Tilastotieteen perusteet

Esim. Arvostettu juoma-asiantuntija ekonomisti E osallistuu juomien makutestiin, jossa voi saada voi saada arvonimen

Melko Suuri Maistaja (MSM *), Suuri maistaja (SM**) tai Erittäin Suuri Maistaja ESM***).

Tulos määräytyy kokeella, jossa maistaja yrittää erottaa Erityisjuoman kolmen vastaavanlaisen juoman joukosta.

Maistaminen toistetaan 13 kertaa ja tulos on oikein tunnistettujen kertojen määrä.

Maistaja on

MSM*, jos tuloksen saavuttaa arvaamalla alle 5 % tapauksista

SM**, jos tuloksen saavuttaa arvaamalla alle kerran sadasta

ESM***, jos tuloksen saavuttaa arvaamalla vain alle 0.1 % tapauksista.

E tunnistaa 13 yrityksestä 10 kertaa juoman oikein.

Tulos 10 oikein on hyvä, mutta mihin se riittää?

(Tästä myöhemmin: hypoteesien testaaminen)

Esim. Alueen kotitalouksista poimittiin markkinatutkimusta varten 1500 suuruinen otos. Otokseen osuneista

kotitalouksista 30 % käyttää hyödykettä Ö.

- Intuitio sanoo, että otoksesta saatu arvo 30 % ”pyörii oikea tuloksen ympärillä”.

- Kuinka hyvin otoksesta saatu arvo ”vastaa” oikeaa arvoa? Siis

kuinka **tarkka** ja **luotettava** tämä arvio on?

(Tästä myöhemmin: estimointi, luottamusväli)

Mainoskampanjan jälkeen tehtiin uusi tutkimus, jossa

1400 suuruisessa **otoksessa** 34 % kotitalouksista käytti Ö:tä.

- Voidaanko tämän perusteella väittää, että Ö:n käyttö on suurentunut kaikkien **perusjoukon** kotitalouksien joukossa?

Jos näin väitetään, kuinka suuri on **riski,** että otoksista havaittu neljän %-yksikön ero onkin vain ”sattuman leikkiä” eikä Ö:n käyttäjien määrä ole todellisuudessa muuttunut kaikkien kotitalouksien joukossa?

(Tästä myöhemmin: suhteellisen osuuden testi)

Esim. Tutkittiin uuden kolesterolilääkkeen vaikutusta.

Kolmen kuukauden ajan lääkittiin **koeryhmän** 200 ja **vertailuryhmän** 150 korkeasta kolesterolista kärsivää koehenkilöä.

Kokeen jälkeen olivat

koeryhmän 200:n lääkettä saaneen kolesteroliarvojen

keskiarvo $\bar{x}_1 = 5.2$ mmol ja keskihajonta $s_1 = 1.2$ mmol ja

vertailuryhmän 150:n lumelääkettä saaneen

keskiarvo $\bar{x}_2 = 5.7$ mmol ja keskihajonta $s_2 = 1.6$ mmol

- Uskallatko päätellä tämän perusteella, että paranemisessa on eroa ja päästää lääkkeen markkinoille?

- Kuinka suuri on **riski**, että aineistossa havaittu ero 0.5 mmol keskiarvoissa on tullut kokeessa ”vain sattumalta”.

Jos lääkkeellä ei todellisuudessa ole mitään vaikutusta paranemiseen, tämä kyllä paljastuu suuresta käyttäjien määrästä. Tällaisella virheellä on luonnollisesti vakavia taloudellisia ym. seurauksia.

(testaamisongelma)

Esim. Yrityksen Y työntekijöistä poimittiin otos, jonka avulla selvitettiin suhtautumista tulospalkkauksen käyttöön ottoon yrityksessä. Saatiin tulokset:

	Kielteinen	Neutraali	Myönteinen
Alle 40-v.	96	174	159
Yli 40-v.	117	155	122

- Voidaanko tämän perusteella päätellä, riippuuko suhtautuminen tulospalkkaukseen työntekijän iästä?

(testaamisongelma)

Esim. Aiemman tutkimusaineiston perusteella arvioidaan, että

sairaudelle S altistava tekijä on 15 % väestöstä.

Alttiuden pikaseulontaa varten on kehitetty automaattinen menetelmä, ja selvitellään, onko se kelvollinen laboratorioille myytäväksi.

Pikatestin tuloksia verrattiin tarkan (kalliimman) menetelmän tuloksiin ja havaittiin, että menetelmä hälyttää positiivisella tuloksella

- 90 prosentissa tapauksista, jos altistava tekijä todella on olemassa,

- mutta myös 2 prosentissa tapauksista, joissa tekijää ei olekaan.

Tilaajat vaativat, että

1) positiivisista hälytyksistä saa korkeintaan 5 % olla vääriä (siis pelästytetään turhaan altistumaton ihminen) ja

2) negatiivisista tuloksista saa korkeintaan 2 % olla vääriä (siis menetelmä ei huomaa altistusta).

- Onko menetelmä myyntikelpoinen?

Tällaisiin kysymyksiin saadaan vastaukset **tilastollisen päättelyn** ja sen pohjana olevan **todennäköisyyslaskennan** avulla.

Empiirinen ilmiö voi olla **deterministinen** tai **satunnaisilmiö**.

