoli selitetään jaksossa 15.2.2.)

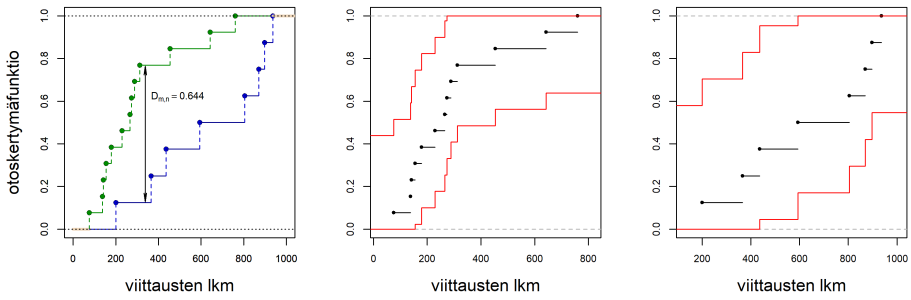
NSM3-paketin komennolla `ecdf.ks.CI(x)` kertymäfunktion 95 %:n samanaikaisen luottamusvyöhykkeen laskeminen ja piirtäminen käy kätevästi:

```
x <- c(200, 366, 436, 594, 804, 870, 898, 936)
ecdf.ks.CI(x)
ecdf.ks.CI(x)$upper
## [1] 0.57927 0.70427 0.82927 0.95427 1.00000 1.00000 1.00000 1.00000
ecdf.ks.CI(x)$lower
## [1] 0.00000 0.00000 0.00000 0.04573 0.17073 0.29573 0.42073 0.54573
```

Lukuarvot 200, 366 jne. ovat viittausten lukumäärät kullekin 8 tutkimukselle, joita ei saatu toistettua. Ne on luettu suuruusjärjestyksessä R:ään, mutta lukujärjestyksellä ei ole väliä. Oikeanpuoleisin osa kuvaa 15.2 on piirretty koodin kahdella ensimmäisellä rivillä (ja muutamalla kuvasta merkintöjä poistavalla

tarkennuksella). Komennot `ecdf.ks.CI(x)$upper` ja `ecdf.ks.CI(x)$lower` palauttavat punaisella piirretyt 95 %:n samanaikaisen luottamusvyöhykkeen ylä- ja alarajat. Keskimmäisin osa kuvaa 15.2 on tuotettu vastaavalla `ecdf.ks.CI`-komennolla. R on laskenut vyöhykkeet käyttäen pienillekin havaintomäärille pätevää kaavaa (15.3). Luottamusvyöhykkeet ovat hyvin leveitä, koska havaintoja on vähän.

Toistettujen tutkimusten otoskertymäfunktio sijoittuu vasemmalle toistamattomien tutkimusten otoskertymäfunktiosta. Toistettuihin tutkimuksiin on viitattu keskimäärin vähemmän. Serra-Garcia ja Gneezy (2021) argumentoivat (monen aineiston ja kattavampien analyysien perusteella), että artikkeleihin, joissa on julkaistu tuloksia, jotka on voitu toistaa, viitataan tieteessä vähemmän kuin tuloksiin, joita ei ole voitu toistaa. Mahdollinen selitys on, että viittauksia keräävät tulokset ovat odottamattomia ja huomiota herättäviä mutteivät välttämättä tosia. □



Kuva 15.2: Toistettujen ja toistamattomien tutkimusten viittausten lukumäärien otoskertymäfunktiot sekä kertymäfunktioiden 95 %:n samanaikaiset luottamusvyöhykkeet.

Likiarvoistuksen (15.2) $z^*_{1-\alpha}$-arvoja voidaan laskea lisää kaavalla $[-0.5 \times \log(\alpha/2)]^{1/2}$ (Lindgren 1976, 494).

Miller (1956) on taulukoinut tunnusluvun D_n kvantiilin $d_{1-\alpha}$ arvoja pienillä havaintomäärillä (myös Dickinson Gibbons ja Chakraborti 2021, taulukko F). Kaava (15.3) likimäärin toistaa taulukon 0.95. kvantiilit. Taulukko mahdollistaa kertymäfunktion samanaikaisen luottamusvälin määrittämisen tarkasti pienillä havaintomäärillä tavanomaisilla luottamustasoilla.

Tunnusluvulla D_n voidaan testata nollahypoteesia tietystä jakaumasta. Dickinson Gibbons ja Chakraborti (2021, jaksot 4.4–4.9) sekä Hollander ym. (2014, jakso 11.5) selittävät testit. Niiden etuja: Jos satunnaismuuttuja on jatkuva-arvoinen, kertymäfunktioon perustuvan testin pitäisi olla voimakkaampi kuin χ^2-testin (luku 12.2), koska jälkimmäisessä menetetään informaatiota luokittelulla. Luokittelu ja siten testin tulos eivät ole yksikäsitteisiä. Edellisen testin voi toteuttaa yksisuuntaisesti; jälkimmäistä ei. Tunnusluvun D_n jakauma tunnetaan pienilläkin havaintomäärillä.

Conover (2006, jakso 6.1), Dickinson Gibbons ja Chakraborti (mts. 126) sekä Hollander ym. (mts:t 573 ja 577) opastavat, kuinka päätellä tilastollisesti, jos satunnaismuuttuja on diskreetti tai luokiteltu ja aineistossa on tasahavaintoja. Conoverin (mts. 435) mukaan diskreettiyden huomioimisen vaikutus voi olla päättelyssä suuri. dgof-paketti sallii diskreetin nollajakauman.

15.2.2 Kahden jakauman testaus

Tarkastellaan kahta jatkuva-arvoista satunnaismuuttujaa, joiden kertymäfunktiot ovat $F(x)$ ja $G(x)$. Satunnaismuuttujista on käytössä riippumattomat m:n ja n:n kokoiset satunnaisotokset. *Kahden otoksen Kolmogorovin–Smirnovin tunnusluku*

$$D_{m,n} = \sup_{-\infty<x<\infty} |\hat{F}(x) - \hat{G}(x)|$$

on otoskertymäfunktioiden $\hat{F}(x)$ ja $\hat{G}(x)$ erotus (pystysuunnassa) suurimmillaan. Nollahypoteesi on, että kertymäfunktiot $F(x)$ ja $G(x)$ ovat samat. Tunnusluvun eli testisuureen suuret arvot viittaavat nollahypoteesin pätemättömyyteen.

Esimerkki. Tutkimusviittaukset II. Toistettujen ja toistamattomien tutkimusten otoskertymäfunktioiden arvojen erotus on suurimmillaan 0.644. Suurimman erotuksen kohta on merkitty kaksipäisellä nuolella kuvassa 15.2. Erotus on suuri. Suuruus tilastollisessa mielessä selviää harjoitustehtävässä. □

Nikolai Smirnov on todistanut, että otoskokojen m ja n suuretessa suhteen m/n pysyessä samana tunnusluvun $\sqrt{mn/(m+n)}\, D_{m,n}$ jakauma yhtyy tunnusluvun $\sqrt{n}\, D_n$ jakaumaan. Tällöin suurilla havaintomäärillä $D_{m,n}$:n kvantiili $d_{1-\alpha}$ on

likimain

$$d_{1-\alpha} \approx \frac{z^*_{1-\alpha}}{\sqrt{\frac{mn}{m+n}}} \tag{15.4}$$

tuloksen (15.2) tapaan. Sen yhteydessä määriteltiin $z^*_{1-\alpha}$. Mikäli otoskoot m ja n eroavat, niiden tulee olla hyvin suuria, jotta tulos (15.4) olisi käyttökelpoinen. Myös tunnusluvun pienotosjakauma tunnetaan (Dickinson Gibbons ja Chakraborti 2021, 258, taulukko I).

Kahden otoksen Kolmogorovin–Smirnovin testi on jakaumasta riippumaton. Voidaan osoittaa, että testisuure riippuu yksinomaan havaintojen sijaluvuista suuruusjärjestyksessä (*ranks*) yhdistetyssä otoksessa havainnoista (Lehmann 1975, 35–36). Mikäli aineistossa on tasahavaintoja (jakso 8.2), testin todellinen koko voi olla pienempi kuin tarkoitettu (Lehmann 1975, 39).

Testi on erityisen sopiva, kun mielenkiinto on jakaumien muissakin eroavaisuuksissa kuin odotusarvojen erisuuruudessa. Jälkimmäisessä tilanteessa odotusarvojen eroihin fokusoivat testit (esim. jakso 12.4) tapaavat olla voimakkaampia.

R-komento `ks.test` laskee testisuureen $D_{m,n}$ ja sen tarkan p-arvon pienten otosten teoriaan perustuen, jos $nm < 10\,000$ eikä aineistoissa ole tasahavaintoja. Muulloin komento laskee p-arvon suurten otosten teorian avulla edellä esitettyyn tapaan.

Esimerkki. Vihkimiset 2013 III. Kirkolliset vihkimiset ajoittuvat siviilivihkimisiä useammin kesään (kuva 8.1), joten vihkimisten varianssit eronnevat. Vihkimisten vaihteluväli on rajattu ($[1, 365]$). Odotusarvon muuttuessa jakauma ei voi siirtyä, joten sen muodon täytyy muuttua. Vihkimisten jakaumat voivat erota monella mielenkiintoisella tavalla, jotka eivät välttämättä ole tiivistettävissä poikkeamiin yhdessä tai kahdessa parametrissa. Kahden otoksen Kolmogorovin–Smirnovin testi sopii moninaisten erojen jakaumien välillä todentamiseen. Sovelletaan sitä.

Sijoitetaan havainnot kirkollisista vihkimisistä ja siviilivihkimisistä muuttujiin `x` ja `y`. R-komento `ks.test(x,y)` palauttaa $D_{m,n}$-testisuureen arvoksi 0.119, p-arvoksi luvun, joka poikkeaa 0:sta 16. desimaalissa ja varoittaa aineiston tasahavainnoista:

```
##    Two-sample Kolmogorov-Smirnov test
## data:  x and y
## D = 0.11868, p-value < 2.2e-16
## alternative hypothesis: two-sided
## Warning message:
## In ks.test(x, y) : cannot compute correct p-values with ties
```

Tasahavainnot tekevät testistä konservatiivisen eli heikentävät sitä. Tasahavaintojen heikentämänäkin testi hylkää nollahypoteesin jakaumien samuudesta kaikilla totutuilla merkitsevyystasoilla.

Jakaumat eroavat. Aineistoissa eroa on sekä keskimääräisessä vihkipäivässä (14.7. tai 10.7.) että vihkipäivien keskihajonnassa (77.5 tai 97.3):

```
mean(x); mean(y)
## [1] 194.9068
## [1] 191.0354
sd(x); sd(y)
## [1] 77.47133
## [1] 97.31096
```

Testi tukee kuvan 8.1 vaikutelmaa kirkollisten vihkimisten painottumisesta kesäpäiviin siviilivihkimisiä enemmän. □

Lehmann (1975, taulukko F) taulukoi lukuisia tunnusluvun $[mn/(m+n)]^{1/2}\, D_{m,n}$ jakauman kvantiileja suurilla otoskoilla. Dickinson Gibbons ja Chakraborti (2021, 262) ehdottavat tapaa tehdä testi tasahavaintotilanteessa. Hollander ym. (2014, 192) osoittavat lähteitä tasahavaintotilanteisiin.

NSM3-paketin `cKolSmirn-` ja `pKolSmirn`-komennot palauttavat pienillä otoskoilla normeeratun testisuureen $(nm/d)D_{m,n}$ kriittiset arvot ja p-arvon (Hollander ym. 2014, 191). Tässä d on m:n ja n:n suurin yhteinen tekijä. Suurilla otoskoilla komennot toimivat edellä esitetyn suurten otosten teorian pohjalta.

15.3 Yhteysmitat ja riippumattomuus

Satunnaismuuttujien yhteyksien tai vaihtoehtoisesti riippumattomuuden tutkiminen on tilastotieteen ydintä. Tutuin *yhteysmitta* (*measure of association*) on kahden satunnaismuuttujan X ja Y lineaarisen yhteyden vahvuutta kuvaava korrelaatiokerroin (jakso 6.5) ja sen otosvastine (jakso 8.2). Edellä on tutkittu satunnaisotoksia välimatka-asteikollisista satunnaismuuttujista. Tässä jaksossa aineisto $(x_1, y_1), \ldots, (x_n, y_n)$ voi olla satunnaisotos myös järjestysasteikollisista satunnaismuuttujista.

Jaksossa tutustutaan yhteysmittoihin ja riippumattomuustesteihin. Jälkimmäinen käsite voi herättää liioiteltuja toiveita. Satunnaismuuttujat eivät ole riippumattomia, jos yhteysmitan teoreettinen arvo poikkeaa 0:sta. Satunnaismuuttujat eivät silti välttämättä ole riippumattomia, jos arvo on 0. Tilastollisia epälineaarisia yhteyksiä on niin monenlaisia, että yhteysmitat eivät pysty

tunnistamaan kaikkia. Tilanne vertautuu aiemmin opittuun, kuinka satunnaismuuttujilla voi olla vahvoja epälineaarisia yhteyksiä, vaikka niiden välinen korrelaatio on 0 (jaksot 6.5 ja 8.2). Selkeää sääntöä, milloin riippumattomuutta kannattaisi testata milläkin yhteysmitalla, ei myöskään ole. Mitat heijastavat kaikki riippuvuutta omalla tavallaan.

Määritellään riippumattomuus kertymäfunktioiden avulla. Tapahtumat A ja B ovat riippumattomia, jos todennäköisyys, että molemmat tapahtuvat, on molempien todennäköisyyksien tulo: $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$ (yhtälö (4.10)). Valitaan tapahtumiksi $X \leq x$ ja $Y \leq y$, joissa x ja y ovat reaalilukuja. Satunnaismuuttujat X ja Y ovat riippumattomia, jos yhtäsuuruus

$$\mathsf{P}(X \leq x \cap Y \leq y) = \mathsf{P}(X \leq x)\mathsf{P}(Y \leq y) = F_X(x)F_Y(y) \tag{15.5}$$

pätee kaikille (x, y)-pareille. Yllä $F_X(x)$ ja $F_Y(y)$ ovat satunnaismuuttujien X ja Y kertymäfunktiot kehitettynä lukuarvoilla x ja y.

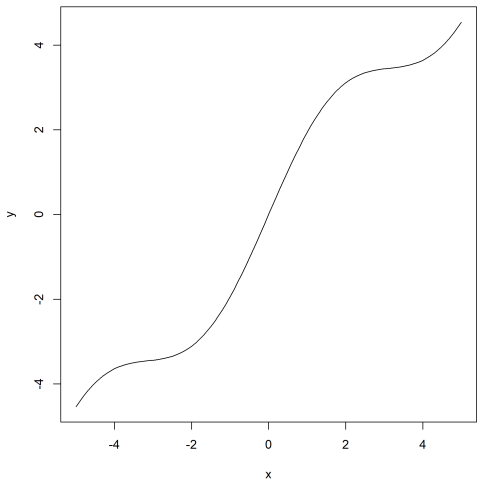
Jakson yhteysmitat saavat arvoja väliltä $[-1, 1]$. Arvot ovat sitä lähempänä 1:tä tai -1:htä, mitä paremmin satunnaismuuttujien välinen suhde on kuvattavissa epälineaarisella monotonisella funktiolla. Sellaisella y suurenee tai pienenee aina x:n suuretessa. Kuvassa 15.3 on esimerkki epälineaarisesta monotonisesti kasvavasta funktiosta. Otoskorrelaatiokerroin ei saa arvoa 1, jos aineiston havaintoparit ovat kuvan käyrällä (ylipäänsä ja jos $n > 2$).

Yhteysmitat jaksossa ovat vakaampia oudokkien suhteen kuin otoskorrelaatiokerroin. Tavaton oudokki voi muuttaa suuresti otoskorrelaatiokerrointa. Oudokin muuntaminen sijaluvuksi (jakso 15.3.1) tai eri- tai samanlaisuusmitaksi (jakso 15.3.2) voi mitöidä virhekirjauksen.

Osoitettavat yhteysmittojen jakaumat pätevät satunnaismuuttujien X ja Y jakaumasta riippumatta, kunhan riippumattomuusehto (15.5) toteutuu.

15.3.1 Spearmanin korrelaatiokerroin

Järjestetään $(x_1, y_1), \ldots, (x_n, y_n)$ -aineiston x_i- ja y_i-havainnot erikseen suuruusjärjestykseen ja merkitään $R_i = R(x_i)$:llä ja $S_i = S(y_i)$:llä sijalukuja $(1, \ldots, n)$ x_i:n ja y_i:n järjestetyissä aineistoissa. (Jos x_3 on toiseksi suurin x_i-arvoista, niin $R_3 = n - 1$.) Lasketaan otoskorrelaatio (8.2) aineistosta $(R_1, S_1), \ldots,$



Kuva 15.3: Monotoninen funktio.

(R_n, S_n):

$$\hat{\rho}_s = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2}}. \tag{15.6}$$

Yllä $\bar{R}$ ja $\bar{S}$ ovat R_i:den ja S_i:den keskiarvot. Tunnusluku $\hat{\rho}_s$ on *Spearmanin järjestyskorrelaatiokerroin* (*rank correlation coefficient*).

Spearmanin korrelaatiokertoimen intuitio on selvä: Jos parissa (R_i, S_i) järjestysluvut tapaavat olla yhtä aikaa suuria, $\hat{\rho}_s > 0$; jos järjestysluvuista toinen tapaa olla suuri toisen ollessa pieni, $\hat{\rho}_s < 0$. Spearmanin korrelaatiokerroin rajoittuu välille $(-1, 1)$, koska se lasketaan otoskorrelaatiokertoimen kaavalla.

Voidaan osoittaa, että $\mathsf{E}(\hat{\rho}_s) = 0$, $\mathsf{V}(\hat{\rho}_s) = (n-1)^{-1}$ ja että $\hat{\rho}_s$ noudattaa normaalijakaumaa suurilla havaintomäärillä, jos riippumattomuusehto (15.5) pätee. Nollahypoteesia riippumattomuudesta voidaan siten suurilla havaintomäärillä testata vertaamalla testisuuretta

$$\frac{\hat{\rho}_s}{1/\sqrt{n-1}} = \sqrt{n-1}\hat{\rho}_s$$

standardinormaalijakaumaan. Kaavojen (12.10) ja (13.13) kaltainen jakaumalikiarvoistus

$$\sqrt{n-2}\frac{\hat{\rho}_s}{\sqrt{1-\hat{\rho}_s^2}} \sim \mathtt{t}(n-2) \tag{15.7}$$

on mahdollisesti parempi (Dickinson Gibbons ja Chakraborti 2021, 447). Myös testisuureen eksakti jakauma (pienille n) on laskettavissa. Jos nollahypoteesia riippumattomuudesta ei hylätä, ei tule päätellä, että satunnaismuuttujat olisivat riippumattomia. Spearmanin korrelaatiokerroin ei tunnista kaikkia mahdollisia satunnaismuuttujien välisiä epälineaarisia riippuvuuksia. Nollahypoteesi hylättäessä päätellään, että satunnaismuuttujat eivät ole riippumattomia.

Jos aineistossa on tasahavaintoja (jakso 8.2), ne korvataan *keskisijaluvuilla* (*mid-rank*) eli tasahavaintojen sijalukujen keskiarvolla. Riippumattomuustesti tehdään muuten kuten edellä.

Esimerkki. Tasahavainnot ja keskisijaluvut. Olkoon x-aineisto $7.2, 4.1, 10.3, 8.4$ ($n = 4$). R-sijalukuaineisto on $2, 1, 4, 3$. Olkoon x-aineisto $7.2, 4.1, 10.3, 7.2$. R-sijalukuaineisto on tällöin $2.5, 1, 4, 2.5$, koska tasahavaintojen sijalukujen keskiarvo on $(2 + 3)/2 = 2.5$. □

Jos tasahavaintoja ei ole, Spearmanin järjestyskorrelaatiokerroin voidaan laskea kaavalla

$$1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

(harjoitustehtävä). Yllä $d_i = R_i - S_i$. Tämä kaava tuottaa eri luvun kuin kaava (15.6), jos tasahavaintoja on. Kaava tässä on laskennallisesti helpompi muttei nykypäivänä tarpeen. Se esitetään, koska siihen törmää kirjallisuudessa.

Jos aineisto on välimatka-asteikollinen, muutoksessa järjestysasteikolliseksi häviää informaatiota. Peukalosääntö on siten, että Spearmanin korrelaatiokerroin kannattaa laskea, jos halutaan tutkia lineaarista riippuvuutta yleisempää riippuvuutta tai jos aineisto on alun perin järjestysasteikollinen.