- **Deterministinen**: Vain systemaattiset tekijät vaikuttavat siihen.

- **Satunnaisilmiö**: Myös sattuma vaikuttaa siihen.

Sattuma vaikuttaa empiiriseen ilmiöön, jos tulokset eivät ole aina samat, vaikka ilmiö toistuisi samoissa olosuhteissa.

Tilastotiedettä tarvitaan erityisesti satunnaisilmiöiden tutkimisessa.

Yksittäisen satunnaisilmiön (-kokeen) tulosta ei voida ennustaa tarkkaan.

Usein kuitenkin havaitaan, että sattuman käyttäytymisessä on selvää lainalaisuutta.

Todennäköisyyslaskenta ja tilastotiede tutkivat tätä.

Empiiriseen ilmiöön voi liittyä

- vain systemaattisia tekijöitä ja ilmiön käyttäytyminen pystytään täsmällisesti ennustamaan.

Esim. annuiteettilaina maksuerän suuruus, kun lainasumma, maksuajan pituus ja korkoprosentti on sovittu.

- vain satunnaisia tekijöitä

Esim. ensi vuonna ensimmäisenä Suomessa syntyvän lapsen sukupuoli

- sekä systemaattisia että satunnaisia tekijöitä

Esim. Ensi vuonna syntyvien lasten lukumäärä, bkt:n suuruus, yrityksen liikevaihto, EU:n kannattajien osuus 1000 suuruisessa otoksessa

Tilastotieteen ”materia” on tilastoyksiköistä mitattujen muuttujien arvojen sisältämä informaatio, ja tilastollinen analyysissä **informaatio tiivistetään tarkoituksenmukaisella tavalla empiirisestä aineistosta**.

Havaintoaineistoa tarkastellaan tilastollisten operaatioiden avulla erilaisista ”näkökulmista”, joiden valintaan vaikuttavat:

- Aineistolle tehdyn operaation tulos kertoo siitä ainakin kohtuullisen hyvin jotain suoraan (**kuvaileva tilastotiede**)

Esim. keskiarvo, keskihajonta, korrelaatiokerroin

- Menetelmän on oltava matemaattisesti niin hyvin toimiva, että analyysissa ”nähdään käsiteltävän aineiston läpi” tutkittavan ilmiön rakenteisiin. (**tilastollinen päättely**)

Kun poimitaan (esim.) 1000 suuruinen otos kuluttajista ja tutkitaan otokseen osuneiden kulutuksen rakennetta, sosioekonomista asemaa, asenteita yms. tiivistetään otoksen sisältämä informaatio erilaisiksi tunnusluvuiksi (esim. muuttujien arvojen keskiarvot, keskihajonnat, jne.).

Ne kertovat suoraan jotain juuri näistä 1000 ihmisestä. Kuitenkin varsinaisen päämääränä on, että näiden ”tiivistelmien” avulla tehdään arviota kaikista perusjoukossa olevista kuluttajista.

Aina kun aineisto hankitaan otannan (tai koesuunnittelun) avulla, sattuma vaikuttaa siihen.

Tilastollisen analyysin päämääränä on **systemaattisten ja satunnaisten tekijöiden tunnistaminen ja erottaminen toisistaan** satunnaisilmiöissä.

- **Deskriptiivisessä tilastotieteessä** käsitellään menetelmiä, joiden avulla tiivistetään informaatio (otannassa) aineistoon osuneista tilastoyksiköistä mitatuista muuttujien arvoista.

- Jotta tämän tunnuslukuihin tiivistetyn informaation avulla voidaan tehdä yleistyksiä, **tilastollista päättelyä**, koko perusjoukkoon on tunnettava

sattuman käyttäytymiseen liittyvät lainalaisuudet.

Niitä tutkii todennäköisyyslaskenta.

1 Todennäköisyyslaskentaa

1.1 Empiirinen todennäköisyyskäsite

Esim. Aiotaan ostaa arpa.

- Arpalaatikossa 1. on 100 arpaa, joista 15 on voittavia,
- arpalaatikossa 2. on 200 arpaa, joista 30 voittaa, ja
- arpalaatikossa 3. on 467 arpaa, joista 60 on voittavia.

Mistä laatikosta arpa kannattaa valita?

Vaihtoehtojen vertailussa tärkeät asiat ovat

- voittavien arpojen lukumäärä,
- arpojen kokonaismäärä ja
- valinnan umpimähkäisyys.

2. vaihtoehdossa voittavien arpojen määrä on kaksinkertainen

1. vaihtoehtoon verrattuna, mutta myös kokonaismäärä on kaksinkertainen.

Silloin voittavia arpoja on samassa suhteessa eli 15 % kummassakin laatikossa ja 1. ja 2. vaihtoehto ovat samanarvoisia.

Tällöin sanotaan, että voiton **todennäköisyys** on yhtä suuri.

- 3. laatikossa voittavia on eniten, mutta myös kokonaismäärä on suurin. Tässä voittavia arpoja on ”huonommassa suhteessa”

$60/467 \approx 12.8$ %,

ja voiton todennäköisyys on pienempi kuin laatikoissa 1. ja 2.

Todennäköisyyslaskennan tehtävänä on

sattuman käyttäytymiseen liittyvien lainalaisuuksien kuvaaminen täsmällisten käsitteiden avulla.