R-komento `cor(x,y,method="spearman")` palauttaa Spearmanin korrelaatiokertoimen ja `cor.test(x,y,method="spearman",exact=FALSE)` testisuureen (15.7) p-arvon $\mathtt{t}(n-2)$-jakaumasta, kun nollahypoteesi on riippumattomuus. Asettamalla jälkimmäisessä komennossa `exact=TRUE` saadaan testisuureen eksakti eli pienotosjakaumasta laskettu p-arvo. Dickinson Gibbons ja Chakraborti (2021) taulukoivat $\hat{\rho}_s$:n pienotosjakaumaa otoskokoon 30 asti. Jakaumasta $\mathtt{t}(n-2)$ laskettu ja eksakti p-arvo eivät välttämättä eroa paljoa pienemmilläkään otoskoilla. Spearmanin korrelaatiokertoimen luottamusväli selviää DescTools-paketin `Assocs`-komennolla.

15.3.2 Kendallin τ, Goodmanin ja Kruskalin γ ja Stuartin τ_c

Oletetaan aluksi, että x_i-havainnot ovat erisuuria ja y_i-havainnot samoin. Havaintoparit (x_i, y_i) ja (x_j, y_j) ovat *samanlaiset* (*concordant*), jos $(x_i - x_j)(y_i - y_j) > 0$ ja *erilaiset* (*disconcordant*), jos $(x_i - x_j)(y_i - y_j) < 0$ $(i \neq j)$. Parit ovat siis samanlaiset, jos erotusten $(x_i - x_j)$ ja $(y_i - y_j)$ etumerkki on sama ja erilaiset, jos erotusten etumerkki on eri. Samanlaisilla pareilla *i.* havaintoparin molemmat komponentit ovat suurempia tai pienempiä kuin *j.* havaintoparin vastaavat komponentit.

Kendallin otosjärjestyskorrelaatiokerroin τ on

$$\tau = \frac{K - D}{n(n-1)/2}. \tag{15.8}$$

Tässä K ja D ovat samanlaisten ja erilaisten parien lukumäärä aineistossa $(x_1, y_1), \ldots, (x_n, y_n)$ ja $K + D = n(n-1)/2$. Mitä suurempi erotuksen $K - D$ itseisarvo on, sitä vahvempi on yhteys muuttujien x ja y välillä.

Perustelu: Erilaisista n:stä havainnosta voidaan poimia binomikertoimen

$$\binom{n}{2} = \frac{n!}{(n-2)!2!} = \frac{n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1}{(n-2) \times (n-1) \times \cdots \times 2 \times 1 \times (2 \times 1)} = \frac{n(n-1)}{2}$$

osoittama määrä erilaisia kahden alkion osajoukkoja eli pareja (kaava (5.3)). Binomikertoimen tulkinta tässä yhteydessä on, että havaintopareista voidaan tehdä $n(n-1)/2$ vertailua, että K:n ja D:n maksimiarvo on $n(n-1)/2$ (toisen saadessa arvon 0) ja että kaavassa (15.8) nimittäjä skaalaa tunnusluvun niin, että τ rajoittuu välille $(-1, 1)$. Jos samanlaisia pareja on enemmän kuin erilaisia, $K - D > 0$ ja $\tau > 0$. Päinvastaisessa tilanteessa $K - D < 0$ ja $\tau < 0$. Toisin sanoen $\tau > 0$, jos aineistossa x_i ja y_i ovat usein yhtä aikaa "suuria" tai "pieniä" suhteessa muihin havaintopareihin, ja $\tau = 1$, jos niin on aina. Vastaavasti $\tau < 0$, jos aineistossa x_i:n ollessa suuri y_i on usein pieni (ja päinvastoin) suhteessa muihin havaintopareihin, ja $\tau = -1$, jos niin on aina. Spearmanin korrelaatiokertoimen tapaan $\tau = 1$ tai $\tau = -1$, jos x:t ja y:t sijoittuvat monotonisesti kasvavalle tai vähenevälle funktiolle.

Testi riippumattomuudelle voidaan perustaa (tässä todistamattomille) tuloksille $\mathsf{E}(K - D) = 0$, $\mathsf{V}(K - D) = n(n-1)(2n+5)/18$ ja erotuksen $K - D$ normaalisuus suurilla havaintomäärillä. Riippumattomuusehdon (15.5) vallitessa

$$\frac{K - D}{\sqrt{n(n-1)(2n+5)/18}}$$

noudattaa siis standardinormaalijakaumaa satunnaismuuttujien X ja Y jakaumista riippumatta. Testisuure kirjoitetaan monesti yhtäpitävässä muodossa

$$\frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}}$$

(harjoitustehtävä). Testi voidaan tehdä myös tukeutuen erotuksen $K - D$ tunnettuun pienotosjakaumaan. Kuten Spearmanin järjestyskorrelaatiota käytettäessä, riippumattomuusnollahypoteesin hylkäämättömyydestä ei voi päätellä riippumattomuutta mutta hylkäämisestä voi päätellä riippuvuuden.

Aineistossa on tasahavaintoja, jos $x_i - x_j = 0$ tai $y_i - y_j = 0$, jolloin $(x_i - x_j)(y_i - y_j) = 0$, vaikka $i \neq j$. Tällöin erotuksen $K - D$ varianssi on eri kuin kaavassa edellä. Erotus $K - D$ jaetaan tasahavainnot huomioivalla keskihajonnan kaavalla (esim. Hollander ym. 2014, 397), ja riippumattomuustesti voidaan jälleen perustaa standardinormaalijakaumalle.

Tasahavaintotilanteessa K tai D ovat väistämättä pienempiä kuin $n(n-1)/2$. Tällöin $|\tau| < 1$, vaikka kaikki havaintoparit osuisivat vaikkapa ja jopa kuvan 15.3 käyrälle eli riippuvuus olisi täydellistä (tasahavainnot olisivat parien $(x_i, y_i) = (x_j, y_j)$ kaltaisia).

Goodmanin ja Kruskalin γ

$$\gamma = \frac{K - D}{K + D}$$

voi saada arvon -1 tai 1, vaikka aineistossa olisi tasahavaintoja. Nimittäjässä on samanlaisten ja erilaisten parien lukumäärä, eli tasahavaintoja ei huomioida. Jos tasahavaintoja ei ole $K + D = n(n-1)/2$ ja $\gamma = \tau$.

Stuartin τ_c on

$$\tau_c = \frac{2\min(I, J)(K - D)}{n^2[\min(I, J) - 1]} = \frac{K - D}{n^2[\min(I, J) - 1]/[2\min(I, J)]}.$$

Se huomioi sekä tasahavainnot että tilanteen, jossa järjestysasteikollinen aineisto on $I \times J$ -taulukko, jossa rivejä ja sarakkeita on eri määrä ($I \neq J$). Yllä $\min(I, J)$ on pienempi luvuista I ja J ($\min(I, K) = I = J$, jos $I = J$). Stuartin τ_c yhtyy jakson muihin yhteysmittoihin, jos tasahavaintoja ei ole ja taulukko on $n \times n$ -muotoa ($n = I = J$):

$$\tau_c = \frac{K - D}{n^2[\min(I, J) - 1]/[2\min(I, J)]} = \frac{K - D}{n^2(n-1)/2n} = \frac{K - D}{n(n-1)/2}$$

ja

$$\tau = \gamma = \tau_c.$$

Erotus $K - D$ on helposti laskettavissa visualisoimalla aineisto, jos se ei ole suuri. Esimerkki selittää parhaiten.

Esimerkki. Havainnollistetaan jääkiekon SM-liigan joukkueittaisia järjestyspareja vuosina 2008 ja 2009

joukkue	sija 2009	sija 2008
JYP	1	5
Blues	2	2
HPK	3	12
Jokerit	4	3
Kärpät	5	1
Kalpa	6	13
HIFK	7	7
Ilves	8	8
Pelicans	9	6
TPS	10	10
Lukko	11	9
Ässät	12	14
Tappara	13	4
Saipa	14	11

$*$-kuviolla (ylempi $*$-kuvio).[184] Samanlaisten (K) ja erilaisten (D) parien lukumäärä voidaan laskea $*$-kuviosta kätevästi käymällä se läpi ylhäältä alas. Kunkin asteriskin ($*$) kohdalla lasketaan siitä alaoikealle ja-vasemmalle sijoittuvien asteriskien lukumäärät. Summaamalla niiden erotukset saadaan $K - D$:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1														
2			*											
3				*										
4					*									
5	*													
6							*							
7								*						
8									*					
9						*								
10										*				
11									*					
12				*										
13											*			
14												*		

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		*												
2			*											
3				*										
4					*									
5						*								
6							*							
7								*						
8									*					
9										*				
10											*			
11												*		
12													*	
13														*
14	*													

$(1,5)$:	$9-4$	$=$	5
$(2,2)$:	$11-1$	$=$	10
$(3,12)$:	$2-9$	$=$	-7
$(4,3)$:	$9-1$	$=$	5
$(5,1)$:	$9-0$	$=$	9
$(6,13)$:	$1-7$	$=$	-6
$(7,7)$:	$5-2$	$=$	3
$(8,8)$:	$4-2$	$=$	2
$(9,6)$:	$4-1$	$=$	3
$(10,10)$:	$2-2$	$=$	0
$(11,9)$:	$2-1$	$=$	1
$(12,14)$:	$0-2$	$=$	-2
$(13,4)$:	$1-0$	$=$	1
Σ:	$K-D$	$=$	24

Aineistossa ei ole tasahavaintoja ja taulukko on $n \times n$ -muotoa, joten tunnusluvut ovat yhtäsuuria:

$$\tau = \gamma = \tau_c = \frac{K-D}{n(n-1)/2} = \frac{24}{14 \times 13/2} = 0.2637363.$$

Sekä $*$-kuvion että Kendallin τ:n suuruuden 0.264 mukaan SM-liigan vuosien 2009 ja 2008 paremmuusjärjestyksien yhteys on heikohko. Täydellisen riippuvuuden tilanteessa τ olisi 1 ja $*$-kuvio näyttäisi alemmalta $*$-kuviolta. Aineistoa tutkitaan lisää harjoitustehtävässä. □

Laskuesimerkki ja $*$-kuviot ovat opettavaisia. Käytännössä Kendallin τ, Goodmanin ja Kruskalin γ ja Stuartin τ_c lasketaan tilasto-ohjelmistolla.

Kendallin τ:n ja sen p-arvon voi laskea R-komennoilla `cor(x,y,method="kendall")` ja `cor.test(x,y,method="kendall",exact=FALSE)` (nollahypoteesin riippumattomuudesta pätiessä). Asettamalla jälkimmäisessä komennossa `exact=TRUE` seuraa testisuureen eksakti eli pienotosjakaumasta laskettu p-arvo. Dickinson Gibbons ja Chakraborti (2021) taulukoivat τ:n pienotosjakaumaa otoskokoon 30 asti. Likimääräinen ja eksakti p-arvo eivät välttämättä eroa suuresti, vaikka havaintoja olisi vähemmän. Kendallin τ:n luottamusväli selviää R:n DescTools-paketin `Assocs`-komennolla.

Goodmanin ja Kruskalin γ:n, sen p-arvon (riippumattomuusnollahypoteesi) sekä vastaavan luottamusvälin voi laskea MESS-paketin `gkgamma`-komennolla. Luottamusvälin voi laskea myös vcdExtra- ja DescTools-pakettien komennoilla `GKgamma` ja `Assocs`. Jälkimmäinen palauttaa samalla Stuartin τ_c:n ja siihen liittyvän luottamusvälin.

Viitteet

[1]N. Cusanus (1440, 2020): *De Docta Ignorantia.* Oppineesta tietämättömyydestä. Johdanto, suomennos ja selitykset Juha Pihkala. Suomalaisen teologisen kirjallisuusseuran julkaisuja 287. Suomalainen teologinen kirjallisuusseura. S. 51.

[2]*An Enquiry Concerning Human Understanding.* `http://www.earlymoderntexts.com/assets/pdfs/hume1748.pdf` (haettu 26.12.2020). (Eino Kailan suomennos 1938: *Tutkimus inhimillisestä ymmärryksestä.* WSOY.)

[3]`https://www.sttinfo.fi/tiedote/tutkimus-suomalaismuseoiden-koetut-hyvinvointivaikutukset-ovat-satojen-miljoonien-arvoiset?publisherId=65178636&releaseId=69963585` (haettu 31.1.2023).

[4]Yle Uutiset 22.9.2017 (`https://yle.fi/uutiset/3-9845228`; haettu 22.9.2017).

[5]Data voidaan määritellä laajemmin tai suppeammin. Luentomateriaalissa data samaistetaan numeroista koostuvaan aineistoon.

[6]A.–M. Staicu (2017): Interview with Nancy Reid. *International Statistical Review*, 85, 381–403. Kiitän Nancy Reidiä kuvasta 1.1 ja luvasta julkaista se.

[7]`https://www.goodreads.com/quotes/655987-if-we-have-data-let-s-look-at-data-if-all` (haettu 26.2.2023).

[8]K. Kafadar (2020): Reinforcing the Impact of Statistics on Society. *Journal of the American Statistical Association.* DOI: 10.1080/01621459.2020.1761217.

[9]`https://fi.wikipedia.org/wiki/Luettelo_musiikin_tyylilajeista` (haettu 27.6.2021).

[10]W.F. Bynum ja R. Porter (2006; toim.): Aristotle 384–22 BC. Greek Natural Philosopher. *Oxford Dictionary of Scientific Quotations.* OUP. Suomentaja tuntematon.

[11]Myös Mac- ja Linux-käyttöjärjestelmille löytyvät linkit täältä.

[12]Laplace (1840/1902, 1).

[13]J. Laine (toim.; 1989): *Suuri sitaattisanakirja.* Otava. S. 18.

[14]Helsingin Sanomat 27.5.2020 (`https://www.hs.fi/kulttuuri/art-2000006519386.html` (haettu 27.5.2020).

[15]Pekka Saurin twiitti 24.5.2020.

[16]Yrjö Kukkapuron 90-vuotishaastattelu Helsingin Sanomissa 5.4.2023.

[17]R.W. Hamming (1998): Mathematics on a Distant Planet. *The American Mathematical Monthly*, 105, 640–650.

[18]`https://yle.fi/uutiset/3-6096707` (haettu 1.12.2018).

[19]Yle Uutiset 24. ja 27.1.20220 (`https://yle.fi/a/3-11174149` ja `https://yle.fi/a/3-11181717`; haut 18.1.2023).

[20]`https://www.hs.fi/talous/art-2000006260346.html` (haettu 3.10.2019).

[21]Mikäli otosavaruudessa on ääretön määrä tulosvaihtoehtoja, tarvitaan neljäs oletus: Olkoot tapahtumat $A_1, A_2, \ldots$ erillisiä ($A_i \cap A_j = \emptyset$ kaikille $i \neq j$) ja kuulukoot ne otosavaruuteen S. Tällöin yhdistetyn tapahtuman $\cup_{i=1}^{\infty} A_i$ todennäköisyys on $\mathsf{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$. Katso esimerkiksi tässä lähteenä käytetty Larsenin ja Marxin (2001, 31–33) kirja. Samantapaisia todistuksia löytyy Blitzsteinin ja Hwangin (2015) sekä Venkateshin (2013) kirjoista.

[22]Blitzstein ja Hwang (2015, 26).

[23]`http://yle.fi/uutiset/3-9483513` (haettu 1.3.2017).

[24]Larsen ja Marx (2001, 53).

[25]C symboloi rivillä tapahtumaa ja yläindeksissä vastatapahtumaa. Tässä C on tapahtuma.

[26]Blitzstein ja Hwang (2015, 57).

[27]Larsen ja Marx (2001, 70).

[28]Todellisuudessa poikavauvan todennäköisyys on noin 0.52 (esim. Pawitan 2013, 75).

[29]Venkatesh (2013, 37).

[30]Asiaa pohditaan tarkemmin kirjoissa Blitzstein ja Hwang (2015, 57), Larsen ja Marx (2001, jakso 2.7) sekä Venkatesh (2013, luku 3).

[31]`https://yle.fi/uutiset/3-5356349` (haettu 10.1.2021).

[32]Riippumattomuusoletus kuvannee todellisuutta hyvin tässä yhteydessä. Maanjäristyksiä pidetään vaikeina ellei mahdottomina ennustaa. (Esim. `https://en.wikipedia.org/wiki/Earthquake_prediction` (haettu 10.1.2021).

[33]Tämä ja seuraavaa esimerkki ovat Blitzsteinin ja Hwangin (2015, 46–47) kirjasta.

[34]Tämä ja kaksi seuraavaa esimerkkiä ovat kirjasta Blitzstein ja Hwang (2015, 58).

[35]Larsen ja Marx (2001, 62–63), `https://www.hs.fi/kotimaa/art-2000002884170.html`) ja `https://www.hs.fi/kaupunki/art-2000009941035.html` (haut 23.10.2023).

[36]Huom! Bayes äännetään samaan tapaan kuin *days*.

[37]Lähteet: `http://www.ktl.fi/ttr/gen/rpt/hivsuo.html`, `http://www.ktl.fi/ttr/gen/rpt/hivaidskuo.html`, `http://www.tilastokeskus.fi/til/vaerak/2008/vaerak_2008_2009-03-27_tie_001_fi.html` ja `https://en.wikipedia.org/wiki/Diagnosis_of_HIV/AIDS` (haut 2010 ja 18.11.2016). Luvut ovat suuntaa antavia, muun muassa koska kaikki HIV-infektiot eivät ole tiedossa. Arviolta joka kolmas tartunta on diagnosoimaton (Suomen hiv-strategia 2013–2016. Terveyden ja hyvinvoinnin laitos 7/2012. S. 7).

[38]Kansanterveyslaitoksen HIV-yksikön johtajan Mika Salmisen artikkeli (2006) HIV-riskit kasvaneet (urlhttp://demo.seco.tkk.fi/tervesuomi/item/ktl:11672 (haettu 14.7.2014)) sekä Terveyden ja hyvinvoinnin laitoksen erikoistutkija Henrikki Brummer-Korvenkontio (henkilökohtainen tiedonanto 16.2.2010). Tutkimuksen tuloksia ei voida yleistää koskemaan suomalaisia homo- tai biseksuaalisia miehiä ylipäänsä mm., koska lehden lukijakunta voi olla valikoitunutta. Esimerkiksi lehden lukijakunta painottuu yli 30-vuotiaisiin miehiin. Myös voi olla, että vastaajiksi valikoitui erityisen riskialttiisti käyttäytyviä homoseksuaaleja, joille kysely tarjosi kätevän tilaisuuden HIV-testaukseen. Kyselyyn vastasi 410 (vastausprosentti 29) ja sylkinäytteen antoi 368 miestä. (Em. artikkeli.)

[39]Veto (esim. Rita 2004) ja vedonlyöntisuhde ovat vanhempia suomennoksia *oddsille*. Jälkimmäinen voi sekoittua kahden vastasuhteen suhteeseen ristisuhteeseen (*odds ratio*) (Rita ym. 2008). Vetokerroin-suomennosta käytettiin jaksossa 4.2.3, koska siellä pohdittiin vedonlyöntitilannetta.

[40]Lastensuojelu 2022. Tilastoraportti 24/2023. Terveyden ja hyvinvoinnin laitos. Kuvio 16 ja taulukko 8 `https://www.julkari.fi/handle/10024/146573` (haettu 16.5.2023).

[41]Tässä ja seuraavissa esimerkeissä laskut on tehty suuremmalla tarkkuudella kuin tekstistä ilmenee. Raportointitarkkuudella tehtyjen laskujen lopputulos poikkeaisi paikoin esitetystä.

[42]Mellin (1996, 55–57). Jakson puukaaviot on piirretty Tikz-koodilla (`https://sourceforge.net/projects/pgf/`). Koodi on oleellisesti sivulta `http://www.texample.net/tikz/examples/coin-flipping/`. (Viittaukset 26.1.2017.)

[43]Ioannidis (2005, 2019). De Long ja Lang (1992). Esimerkki perustuu lähteisiin J.A.C. Sterne ja G.D. Smith (2001): Sifting the Evidence — What's Wrong with Significance Tests? *British Medical Journal*, 322, 226–231. D.R. Cox (2001): Another Comment on the Role of Statistical Method. *British Medical Journal*, 322, 231. Agresti ja Finlay (2009, 165–166).

[44]`https://www.hs.fi/kotimaa/art-2000002879992.html`, `https://www.kaleva.fi/eduskuntaan-tehty-kyberhyokkays-on-vakava-isku-suo/3223418`, `https://www.hs.fi/politiikka/art-2000008137465.html` ja `https://yle.fi/uutiset/3-12292218` (haut 20.7.2021 ja 28.1.2022).

[45]Jakso perustuu Blitzsteinin ja Hwangin (2015) kirjaan ja artikkeleihin Baker ja Kramer (2001), Norton ja Divine (2015) sekä Wainer ja Brown (2004).

[46]Blitzstein ja Hwang (2015, 67–69).

[47]P.J. Bickel, E.A. Hammel ja J.W. O'Connell (1975): Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187, 398–404. W.G. Cochran (1968): The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, 295–313. S. Holm (2013): Tutkielmien arvosanat Helsingin yliopistossa. Pro gradu -tutkielma (tilastotiede). Matemaattis-luonnontieteellinen tiedekunta. Helsingin yliopisto. `https://helda.helsinki.fi/handle/10138/39493` (haettu 28.10.2016).