Empiiristä ilmiötä, johon vaikuttaa sattuma, sanotaan

satunnaisilmiöksi tai **satunnaiskokeeksi**.

Satunnaiskokeessa

- yksittäisen satunnaiskokeen lopputulosta ei pystytä ennustamaan, mutta

- tuloksissa on selvää säännönmukaisuutta, kun koe toistuu useita kertoja samoissa olosuhteissa.

Esim. Markkinatutkimusta varten aiotaan poimia kunnan asukkaista otos. Taustatietona tiedetään (mm.), että väestöstä on 45 % naisia.

- 1., 2., 3., jne. otokseen umpimähkään arvottavan yksittäisen henkilön sukupuolta ei pystytä varmasti ennustamaan,

- mutta otannan edistyessä naisten osuus asettuu 45 % tuntumaan.

Todennäköisyyden käsitteen avulla kuvataan (mitataan)

”kuinka suuri mahdollisuus” jollain tapahtumalla on olla tuloksena jossain satunnaiskokeessa.

Todennäköisyyden suuruus esitetään suhteellisena osuutena (prosenttilukuna).

”Kuinka monta mahdollisuutta 100:sta.”

Esim. a) Laatikossa on 200 arpaa, joista 30 on voittavia.

Kun satunnaiskokeena on arvan valinta umpimähkään,

voiton todennäköisyys on 30/200 = 0.15 = 15 %.

b) Kunnan väestöstä on 45 % naisia.

Todennäköisyys, että umpimähkään valittava henkilö on nainen, on 45 %.

Tällöin ajatellaan:

- Vaikka kenenkään yksittäisen (otokseen) valittavan henkilön sukupuolta ei pystytä etukäteen varmasti ennustamaan, niin

- ”45 %:ssa tapauksista on odotettavissa” nainen.

c) Kun aiotaan heittää rahaa, voidaan ilmeisesti järkevästi sanoa,

että on noin 50-prosenttinen mahdollisuus eli todennäköisyys, että saadaan tuloksena klaava.

Näissä esimerkeissä todennäköisyyksien suuruudet pääteltiin **ilman empiiristä kokemusta** pelkästään asetelman rakenteen perusteella.

Tällaisesta päättelyä sanotaan **klassisen todennäköisyyden** laskemiseksi.

Tilanne voi olla myös toisenlainen:

Esim. Väestötilastosta nähdään syntyneistä lapsista:

Vuosi	Lapsia	Poikia	Poikien %-	osuus
1990	65549	33531	0.5115	≈ 51 %
1991	65395	33261	0.5086	≈ 51 %
1992	66731	34147	0.5086	≈ 51 %
1993	64826	33001	0.5091	≈ 51 %
1994	65231	33180	0.5087	≈ 51 %

Tässä tilastoista saatavan **empiirisen tiedon perusteella** on järkevää päätellä, että

noin 51 % todennäköisyydellä seuraava syntyvä lapsi on poika.

Tässä todennäköisyys arvioidaan **frekventistisen todennäköisyyden** käsitteen avulla.

Vaikka tässä lähtökohta on erilainen, ajatellaan kuten edellä:

- Vaikka kenenkään yksittäisen tulevaisuudessa syntyvän lapsen sukupuolta ei pystytä (ilman lisätutkimuksia) etukäteen varmasti ennustamaan, niin

- ”On odotettavissa, että” noin 51 % tulevaisuudessa syntyvistä lapsista tulee olemaan poikia.

Todennäköisyyksien arvioiminen voi olla järkevää ja hyödyllistä myös **subjektiivisina todennäköisyyksinä**, millä tarkoitetaan

”rationaalisen henkilön” arviota tapahtuman esiintymismahdollisuuden suuruudesta.

Esimerkiksi liiketoiminnassa hyvin ”sisällä” olevilla voi olla järkevä ”tuntuma” toimialansa suhdannevaiheiden muuttumisesta.

Edellä tapahtumien todennäköisyydet on nähty suoraan.

Tilanne voi olla kuitenkin monimutkaisempi:

Esim. Kuinka todennäköistä on, että

- arpajaisesimerkissä ostettavista viidestä arvasta vähintään 1 voittaa,

- perheeseen syntyy 2 poikaa ja 3 tyttöä,

- 1000 suuruiseen otokseen tulee osumaan yli 50 % naisia, kun heitä on kunnassa (perusjoukossa) 45 %?

Tällaisiin kysymyksiin saadaan varsin helposti vastaukset todennäköisyyskäsitteen perusominaisuuksien avulla.

Usein todennäköisyyttä laskettaessa on selvitettävä

”kuinka monta vaihtoehtoa on olemassa”.

Esim. Kuinka todennäköistä on saada lotossa kaikki 7 numeroa (palloa)

40:stä oikein tai 5 oikein?

Tätä varten on selvitettävä

- montako lottoriviä (7 pallon yhdistelmää) on yhteensä ja

- montako sellaista riviä on, joissa on 5 oikeaa ja 2 muuta numeroa?

Tämä ei näy suoraan, mutta vastaus saadaan helposti **kombinatoriikan** avulla.

Huom. Edellisen esimerkin ”lotto-ongelma” ei ole kovin kiinnostava sellaiselle, joka ei pelaa rahapelejä.