[48]H. Poincaré (1905): *Science and Hypothesis*. Hayes Barton Press. Luku 11.

[49]Edgar Allan Poe: Miten kirjoittaa Blackwoodin artikkeli. Kirjassa *Edgar Allan Poe. Kootut kertomukset*. Teos. Suomentanut Jaana Kapari-Jatta. Lainaukseen on lisätty kursivoinnit.

[50]Jako riittää tähän esitykseen. Satunnaismuuttuja, jonka jakauma on singulaarinen, ei ole diskreetti eikä jatkuva.

[51]Kattavampi määritelmä on mahdollinen.

[52]Epäjatkuvuuskohtia voidaan sallia rajallinen määrä.

[53]Kreikkalainen kirjaimisto kuvataan Wikipediassa: `https://fi.wikipedia.org/wiki/Kreikkalainen_kirjaimisto` (haettu 14.9.2022).

[54]Kvartiileilla ja desiileillä viitataan joskus myös niiden määrittämiin osajoukkoihin.

[55]Normaali-, χ^2-, t- ja F-jakaumia havainnollistavat kuvat 6.1, 7.2.2, 6.3 sekä 7.5–7.8 on piirretty muokkaamalla Crawleyn (2013, 206, 273 ja 290) R-koodeja.

[56]H. Rosling (2010): The Joy of Stats. `https://www.open.edu/openlearn/science-maths-technology/mathematics-and-statistics/mathematics/the-joy-stats-above-average` (haettu 13.6.2019).

[57]Mediaanille löytyy kirjallisuudesta erilaisia määritelmiä. Mediaani voidaan määritellä elegantisti optimointitehtävän ratkaisuna.

[58]Kaasun hinta nousi tavattomasti Venäjän hyökättyä Ukrainaan 2022. Saksa ei sallinut vastaavaa hinnan korotusta saksalaisille kuluttajille. Uniper oli kaatua, ja Fortumille tuli suomalaisen yrityshistorian suurimmat miljardiluokan tappiot. Tässäkin esimerkissä varianssiin liittyi lopulta riski.

[59]`https://blogi.foreca.fi/2019/06/lumisade-vs-megahelle-kumpi-voittaa/` (haettu 8.6.2019).

[60]Y. Ding ja K. Savani (2020): From Variability to Vulnerability: People Exposed to Greater Variability Judge Wrongdoers More Harshly. *Journal of Personality and Social Psychology*, 118, 1101–1117.

[61]Tero Valkosen suomennos. Isaacson (2019, 197).

[62]Kuva pohjautuu Alan Arnholtin R-koodiin (`https://github.com/alanarnholt/PASWR2E-Rscripts/blob/master/ChapterScripts/chapter04.R`; haettu 26.2.2016). Koodi on luvusta 4 kirjasta Ugarte ym. (2016).

[63]Lapset määrättiin asumaan isän luona 27.3 %:ssa, äidin luona 65.2 %:ssa ja 7.5 %:ssa päätöksistä molemmilla (lapset "jaettiin" tai määrättiin "vuoroasumisesta"). Havaintoja oli 127. Periaatteessa on mahdollista, että samoista lapsesta olisi riidelty samanvuoden aikana useamman kerran aineistossa. Se mahdollisuus sivuutetaan, ja esimerkissä oletetaan, että havainnot ovat riippumattomia. E. Valkama ja M. Litmala (2006): Lasten huoltoriidat käräjäoikeuksissa. OPTL:n julkaisuja 224. `https://helda.helsinki.fi/handle/10138/152456` (haettu 25.2.2016).

[64]Käytännössä kerran haastateltua ei haastateltaisi uudestaan. Tällöinkin multinomijakaumalla voitaisiin approksimoida vastausten jakaumaa, sillä haastateltujen kansalaisten lukumäärä olisi pieni suhteessa äänestysikäisten suomalaisten lukumäärään.

[65]Lindgren (1976, 168).

[66]Tijms (2012, 136).

[67]Esimerkki juontaa Mika Sutelan laskelmiin liikennekuolemista ja siihen perustuvaan Helsingin Sanomien uutiseen 7.1.2020 "Selvitys: Tampere on selvästi vaarallisin pyöräilijöille ja jalankulkijoille – – " (`https://www.hs.fi/kotimaa/art-2000006365292.html`; haettu 25.1.2020).

[68]Lauseessa oletettiin, että satunnaismuuttujilla X_i on odotusarvo ja varianssi. Niin ei aina ole. Sellaisissa tilanteissa lause ei takaa keskiarvon normaalisuutta.

[69]Galton (1899, 66).

[70]Ramachandran ja Tsokos (2020, 96) sekä Engineering Statistics Handbook (`http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc331.htm` (haettu 6.3.2015).

[71]Armitage, Berry ja Matthews (2002, 76), Larsen ja Marx (2001, 248), Lindgren (1976, 185) ja Ross (2017, 253). Lindgrenin ja Rossin esimerkeissä approksimaatio toimii hyvin jo, kun π on 0.1–0.125 ja n on 8–10.

[72]J.S. Rosenthal (2014): Statistics and the Ontario Lottery Retailer Scandal. *Chance*, 27, 4–9.

[73]Iltalehti 23.12.2015, Helsingin Sanomat 3.4.2016 ja Ilta-Sanomat 20.1.2020 (`http://www.iltale`

hti.fi/digi/2015122320869989_du.shtml, https://www.hs.fi/kotimaa/art-2000002894315.html ja https://www.is.fi/kotimaa/art-2000006378479.html. Haut 24.1.2021).

[74]Kiitän Stephen M. Stigleriä luvasta julkaista ottamansa kuva 7.10. Kuva 7.13: Matemateca (IME/USP)/Rodrigo Tetsuo Argenton (https://commons.wikimedia.org/wiki/File:Galton_box.jpg; haettu 5.7.2017). Galtonin koneen R-kielisen emulaattorin ovat tehneet Y. Xie, L. Yu ja K. O'Rourke (2013): Demonstration of the Quincunx (Bean Machine/Galton Box). http://www.rforge.net/doc/packages/animation/quincunx.html (haettu 10.2.2015). Koneen "parannettu" versio on patentoitu Yhdysvalloissa 1990 (http://www.google.nl/patents/US4900255; haettu 4.3.2016).

[75]Kuvan piirtämisessä on käytetty mnormt-pakettia ja koodia sivulta https://stackoverflow.com/questions/62426391/how-to-plot-the-surface-and-contours-of-a-bivariate-normal-distribution-using-fe (haettu 14.12.2021).

[76]Vertauksen esittäjäksi todetaan monesti George Gallup nimeämättä lähdettä. Bestin ja Radcliffin (2005, 711–712) mukaan vertaus on painettu Amerikan historiallisen yhdistyksen GI Roundtable -lehtisessä "Are Opinion Polls Useful?", jonka kirjoittajia ei tiedetä. Sivun https://www.historians.org/about-aha-and-membership/aha-history-and-archives/gi-roundtable-series/pamphlets mukaan ko. lehtisen ensimmäisen käsikirjoituksen on laatinut R. Nafziger ja lehtinen on julkaistu 1946. Vertaus löytyy hieman pidemmässä muodossa sivulta https://www.historians.org/about-aha-and-membership/aha-history-and-archives/gi-roundtable-series/pamphlets/em-4-are-opinion-polls-useful-(1946)/how-are-polls-made. (Haut 19.2.2019.)

[77]Johannsen (1920, 315).

[78]I. Anttila (1966): Piilorikollisuus ja ilmirikollisuus. Teoksessa I. Anttila ja R. Jaakkola (toim.): Piiloon jäävä rikollisuus. Kansalaiskasvatuksen keskus. Monistesarja 12/66.

[79]Histogrammin määritelmä vaihtelee. Histogrammit voidaan jakaa frekvenssihistogrammeihin (*frequency histogram*) ja suhteellisiin frekvenssihistogrammeihin (*relative frequency histogram*). Kuvissa FVihkimiset2013 ja 8.3 on frekvenssihistogrammi.

[80]Kiitän Kirkkohallituksen tilastoasiantuntija Pasi Mäkelää aineiston luovuttamisesta 27.3.2014.

[81]Kiitän Peter Nybergiä ja Mika Vaihekoskea indeksin luovuttamisesta opetuskäyttöön. Indeksin laskutapa selitetään artikkelissa P. Nyberg ja M. Vaihekoski (2010): A New Value-Weighted Total Return Index for the Finnish Stock Market. *Research in International Business and Finance*, 24, 267–283.

[82]Kiitän Osmo Kontulaa aineiston luovuttamisesta 6.5.2019. Aineistoa kuvataan tarkemmin artikkelissa O. Kontula (2009): Between Sexual Desire and Reality: The Evolution of Sex in Finland. The Family Federation of Finland: The Population Research Institute D49/2009. Helsinki.

[83]Kuuluisa esimerkki on Gregor Mendelin 1865 julkaisema artikkeli perinnöllisyyskokeista. Ronald Fisher (1936) argumentoi, että aineisto ei voi olla aito. Ks. myös Ross (2017, 585–588, 593). Jaksossa 14.2.2 on esimerkki ymmärtämättömyydestä johtuvasta osa-aineiston analyysista, joka johti seitsemän ihmisen kuolemaan.

[84]Esim. V. Angelini, M. Bertoni ja L. Corazzini (2017): Unpacking the Determinants of Life Satisfaction: A Survey Experiment. *Journal of the Royal Statistical Society, A*, 180, 225–246. A. Deaton ja A.A. Stone (2016): Understanding Context Effects for a Measure of Life Evaluation: How Responses Matter. *Oxford Economic Papers*, 68, 861–870. J.P. Schuldt, S.H. Konrath ja N. Schwarz (2011): "Global Warming " or "Climate Change"? Whether the Planet is Warming Depends on Question Wording. *Public Opinion Quarterly*, 75, 115–124.

[85]Esim. S. Laaksonen ja M. Heiskanen (2014): Comparison of Three Modes for a Crime Victimization Survey. *Journal of Survey Statistics and Methodology*, 2, 459–483. J. Kuha ja J. Jackson (2014). The Item Count Method for Sensitive Survey Questions: Modelling Criminal Behaviour. *Journal of the Royal Statistical Society, C*, 63, 321-341. T.F. Crossley, T. Schmidt, P. Tzamourani ja J.K. Winter (2021): Interviewer Effects and the Measurement of Financial Literacy. *Journal of the Royal Statistical Society, A*, 184, 150–178.

[86]F. Galton (1866): On an Error in the Usual Method of Obtaining Meteorological Statistics of the Ocean. Report of the British Association for the Advancement of Science, 36, 16–17. http://galton.org/essays/1860-1869/galton-1866-rba-error-ocean.pdf (haettu 11.9.2020). Bulmer (2003, 31).