Tässä ovat kuitenkin pelkistettyinä ensimmäiset askeleet otoksesta havaittavan tilastoyksiköiden jonkin ominaisuuden

suhteellisen osuuden suuruuden määräytymisen lainalaisuuksien tutkimista varten:

- Tarkasteltavana on **äärellinen perusjoukko** (tässä N=40 palloa), jossa on **kahdenlaisia tilastoyksiköitä** (tässä 7 oikeaa ja 33 muuta).

- Perusjoukosta poimitaan **otos palauttamatta** (tässä n=7 palloa).

- Selvitetään kuinka todennäköistä on, että otokseen osuu juuri tietty yhdistelmä ”toisenlaisia” ja ”toisenlaisia” tilastoyksiköitä (esim. tässä k = 5 oikeaa ja n - k = 2 muuta).

- Muutenkin tarvittavien perusteiden opettelu pelkistettyjen esimerkkien avulla voi koetella kärsivällisyyttä, ennen kuin päästään ”varsinaiseen asiaan”.

Tilastollisen päättelyn logiikka ei kuitenkaan avaudu ilman todennäköisyyslaskennan alkeita.

- Todennäköisyyslaskenta alkoi kehittyä 1600-luvulla kortti- ja noppapelien tutkimisesta, ja siitä asti nämä uhkapelurin työkalut ovat olleet apuna ja rasitteena todennäköisyyslaskennan sääntöjen opettelemisessa.

- Tällaiset esimerkit toimivat sinänsä hyvin, mutta tässä niiden käyttämistä yleensä vältetään. Samat rakenteet näkyvät ”otantatilanteista”, joissa tosin aluksi ”otos” ja perusjoukko, josta otos poimitaan, ovat hyvin pieniä.

1.2 Joukko-oppia ja kombinatoriikkaa

Joukko-oppia

tarvitaan todennäköisyyslaskennan merkinnöissä apuvälineenä.

Joukko on kokoelma joitakin objekteja a, b, c, …

Näitä olioita sanotaan joukon **alkioiksi.**

Merkintä: Esim. alkioiden a, b, c ja d joukosta käytetään merkintää

{a, b, c, d}.

Joukot nimetään (tarvittaessa) yleensä isoilla kirjaimilla:

A, B, C … tai A_1, A_2, A_3,…

Esim. Voi olla A = {1, 2, 3, 4, 5}, B = {3, 5, 7, 8} jne.

Jos x on joukon A alkio, sanotaan, että **x kuuluu joukkoon** A ja tästä käytetään merkintää x∈ A.

Jos y ei ole A:n alkio eli **y ei kuulu joukkoon** A merkintä on y $\notin$ A.

Edellisessä esimerkissä 2∈ A, mutta 2 $\notin$ B.

Tyhjä joukko Φ on joukko, jossa ei ole yhtään alkiota.

Joukot A ja B ovat samoja eli A = B, jos niissä on täsmälleen samat alkiot.

Alkioiden esittämisjärjestyksellä ei ole väliä.

Esim. {a, b} = {b, a}

Jokainen alkio merkitään joukkoon vain kerran.

Esim. {a, a, b} = {a, b}

Perusjoukko on kaikkien mahdollisten mielenkiinnon kohteena olevien alkioiden muodostama joukko, jota merkitään usein E:llä.

Esim. Aiotaan heittää noppaa.

(Hyvin pelkistetysti) mahdolliset tulosvaihtoehdot voidaan esittää perusjoukkona

$E = \{1, 2, 3, 4, 5, 6\}$

Osajoukko:

Joukko B on joukon A osajoukko, jos jokainen B:n alkio kuuluu myös A:han. Merkintänä käytetään $B \subset A$.

Jos C ei ole A:n osajoukko, merkitään $C \not\subset A$.

Esim. **Otanta palauttamatta:**

Perusjoukkona E ovat täysikäiset suomalaiset, joista arvotaan (palauttamatta) 1000 henkilöä haastateltaviksi. Jokainen tällainen järjestämätön **otos**

A, B, C, … on E:n osajoukko.

Tämä on tärkeä lähtökohta **otantatutkimuksessa**. Kun otantatilanne hahmotetaan tarkoituksenmukaisella tavalla, todennäköisyyslaskennan menetelmien avulla voidaan (monissa tärkeissä tilanteissa) selvittää otoksesta tehtävien päätelmien **tarkkuus** ja **luotettavuus**.

Usein joukon A määrittelyssä esitetään ehto, joka perusjoukon E alkioiden on toteutettava, jotta ne kuuluvat A:han. Tämä merkitään:

$A = \{x \in E \mid$ ”ehto”$\}$

Esim. (nopan heitto)

”Tulos on yli 4” voidaan esittää (, jos sitä pidetään tarpeellisena)

joukkona $B = \{ x \in E \mid x > 4\}$ $(= \{5, 6\})$

Esim. (jatkoa, otanta palauttamatta)

$A = \{x \in E \mid$ alle 30 % otoksen henkilöistä x omistaa älypuhelimen$\}$

(Tässä E:n osajoukon määrittely ei ole 1-käsitteinen, vaan kertoo, minkälaisista 1000 suuruisista otoksista A ollaan kiinnostuneita.)

Komplementti: Joukon A komplementissa A^c ovat ne perusjoukon E alkiot, jotka eivät kuulu A:han.

(Komplementista käytetään myös merkintää $\overline{A}$.)