[87]Salsburg (2001, 170).

[88]C.D. Haines, E.M. Rose, K.J. Odom ja K.E. Omland (2020): The Role of Diversity in Science: A Case Study of Women Advancing Female Birdsong Research. *Animal Behaviour*, 168, 19–24. K.J. Odom, M.L. Hall, K. Riebel, K.E. Omland ja N.E. Langmore (2014): Female Song is Widespread and Ancestral in Songbirds. *Nature Communications*. DOI: 10.1038/ncomms4379.

[89]Helsingin Sanomat 5.1.2020 (`https://www.hs.fi/kotimaa/art-2000006363708.html`; haettu 5.1.2020).

[90]Topelius (1905b, luku Sattumako vai kaitselmus).

[91]Freedman (1978, 302–304) sekä `https://en.wikipedia.org/wiki/The_Literary_Digest`, `https://en.wikipedia.org/wiki/George_Gallup` ja `https://en.wikipedia.org/wiki/Gallup_company` (haettu 16.3.2016).

[92]Robinsonin (1937) mukaan jälkimmäisen otoksen koko oli 125 000. C.E. Robinson (1937): Recent Developments in the Straw-Poll Field. *Public Opinion Quarterly*, 1, 45–56.

[93]Student (1931): The Lanarkshire Milk Experiment. *Biometrika*, 23, 398–406.

[94]FT Magazine 14.4.2016 (`https://www.ft.com/content/2e43b3e8-01c7-11e6-ac98-3c15a1aa2e62`; haettu 30.12.2020). M. Tolvi (2020): The Weekend Effect and Readmissions in the Helsinki and Uusimaa Hospital District. Väitöskirja. Lääketieteellinen tiedekunta. Helsingin yliopisto.

[95]Spiegelhalter (2015, 8, 10, 72 ja 317) ja `http://www.hiteresearchfoundation.org` (haettu 16.3.2016).

[96]L. Demirdjian (11.12.2017): The Promise: When Truth Overshadows Power. `bit.ly/2m7KaxB` (Significance-lehden nettisivut). *Internet Movie Database*'in *The Promise* -elokuvan sivu `http://www.imdb.com/title/tt4776998/`. Haut 8.2.2018. Ruotsi katsoo kansanmurhan tapahtuneen. Yhdysvaltojen edustajainhuone, kongressi ja presidentti tunnustivat kansanmurhan marras-joulukuussa 2019 ja huhtikuussa 2021. (Helsingin Sanomat 3.11. ja 13.12.2019, Yle 24.4.2021. `https://www.hs.fi/ulkomaat/art-2000006294195.html`, `https://www.hs.fi/ulkomaat/art-2000006341721.html`, `https://yle.fi/uutiset/3-11900750`. Haut 24.4.2021.) Euroopan parlamentti, paavi Franciscus ja ylipäänsä 31 maata tunnustavat kansanmurhan. Kansanmurhan muistomerkkejä on pari sataa. (`https://en.wikipedia.org/wiki/Armenian_genocide`, `https://en.m.wikipedia.org/wiki/List_of_Armenian_genocide_memorials`. Haut 11.6.2022.)

[97]K. Aromaa ja M. Heiskanen (2006): Kansainvälinen rikosuhritutkimus vaikeuksissa. *Haaste*, 3/2006, 16–17.

[98]M. Keyriläinen (2019): Perhevapaan vaikutus naisten urakehitykseen kielteisempi korkeakoulutetuilla. *Tieto & Trendit*, 5.12.2019. `http://www.stat.fi/tietotrendit/artikkelit/2019/perhevapaan-vaikutus-naisten-urakehitykseen-kielteisempi-korkeakoulutetuilla/`. Helsingin Sanomat 5.12.2019: Monet suomalaisnaiset kokevat raskaussyrjintää: tutkimus selvitti, miten työn ja perheen yhdistäminen onnistuu Suomessa. `https://www.hs.fi/kotimaa/art-2000006333527.html`. Haut 9.12.2019.

[99]`http://www.pewforum.org/2019/02/06/the-evolution-of-pew-research-centers-survey-questions-about-the-origins-and-development-of-life-on-earth/` (haettu 10.2.2019).

[100]Eurostatin tiedote 4.2.2019 sivulla `https://ec.europa.eu/eurostat/news/news-releases`, `https://www.hs.fi/kotimaa/art-2000006007176.html` ja `http://www.stat.fi/tietotrendit/blogit/2019/maiden-valisia-terveyseroja-tulee-vertailla-varoen/` (viitaukset 21.2.2019).

[101]P. Suhonen (2016): Epäluotettavia tutkimustuloksia asenteista eutanasiaan. *Yhteiskuntapolitiikka*, 81, 732–733.

[102]H. Lehti ja M. Laaninen (2021): Perhetaustan yhteys oppimistuloksiin Suomessa PISA- ja rekisteriaineistojen valossa. *Yhteiskuntapolitiikka*, 86, 520–532.

[103]`http://www.vaestoliitto.fi/tieto_ja_tutkimus/vaestontutkimuslaitos/seksologinen_tutkimus/suomalaisten-seksuaalisuus-finse` ja `news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx`. Viittaukset 13.1.2018. G.E. Derrick, P.-Y. Chen, T. van Leeuwen, V. Larivière ja C.R. Sugimoto (2022): The Relationship between Parenting Engagement and Academic Performance. *Scientific Reports*, 12, 22300. `https://doi.org/10.1038/s41598-022-26258-z`. M.-L. Halko, M. Kaustia ja E. Alanko (2011): The Gender Effect in Risky Asset Holdings. *Journal of Economic Behaviour & Organisation*, 83, 66–81.

[104]M. Ritakorpi, M. Kaunonen, M. Kaila, S. Paldanius ja N. Seilo (2019): Sähköiseen terveysky-

selyyn vastaamatta jättäneet yliopisto-opiskelijat — katoanalyysi. *Sosiaalilääketieteellinen aikakauslehti*, 56, 42–52.

[105]Helsingin Sanomat 22.2.2018 (`https://www.hs.fi/hyvinvointi/art-2000005575276.html`; haettu 22.2.2018) ja J. Kopra (2018): Statistical Modelling of Selective Non-Participation in Health Examination Surveys. Department of Mathematics and Statistics, University of Jyväskylä. Report 164.

[106]A. Kalenius (2022): Koulutustasovertailu työvoimatutkimuksen uudistuttua. Opetus- ja kulttuuriministeriön politiikka-analyysejä 2022:1. Suomalaisten koulutustaso. Opetus- ja kulttuuriministeriö. `http://urn.fi/URN:ISBN:978-952-263-804-5` (haettu 17.10.2022).

[107]L.E. Jones ja N.R. Ziebarth (2016): Succesful Scientific Replication and Extension of Levitt (2008): Child Seats Are Still No Safer than Seat Belts. *Journal of Applied Econometrics*, 31, 920–928. S.D. Levitt (2008): Evidence that Seat Belts Are as Effective as Child Safety Seats in Preventing Death for Children Aged Two and Up. *Review of Economics and Statistics*, 90, 158–163.

[108]Tilastokeskuksen Palkat ja työvoimakustannukset sivu `https://www.tilastokeskus.fi/tup/suoluk/suoluk_palkat.html` (haettu 21.12.2019).

[109]A. Keinänen ja T. Tukiainen (2010): Laittomista työtaistelutoimenpiteistä tuomittujen hyvityssakkojen määräytyminen työtuomioistuinkäytännössä. Edilex 2010/15 (`http://www.edilex.fi/lakikirjasto/asiantuntijakirjoitukset/7008`; haettu 7.5.2016). Kiitän Anssi Keinästä aineiston antamisesta 16.5.2016.

[110]R. W. Hamming (1973): *Numerical Methods for Scientists and Engineers, 2. laitos*. Dover.

[111]Parametrit eivät saa sijaita niin sanotun parametriavaruuden reunapisteessä, jotta tekstissä osoitetut estimaattoreiden jakaumat suurilla havaintomäärillä pätisivät. Esimerkiksi binomijakauman parametrin tulee sijaita välillä $(0, 1)$ (olla nollaa suurempi mutta yhtä pienempi). Jakson 9.1 esimerkinkaltaista tilannetta, jossa $\theta = 1$ ja parametriavaruus on $[0, 1]$, ei sallita.

[112]V.-M. Rissanen (2009; toim. ja suom.): *Marcus Tullius Cicero. Keskusteluja Tusculumissa*. Faros. S. 36.

[113]Kuva pohjautuu Alan Arnholtin R-koodiin (`https://raw.githubusercontent.com/alanarnholt/PASWR2E-Rscripts/master/ChapterScripts/chapter08.R`; haettu 23.3.2016). Koodi on Ugarten ym:iden (2016) kirjan luvusta 8.

[114]S. Nieuwenhuis, B.U. Forstmann ja E.-J. Wagenmakers (2011): Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance. *Nature Neuroscience*, 14, 1105–1107.

[115]Esim. "$n\pi > 10$ ja $n(1-\pi) > 10$", "$n\pi(1-\pi) \geq 10$" tai "$n\pi - 3[n\pi(1-\pi)]^{1/2} > 0$ ja $n\pi + 3[n\pi(1-\pi)]^{1/2} < n$").

[116]Agresti (2013, 604).

[117]`https://en.wikipedia.org/wiki/Rule_of_three_(statistics)` (haettu 24.3.2016). `https://www.hs.fi/kotimaa/art-2000009012531.html` (haettu 19.8.2022).

[118]http://www.stat.ufl.edu/ aa/cda/R/one-sample (haettu 11.2.2021).

[119]Esim. "$n_1 > 30$ ja $n_2 > 30$", "$n_i\hat{\pi}_i \geq 5$ ja $n_i(1-\hat{\pi}_i) \geq 5$, $i = 1, 2$" sekä "$y_i \geq 10$ ja $n_i - y_i \geq 10$, $i = 1, 2$".

[120]M. Huttunen, M. Husso ja J. Hietamäki (2015): Sukupuoliero parisuhdeväkivallan yleisyydessä ja sen havaitsemisessa lasten ja nuorten näkökulmasta. *Janus*, 23, 109–126. Aineisto on vuonna 2008 kerätystä lapsiuhritutkimuksesta.

[121]Newcombe (2013, 118). Alkuperäistutkimus: Boice ja Monson (1977).