Esim. (jatkoa, nopan heitto)

$B^c = \{ x \in E \mid x \leq 4\}$ $(= \{1, 2, 3, 4\})$

Yhdiste (unioni):

Joukkojen A ja B yhdisteeseen $A \cup B$ kuuluvat ne alkiot, jotka kuuluvat joko A:han **tai** B:hen (tai molempiin).

Yhdisteen ja leikkauksen määritelmät yleistyvät myös useammalle kuin kahdelle joukolle.

Esim. Perusjoukko E = {1, 2, 3, …, 10}, A = {3, 8, 9} ja B = {4, 8, 9, 10}.

$A \cup B$ = {3, 4, 8, 9, 10}.

Leikkaus:

Joukkojen A ja B leikkaus $A \cap B$ koostuu alkioista, jotka kuuluvat molempiin joukkoihin A **ja** B.

Esim. (jatkoa edelliseen) A = {3, **8**, **9**} ja B = {4, **8**, **9**, 10}.

$A \cap B$ = {**8**, **9**}

Joukko-opillinen erotus: Joukkojen A ja B erotukseen A\B kuuluvat ne alkiot, jotka kuuluvat A:han, mutta eivät B:hen.

Esim. (jatkoa edelliseen) A = {3, 8, 9} ja B = {4, 8, 9, 10}.

A\B = {3}.

Tulojoukko:

Joukkojen A ja B **tulojoukko** (karteesinen tulo) A x B on joukko

$A \times B = \{(x_1, x_2) \mid x_1 \in A \wedge x_2 \in B\}$.

Siis A x B:n alkioina ovat kaikki **järjestetyt alkioparit (-jonot)**, missä **ensin** on A:n alkio ja **sitten** B:n alkio.

Esim. Joukko A = {1, 2} ja B = {1, 3, 5}.

A x B = {(1, 1), (1, 3), (1, 5), (2, 1), (2, 3), (2, 5)}

ja

B x A = {(1, 1), (1, 2), (3, 1), (3, 2), (5, 1), (5, 2)}

(Piirrä arvoparit kuvioina.)

Siis yleensä $A \times B \neq B \times A$.

Esim. (jatkoa) Noppaa aiotaan heittää kaksi kertaa.

Kun $E_1 = \{1, 2, 3, 4, 5, 6\}$ esittää 1. heiton mahdollisia tuloksia

ja $E_2 = \{1, 2, 3, 4, 5, 6\}$ vastaavasti 2. heiton tulosta,

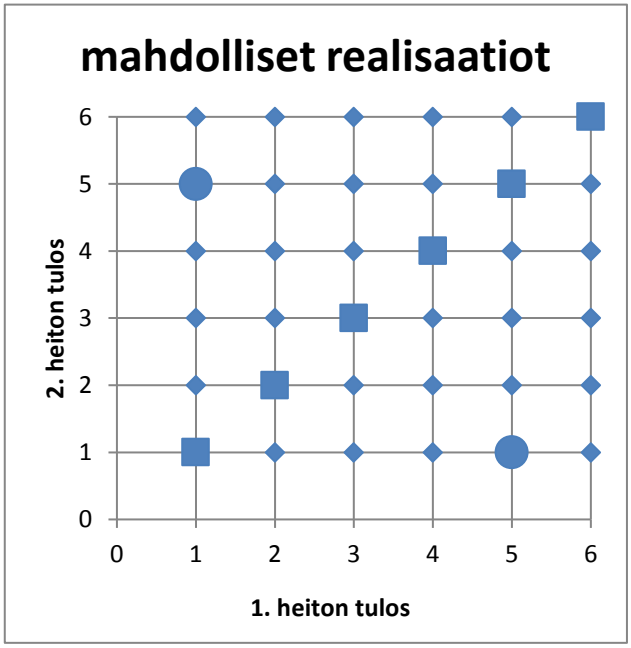
kokeen mahdollisia realisaatioita voidaan kuvata tulojoukkona

$E_1 \times E_2 = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (6,6)\}$

Esim. osajoukko, jossa molemmilla heitoilla tullaan saamaan sama tulos

$$B = \{(x_1,x_2) \in E_1 \times E_2 \mid x_1 = x_2\} = \{(1,1), \ldots, (6,6)\}$$

↙



Järjestykselle annetaan tietoisesti merkitys. Esim. (1,5) on eri realisaatio kuin (5,1).

Tulojoukon määritelmä yleistyy myös kahta useammalle joukolle.

Kombinatoriikkaa

Kombinatoriikan avulla vastataan kysymykseen:

Kuinka monella tavalla jokin operaatio tai operaatiosarja voidaan tehdä?

Usein tällaisten kysymysten selvittäminen pelkistyy tilanteeksi:

Tutkitaan äärellistä joukkoa $E = \{a_1, a_2, \ldots, a_n\}$ ja halutaan tietää

- kuinka monta **erilaista** k:n ($k \leq n$) alkion muodostamaa **jonoa** eli **permutaatiota** näistä alkioista voidaan tehdä tai

- kuinka monta **erilaista** k:n alkion (järjestämätöntä) **osajoukkoa** eli **kombinaatiota** voidaan valita.

Esim. Neljän henkilön (perus-)joukko {A, B, C, D}

a) Kuinka monella tavalla tästä joukosta voidaan

- valita 2 henkilöä ja

- asettaa heidät jonoon?