[122]Kasvukäyrät perustuvat vuosina 1983–2008 syntyneiden espoolaislasten kasvutietoihin. Kasvukäyrät on uusinut professori Leo Dunkelin tutkimusryhmä Itä-Suomen yliopistossa. Keskipituus ja otoskeskihajonta ovat tarkalleen miehille 181.042 ja 6.0609 cm ja naisille 167.4764 ja 5.4047 cm. (Professori Leo Dunkell, henkilökohtainen tiedonanto 16.3.2010.) Tyttöjen pituus ei ole välttämättä normaalijakautunut (A. Pere (2000): Comparison of Two Methods for Transforming Height and Weight to Normality. *Annals of Human Biology*, 27, 35 − 45). Teoreettisesta näkökulmasta pituus ei voi olla normaalijakautunut, sillä pituudet sijoittuvat välille (0,3) metriä eli pituuden vaihteluväli on rajoitettu.

[123]A. Latvala, R. Kuja-Halkola, C. Almqvist, H. Larsson ja P. Lichtenstein (2015): A Longitudinal

Study of Resting Heart Rate and Violent Criminality in More Than 700 000 Men. *JAMA Psychiatry*, 72, 971–978. A. Latvala, R. Kuja-Halkola, C. Rück, B.M. D'Onofrio, T. Jernberg, C. Almqvist, D. Mataix-Cols, H. Larsson ja P. Lichtenstein (2016): Association of Resting Heart Rate and Blood Pressure in Late Adolescence With Subsequent Mental Disorders. A Longitudinal Population Study of More Than 1 Million Men in Sweden. *JAMA Psychiatry*, 73, 1268–1275. O. Choy, A. Raine, P.H. Venables, ja D.P. Farrington (2017): Explaining the Gender Gap in Crime: The Role of Heart Rate. *Criminology*, 55, 465–487. Kiitän Olivia Choyta artikkelissa julkaistua tarkemmista tiedoista 18.8.2017.

[124]Teoksessa P. McCullagh (2023): *Ten Projects in Applied Statistics*. Springer. S. viii

[125]Kuva pohjautuu paljolti Crawleyn (2013, 206 ja 277) R-koodeihin.

[126]Kuva 11.3 pohjautuu R-koodiin 9.1 Ugarten ym:iden (2016) kirjassa.

[127]Tässä esitetyn teorian puitteissa ei tule myöskään todeta, että p-arvo on hylkäysvirheen todennäköisyys. Testin merkitsevyystaso eli hylkäysvirheen todennäköisyys on valittu etukäteen ennen testin suorittamista. Merkitsevyystaso on kiinteä luku; p-arvo on satunnaismuuttuja. Kirjallisuudessa käytetään myös määritelmää, jossa samaistetaan merkitsevyystaso ja p-arvo. Tällöin molemmat ovat satunnaismuuttujia eikä noudateta edellä esitetyn kaltaista ajatusrakennelmaa. (Merkinnöissä edellä ei ole eroteltu p-arvoa satunnaismuuttujana ja sen toteumana.) Teoriakehystä laajentamalla p-arvon tulkintaa voidaan avartaa tekstissä esitetystä (esim. Martin ja Liu 2014).

[128]D. Chavalarias, J.D. Wallach, A.H.T. Li, ja J.P.A. Ioannidis (2016): Evolution of Reporting P Values in the Biomedical Literature, 1990–2015. *Journal of the American Medical Association*, 315, 1141–1148.

[129]J. Kallio (suom. ja taustoittanut 2014): *Mestari Kongin keskustelut. Kungfutselaisuuden ydinolemus*. Gaudeamus. S. 94.

[130]E. Valkama ja M. Litmala (2006): Lasten huoltoriidat käräjäoikeuksissa. OPTL:n julkaisuja 224. `https://helda.helsinki.fi/handle/10138/152456` (haettu 13.2.2021).

Tutkimuslaitos pyysi käräjäoikeuksia lähettämään tiedot ratkaisuistaan sitä mukaa kuin riidat ratkaistiin. Tutkimuslaitos sai 565:n riidan asiapaperit, joista puutteellisina tai muusta syystä hylättiin 36 tapausta (n. 6 %). Laitoksen tutkijat arvioivat olleen ilmeistä, että käräjäoikeudet eivät olleet systemaattisesti lähettäneet kaikkia ratkaistuja päätöksiä tutkimuslaitokselle. Vain osa riidoista koski lasten asumista. Aineiston keruutapa ei ole ongelmaton, mm. koska tutkimuksesta kerrottiin etukäteen käräjäoikeuksille. Tutkimuksen tekijät huomauttavat itse muista epävarmuustekijöistä aineiston edustavuudessa mutta pitävät aineistoa kuitenkin melko luotettavana.

[131]Lapset määrättiin asumaan isän luona 27.3 %:ssa, äidin luona 65.2 %:ssa ja 7.5 %:ssa päätöksistä (127 havaintoa) molemmilla (lapset "jaettiin" tai määrättiin "vuoroasumisesta").

[132]http://users.stat.ufl.edu/ aa/cda/R/one-sample (haettu 11.2.2021).

[133]M. Palo-Repo (2015): Lasten huolto- ja asumisriidat Helsingin hovioikeudessa 2003–2006. Pro gradu -tutkielma (tilastotiede). Valtiotieteellinen tiedekunta. Helsingin yliopisto. `https://helda.helsinki.fi/handle/10138/155254` (haettu 2.1.2016).

[134]Mellin (1996, 176).

[135]Taulukko ja tehtävän tiedot ovat Bulmerin (2013, 154–156) kirjasta. Alkuperäislähde olisi Galton (1869).

[136]Jos n on satunnainen, voidaan päättely ehdollistaa reunalukumäärille n_{i+}. Näin aineistoa voidaan analysoida kuin se olisi saatu riippumattomalla multinomiaalisella otannalla. Jos kaikki reunalukumäärät n_{i+} ja n_{+j} ovat kiinteitä tai päättely ehdollistetaan niille, solulukumäärät noudattavat hypergeometrista jakaumaa. Pienillä havaintomäärillä voi käyttää siihen tukeutuvaa päättelyä kuten keski-p-korjattua Fisherin eksaktia testiä.

[137]E. Callaway (2013): Uncertainty on Trial. *Nature*, 502, 3, 17–18. `http://www.bloomberg.com/news/articles/2013-03-05/ex-intermune-ceo-harkonen-s-wire-fraud-conviction-upheld`, `http://www.iltalehti.fi/ulkomaat/2014090718638829_ul.shtml`, `https://www.hs.fi/tiede/art-2000002827076.html`, `http://suomenkuvalehti.fi/jutut/tiede/miljoonia-maksanut-jatkotutkimus-paljastui-roskaksi-yha-sairaammat-potilaat-julistettiin-parantuneiksi` ja `https://www.hs.fi/hyvinvointi/art-2000005775378.html`. Haut 31.7.2018.

[138]P. Gøtzsche (2014): *Tappavat lääkkeet ja järjestäytynyt rikollisuus. Näin lääketeollisuus on*

turmellut terveydenhoidon. Sitruuna-kustannus. Kirja on Britannian lääkäriliiton palkitsema. Helsingin yliopiston psykiatrian professori Erkki Isometsä arvostelee Gøtzschen näkemyksiä Helsingin Sanomissa 10.3.2017 (`http://www.hs.fi/tiede/art-2000005122873.html`; haettu 12.3.2017).

[139] `http://yle.fi/uutiset/laakefirmojen_tutkimukset_rahoittajilleen_myonteisia/1894597`. F.T. Bourgeois, S. Murthy ja K.D. Mandl (2010): Outcome Reporting Among Drug Trials Registered in ClinicalTrials.gov. *Annals of Internal Medicine*, 153, 158–167.

[140] J. Antfolk ja A. Sjölund (2018): High Parental Investment in Childhood is Associated with Increased Mate Value in Adulthood. *Personality and Individual Differences*, 127, 144–150.

[141] Agresti ja Finlay (2009, 195–196).

[142] Lapsiasiavaltuutetun vuosikirja 2014. Eriarvoistuva lapsuus. Lasten hyvinvointi kansallisten indikaattorien valossa. Lapsiasiavaltuutetun toimiston julkaisuja 2014:3. `http://lapsiasia.fi/wp-content/uploads/2014/12/Vuosikirja-2014.pdf` (haettu 28.3.2015). S. 81.

[143] Kiitän professori emeritus Seppo Laaksosta PISA-tunnuslukujen laskemisesta 24.4.2013. Professori emeritus Laaksonen on korjannut aineistoa niin, että sitä voidaan pitkälti käsitellä niin kuin se olisi tehty yksinkertaisella satunnaisotannalla, vaikka PISA-tutkimus tehdään monimutkaisemmalla otantamenetelmällä.

[144] `https://en.wikipedia.org/wiki/Lawrence_Summers` (haettu 4.4.2017), Wainer (2007) ja A. Ukkola, J. Metsämuuronen ja M. Paananen (2020): Alkumittauksen syventäviä kysymyksiä. Kansallinen koulutuksen arviointikeskus. Julkaisut 10:2020.

[145] Keskivanhemman määritelmä löytyy esimerkiksi Wikipediasta (`http://en.wikipedia.org/wiki/Midparent`; haettu 23.4.2016).

[146] Stigler (1999) kertoo Secristin tutkimuksesta tarkemmin. Ks. myös Wallis ja Roberts (1956, luku 8).

[147] Lisää esimerkkejä on Wainerin (2005) kirjan luvussa 10 sekä Wallisin ja Robertsin (1956) kirjan luvussa 8. Ks. myös Friedman (1992), Jerrim ja Vignoles (2013) sekä Goldstein (2015).

[148] Aineisto on luotu olettaen sekä isien että poikien keskipituudeksi ja -hajonnaksi 181.0 cm ja 6.06 cm. Nämä ovat suomalaisten miesten pituustietojen mukaiset luvut (professori Leo Dunkell, henkilökohtainen tiedonanto 16.3.2010). Isien ja poikien pituuden korrelaatioksi on oletettu 0.5, joka vastaa melko tarkasti todellista korrelaatiota (esim. Pearson ja Lee 1903). Aineisto, kuviot ja analyysit alla tehtiin MASS-paketin (Venables ja Ripley 2002) sekä Crawleyn (2013, 451) koodin avulla.

[149] Syy jakaa JNS $n-2$:lla eikä n:llä liittyy näin saadun σ^2:n estimaatin teoreettisiin ominaisuuksiin ja jakson 13.4.2 jakaumateoriaan. Muistisääntö on, että estimoinnissa havaintoja "menetetään" yksi kutakin estimoitua parametria kohti. Vrt. s^2:n kaava (13.12) monen selittäjän regressiomallin (13.10) yhteydessä.

[150] Tekstissä käytetään samaa merkintää selitysasteelle populaatiossa ja otoksessa.