(Montako erilaista kahden suuruista **järjestettyä otosta** voidaan valita **palauttamatta**?)

Tässä vaihtoehdot voidaan luetella:

AB BA AC CA AD DA BC CB BD DB CD DC, yhteensä 12 kpl.

b) Kuinka monta erilaista kahden henkilön muodostamaa osajoukkoa voidaan valita?

(Montako erilaista kahden suuruista **järjestämätöntä otosta** voidaan poimia **palauttamatta**?)

Nyt vaihtoehtoja on vähemmän, kun henkilöiden järjestyksellä ei ole väliä:

{A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}, yhteensä 6 kpl.

Jos (perus-)joukko ja valittavien alkioiden määrä (otoskoko) on suurempi, vaihtoehtoja ei pystytä luettelemaan.

Esimerkiksi (vain) 15 henkilön joukosta voidaan poimia palauttamatta **erilaisia** 4:n suuruisia

- järjestettyjä otoksia (jonoja) 36760 kpl ja

- järjestämättömiä otoksia (osajoukkoja) 1365 kpl.

Jonojen ja osajoukkojen lukumäärät voidaan kuitenkin laskea helposti:

Kertolasku- ja yhteenlaskuperiaate

Kertolaskuperiaate

Esim. Kaupungista X vie 3 tietä kaupunkiin Y, josta on edelleen 5 tietä kaupunkiin Z. (Piirrä kuvio tilanteesta.)

Kuinka monta **erilaista** reittiä on X:stä Y:n kautta Z:an?

- Jokaista kolmea X → Y aloitusvaihtoehtoa kohti

- voidaan jatkaa Y → Z väli viidellä tavalla,

joten vaihtoehtoja on yhteensä 3·5 = 15 kpl.

Kuten esimerkissä **kertolaskuperiaatteen** mukaan:

- Jos operaatiot 1. ja 2. tehdään **peräkkäin** (tai muuten kiinteästi toisiinsa liittyen) ja

- 1. operaatio voidaan tehdä n_1 tavalla ja 2. operaatio n_2 tavalla,

niin operaatiosarja

1. **ja** 2. kokonaisuutena ajatellen voidaan tehdä $n_1 \cdot n_2$ tavalla.

Esim. (jatkoa) Neljän henkilön joukosta {A, B, C, D} voidaan valita

1. henkilö 4 tavalla (1. operaatio) **ja** 2. henkilö 3 tavalla (2. operaatio),

jolloin yhdistetty operaatio:

”Valitaan jonoon 1. henkilö **ja** 2. henkilö jäljelle jääneistä”

voidaan tehdä 4·3 = 12 tavalla.

Kertolaskuperiaatteen avulla saadaan myös kahta pitempien peräkkäisten operaatioiden sarjojen vaihtoehtojen lukumäärät.

Esim. (jatkoa) 15 henkilöstä voidaan tehdä

$15 \cdot 14 \cdot 13 \cdot 12 = 32760$ erilaista 4 pituista jonoa (järjestettyä otosta).

Esim. (Informaation määrä merkkijonoissa.)

Vaiteliaiden maan kielessä käytetään vain aakkosia a, b, c, d, ja e.

Montako sellaista nelikirjaimista sanaa on olemassa, joissa

a) samaa kirjainta saa käyttää vain kerran ja

b) samaa kirjainta saa käyttää myös useammin?

a) Nelikirjaimisen sanan kirjoittaminen on neljän peräkkäisen ”operaation” sarja, jossa

1. operaatio: Valitaan jonoon 1. kirjain. (5 mahdollisuutta) **ja**

2. operaatio: Valitaan 2. kirjain jäljelle olevista. (4 mahdollisuutta) **ja**

3. operaatio: Valitaan 3. kirjain jäljelle olevista. (3 mahdollisuutta) **ja**

4. operaatio: Valitaan 4. kirjain jäljelle olevista. (2 mahdollisuutta).

Kertolaskuperiaatteen mukaan sanoja (jonoja) on $5 \cdot 4 \cdot 3 \cdot 2 = 120$ kpl.

b) Jos samaa kirjainta voi käyttää useammin, jonoja on $5 \cdot 5 \cdot 5 \cdot 5 = 5^4 = 625$ kpl.

Rinnakkaisien operaatioiden yhdistämiseen käy varsin itsestään selvä

Yhteenlaskuperiaate

Esim. Kaupungista X vie 2 tietä kaupunkiin A ja 4 tietä kaupunkiin B.

(Piirrä kuvio.)

Kuinka monella tavalla X:stä voidaan matkustaa joko A:han **tai** B:hen?

Selvästi vaihtoehtoja on $2+4 = 6$.

Kuten esimerkissä **yhteenlaskuperiaatteen** mukaan:

- Jos operaatiot 1. ja 2. tehdään **rinnakkain** ja niitä ei voi tehdä samanaikaisesti ja

- 1. operaatiossa on n_1 ja 2. operaatiossa n_2 vaihtoehtoa,

niin operaatioyhdistelmä ”tehdään 1. **tai** 2. operaatio”

voidaan tehdä n_1**+**n_2 tavalla.

Esim. Vaiteliaiden maan vaikeasti lausuttavassa kielessä käytetään vain aakkosia a, b, c, d, ja e. Sanat ovat 1-, 2-, 3-, 4- tai 5-kirjaimisia.