[151] Daly, M.C., A.J. Oswald, D. Wilson ja S. Wu (2011): Dark Contrasts: The Paradox of High Rates of Suicide in Happy Places. *Journal of Economic Behavior & Organization*, 80, 435–442. Kiitän Stephen Wuta aineiston luovuttamisesta 4.3.2020. Aineisto eroaa artikkelin kuvasta 2. Siinä maat on nimetty väärin: Itä-Saksan po. Portugali, Portugalin po. Kreikka ja Kreikan po. Italia. Numeeriset analyysit alla täsmäävät artikkelissa raportoitujen kanssa. Artikkelissa ei raportoida jäännöksen keskihajontaa. Se on laskettu tähän erikseen.

[152] Yhden selittäjän tilanteessa jälkimmäinen ehto merkitsee, että kaikki ainoan selittäjän havainnot eivät saa olla samoja. Tätä tilannetta sivuttiin jakson 13.4.1 lopussa. Monen selittäjän tilanteessa ei esimerkiksi ole sallittua, että yhden selittäjän arvot olisivat toisen selittäjän arvoja kerrottuna jollain luvulla.

[153] Systemaattisen osan geometrinen tulkinta ei ole yhtä helppo kuin yhden selittäjän regressiossa, jossa aineistoon sovitettiin suora. Mikäli selittäjiä on kaksi, sovitetaan aineistoon kaksiulotteinen taso.

[154] Jos selittäjän lukuarvot ovat kiinteitä lukuja, vastaavaa populaatiokorrelaatiota ei ole olemassa.

[155] A. Keinänen ja A. Pakarinen (2009): Palkkasyrjinnän todistaminen tilastollisesti. Edilex 2009/5. Keinänen ja Pakarinen eivät selitä, miten työsuhteen laatu -muuttuja on luotu. Selitys yllä on päätelty. Keinänen ja Pakarinen eivät raportoi F-testisuuretta, mutta se on laskettavissa artikkelin

tietojen perusteella.

[156]Taulukkokirjaa käytettäessä täytyy tyytyä esimerkiksi jakauman $\mathsf{F}(4, 100)$ kriittisiin arvoihin. Sen 0.95. kvantiili on 2.463.

[157]Taulukoita käytettäessä testisuuretta voi verrata $\mathsf{t}(100)$-jakaumaan. Sen 0.95. kvantiili on 1.660. Ero $\mathsf{N}(0, 1)$-jakauman 0.95. kvantiiliin 1.645 on pieni, koska havaintoja on paljon.

[158]Kiitän Helsingin yliopiston urapalveluita ja erityisesti LuK Tuukka Kangasta aineiston luovuttamisesta 18.4.2017.

[159]S. Kannasoja (2013): Nuorten sosiaalinen toimintakyky. Jyväskylän yliopisto. Yhteiskuntatieteellinen tiedekunta. Jyväskylä Studies in Education, Psychology and Social Research, 484. Kiitän Sirpa Kannasojaa väitöskirjansa aineiston luovuttamisesta 29.9.2014. Kannasoja (mt.) selittää muuttujien merkitykset ja luonnin. Vaste saa arvoja 0.1:n välein, ja sen vaihteluväli on rajoitettu. Nämä seikat sivuutetaan analyyseissa. Kirjan analyysit ovat itsenäisiä eivätkä väitöskirjasta poimittuja. Esimerkin t-arvot on otettu R:n regressiokäskyn palautteesta. Ne eivät siksi ole aivan samoja kuin esimerkin pyöristetyistä luvuista laskettavissa olevat t-arvot.

[160]Vastaavia kuvia löytyy Foxin (2016) kirjasta. Se on oiva lähde teemaan ja on ollut avuksi jakson kirjoittamisessa.

[161]Lineaarinen estimaattori riippuu selitettävistä havainnoista lineaarisesti eli on muotoa $\sum_{i=1}^{n} b_i y_i$ (b_i:t määräytyvät selittävien muuttujien arvoista). PNS-estimaattori on tällainen estimaattori. Asia jää toteamuksen varaan, koska PNS-estimaattoreille ei johdeta tässä analyyttisia kaavoja.

[162]Havainto 1822 ei ilmeisesti ole kirjausvirhe. Aikasarjassa helmi-kesäkuun lämpötiloista Turussa on tavaton lämpöpiikki n. 1822. J. Halonen (2006): Reconstructions of Past Climates from Documentary and Natural Sources in Finland since the 18th Century. Helsingin yliopisto. Geologian laitoksen julkaisuja D9. S. 15.

[163]T. Vigen (2015): *Spurious Correlations*. Hachette Books.

[164]Demidenkon (2020, 680–681) vastaava empiirinen esimerkki innoittaa analyysit tässä.

[165]Aineisto on Baddeleyn ja Barrowcloughin (2009) kirjasta (s:t 47 ja 63) ja perustuu Yhdistyneiden kansakuntien keräämiin tilastoihin.

[166]Jos selittäjät korreloivat, se vaikeuttaa vertailua lisää.

[167]Esimerkki juontaa kuviosta XII.3 julkaisussa *Selvityksiä raiskausrikoksista*. Selvityksiä ja ohjeita 13/2012. Oikeusministeriö. Kiitän Oikeuspoliittisen tutkimuslaitoksen suunnittelija Petri Danielssonia aineiston luovuttamisesta 11.3.2014. Aineisto kuvataan tarkemmin harjoitustehtävässä.

[168]Esimerkin inspiraatio on V. Berneliuksen (2013) väitöskirja: Eriytyvät kaupunkikoulut. Tutkimuksia 2013:1. Helsingin kaupungin tietokeskus. Helsinki.

[169]C.M. Reinhart ja K.S. Rogoff (2010): Growth in a Time of Debt. *American Economic Review: Papers & Proceedings 100*, 573–578. EKP:n vuosikertomus 2010 (s. 77). Euroopan keskuspankki. `https://www.suomenpankki.fi/fi/media-ja-julkaisut/julkaisut/euroopan-keskuspankin-julkaisuja/vuosikertomus/2011/ekpn-vuosikertomus-2010/`. Euroopan unionin talouskomissaari Olli Rehnin puhe 28.2.2013 "Deeper Integration in the Eurozone and Britain's Place in Europe". Policy Network conference, Lontoo. `http://europa.eu/rapid/press-release_SPEECH-13-174_en.htm`. T. Herndon, M. Ash ja R. Pollin (2014): Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38, 257–279. M. Maziarz (2017): The Reinhart-Rogoff Controversy as an Instance of the 'Emerging Contrary Result' Phenomenon. *Journal of Economic Methodology*, 24, 213–225. `https://en.wikipedia.org/wiki/Growth_in_a_Time_of_Debt`. (Viittaukset 1.2.2018.)

[170]Lisää aiheesta on artikkelissa L. Törnqvist, P. Vartia ja Y. Vartia (1985): How Should Relative Changes Be Measured? *The American Statistician*, 39, 43–46.

[171]Taloustieteen opiskelijat huomaavat, että β on y:n jousto x:n suhteen: (y:n muutos prosenteissa)/(x:n muutos prosenteissa).

[172]$\mathsf{R}^2 = 1 - \mathsf{JNS}/\mathsf{KNS}$. Selitysasteiden vertailu perustuu JNS:n vaihteluun mallien välillä KNS:n ollessa kiinteä. Jos vaste logaritmoidaan, muuttuu KNS, eikä vertailu mallien välillä ole mahdollista.

[173]Lause on englanninnettu kahdessa lähteessä: W. Isaacson (2019): Leonardo da Vinci – The

Scientist. *Substantia*, 3, 59-64. https://doi.org/10.13128/Substantia-636 (haettu 23.8.2021). D. Wootton (2015): *The Invention of Science*. Allen Lane. Isaacson muotoilee *test of experiment*; Wootton (mts. 24) *test of experience*. Suomennoksessa on pyritty lähemmäs jälkimmäistä sanamuotoa, joka kuvannee Leonardo da Vincin ajattelua paremmin.

[174]Kuvien R-koodin ydin: https://stackoverflow.com/questions/26431555/plot-a-function-with-several-arguments-in-r (haettu 7.4.2020).

[175]Muistutus: Riippuu sovelluksesta, onko todella kyse syy-seuraussuhteesta vai vain vasteen ja selitettävän välisestä tilastollisesta yhteydestä.

[176]http://fi.wikipedia.org/wiki/Challenger ja http://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster (viittaukset 11.4.2020).

[177]Aineisto ja suurelta osin esimerkki ovat Agrestin (2013) oppikirjasta (sen tehtävä 5.6). Aineiston fahrenheitasteet on muutettu esimerkkiä varten celsiusasteiksi. Challenger-onnettomuutta ja -aineistoa käsitellään laajemmin Bilderin ja Loughinin (2015) sekä Davisonin (2003) oppikirjoissa sekä artikkeleissa S. Dalal, E. Fowlkes, ja B. Hoadley (1989): Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*, 84, 945–957, T. J. Fisher ja M. W. Robbins (2019): A Cheap Trick to Improve the Power of a Conservative Hypothesis Test. *American Statistician*, 73, 232–242 ja R. Gnanadesikan (1990): Looking Ahead: Cross-Disciplinary Opportunities for Statistics. *American Statistician*, 44, 121–125.

[178]Piirtämisessä apuna oli Crawleyn (2013, 654) R-koodi.

[179]P. Pere, E. Lilja ja A. Sobolev (2017): Maahanmuuttajien mielikuva naisten suosimisesta tuomioistuimissa. *Lakimies*, 115, 90–99. Aineistoa on sittemmin korjattu ja voittoa selitetty useammalla muuttujalla.

[180]Demidenko (2020, xv).

[181]Wilcox (2021, 319–322), osoittaa varoittavia esimerkkejä hyvin vinojen jakaumien ja hyvin poikkeavien oudokkien vaikutuksesta t-testisuureen jakaumaan.

[182]Ilmaisu ei ole matemaattisesti tarkka. Jos otoskertymäfunktio on diskreetti ja kertymäfunktio jatkuva, niin erotuksen suurinta arvoa ei ole olemassa. Suurimman arvon pienin yläraja, *supremum*, on, ja siihen päätekstissä viitataan "suurimmalla arvolla".

[183]M. Serra-Garcia ja U. Gneezy (2021): Nonreplicable Publications Are Cited More than Replicable Ones. *Science Advances*, 7: eabd1705. Kiitän Marta Serra-Garciaa aineiston jakamisesta 2.6.2021.

[184]https://fi.wikipedia.org/wiki/JÃdÃdkiekon_SM-liigakausi_2007-2008 ja https://fi.wikipedia.org/wiki/JÃdÃdkiekon_SM-liigakausi_2008-2009 (haettu 31.10.2023).