Kuinka monta sellaista ”sanaa” on olemassa, joissa

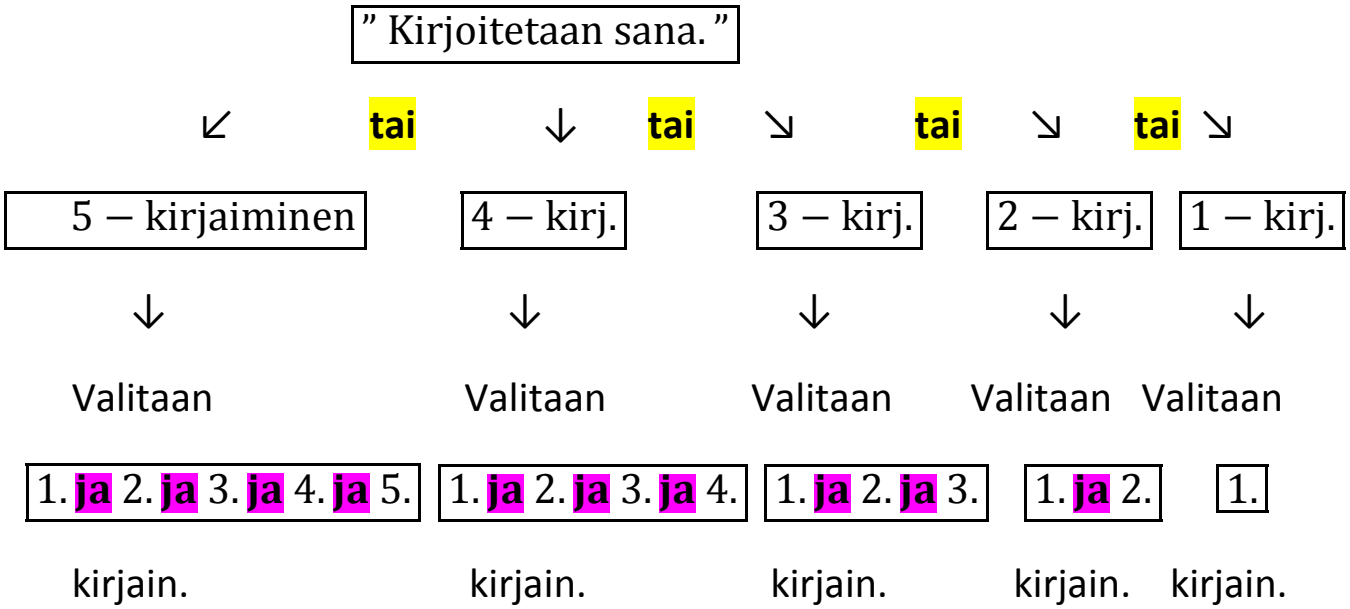
a) samaa kirjainta käytetään vain kerran ja

b) samaa kirjainta saa käyttää myös useammin?

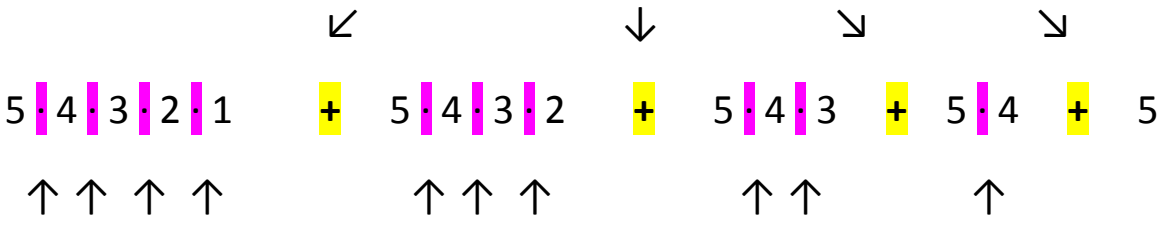
a) Operaatio ”Kirjoitetaan sana” muodostuu

- viidestä **rinnakkaisesta** ”operaatiosta”,

- jotka taas ovat **peräkkäisten** ”operaatioiden” sarjoja:



Yhteenlaskuperiaate



Kertolaskuperiaate

= 120 + 120 + 60 + 20 + 5

= 325 ”sanaa”.

b) Kun samaa kirjainta saa käyttää uudelleen, hajoaa tehtävä samalla tavalla

- ”ylätasolla” samoiksi rinnakkaisiksi vaihtoehdoiksi, joiden lukumäärät lasketaan yhteen ja

- ”alatasolla” peräkkäisiksi kirjainten valinnoiksi, joiden lukumäärät kerrotaan keskenään.

”Sanoja” on silloin

$5 \cdot 5 \cdot 5 \cdot 5 \cdot 5 + 5 \cdot 5 \cdot 5 \cdot 5 + 5 \cdot 5 \cdot 5 + 5 \cdot 5 + 5 = 3905$ kpl.

Muistisääntö: Kun tehdään

1. ”jotain” **ja** ”jotain”, vaihtoehtojen lukumäärät **kerrotaan** keskenään, ja

2. ”jotain” **tai** ”jotain”, vaihtoehtojen lukumäärät **lasketaan yhteen**.

Samanlaiset säännöt ovat myös todennäköisyyksien yhdistelyssä.

Mm. edellisen esimerkin tehtävässä on sama rakenne kuin kokonaistodennäköisyyden kaavassa.

Yksittäisissä laskutehtävissä permutaatioiden lukumäärät voidaan laskea kertolaskusäännön avulla vastaavalla tavalla kuin edellä.

Jatkoa varten sama on voitava esittää myös yleisin merkinnöin (siis ”kaavamuodossa”):

Merkintöjen lyhentämiseksi on sovittu, että merkintää **n!** käytetään tulosta

n! = n·(n-1) ·(n-2)· … ·2·1

↖

Luetaan: ”n-kertoma”

Lisäksi on järkevää sopia, että **0! = 1**.

Esim. 5! = 5·4·3·2·1 = 120.

Edellisessä esimerkissä nähtiin, että 5 aakkosta a, b, c, d, ja e voidaan

asettaa jonoon 5! = 5·4·3·2·1 = 120 tavalla.

Samalla tavalla kertolaskuperiaatteesta seuraa yleisesti, että
n erilaisesta alkiosta voidaan tehdä jonoja *n!* kpl:

Täytetään

1. 2. 3. … (n-1):s n:s paikka

↓ ↓ ↓ ↓ ↓

n · (n-1) · (n-2)· … · 2 · 1 = n! kpl.

Esim. Edellä laskettiin myös, että 5 aakkosesta a, b, c, d, ja e voidaan tehdä 3-kirjaimisia sanoja, joissa kaikki kirjaimet ovat erilaisia,

$$60 = 5\cdot4\cdot3 = \frac{5\cdot4\cdot3\cdot\mathbf{2\cdot1}}{\mathbf{2\cdot1}} = \frac{5!}{(5-3)!} \text{ kpl.}$$

Myös yleisesti kertolaskuperiaatteesta seuraa, että

n erilaisesta alkiosta voidaan tehdä k:n alkion permutaatioita

$$\frac{\boldsymbol{n}!}{(\boldsymbol{n}-\boldsymbol{k})!} \text{ kpl:}$$

Täytetään

1. 2. 3. … k:s paikka

↓ ↓ ↓ ↓

n · (n-1) · (n-2)· … · (n-(**k-1**)) = n·(n-1)·(n-2)· … ·(n-k+1)

↑

k-1 alkiota ”käytetty”

lavennus

↙

$$= \frac{n\cdot(n-1)\cdot(n-2)\cdot \ldots \cdot(n-k+1)\cdot \mathbf{(n-k)\cdot(n-k-1)\cdot \ldots \cdot 2\cdot 1}}{\mathbf{(n-k)\cdot(n-k-1)\cdot \ldots \cdot 2\cdot 1}} = \frac{n!}{(n-k)!} .$$

Esim. 5 aakkosesta a, b, c, d, ja e voidaan tehdä 2-kirjaimisia sanoja, joissa kaikki kirjaimet ovat erilaisia,

$\frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{3\cdot2\cdot1} = 5\cdot4 = 20$ kpl, mikä edellä pääteltiin lyhyemminkin.

Permutaatioiden lukumäärän antava sääntö auttaa (järjestämättömien) osajoukkojen lukumäärän laskemisessa:

Esim. Kuinka monta erilaista 4 suuruista osajoukkoa (kombinaatiota) voidaan valita 15 henkilöstä?

- Edellä pääteltiin 4:n pituisten **jonojen lukumäärä** kertolaskuperiaatteen avulla ”täyttämällä” jonon 1. ja 2. ja 3. ja 4. paikka aina jäljellä olevista henkilöistä.

15·14·13·12 = 32760, mikä saadaan myös

edellisestä säännöstä $\frac{15!}{(15-4)!} = \frac{15!}{11!} = \frac{15\cdot14\cdot13\cdot12\cdot\cancel{11!}}{\cancel{11!}}$ = 5040.

- **Jonojen lukumäärä** määrä voidaan päätellä myös toisella tavalla kahden ”peräkkäisen” operaation avulla:

”Valitaan 4:n suuruinen (järjestämätön) osajoukko 15 henkilöstä.”

(Juuri tämän operaation vaihtoehtojen määrä x halutaan selvittää.)

ja

”Järjestetään valitut 4 henkilöä jonoon.”

Tämä voidaan tehdä 4! tavalla.

Kertolaskuperiaatteen mukaan jonoja on yhteensä x · 4! kpl.

Silloin on

x · 4! = $\frac{15!}{(15-4)!}$, josta saadaan kombinaatioiden lukumäärä

x = $\frac{15!}{(15-4)!\cdot4!} = \frac{15!}{11!\cdot4!} = \frac{15\cdot14\cdot13\cdot12\cdot\cancel{11}!}{\cancel{11!}\cdot4\cdot3\cdot2\cdot1}$ = 1365 kpl.

Lausekkeesta $\frac{15!}{(15-4)!\cdot 4!}$ käytetään merkintää $\binom{15}{4}$ (Luetaan:”15 yli 4:n”),

Siis 15 suuruisesta perusjoukosta voidaan saada palauttamatta poimimalla $\binom{15}{4}$ = 1365 erilaista 4:n suuruista **järjestämätöntä otosta**.

Samalla tavalla päätellään yleisesti:

Joukosta E, jossa on n (erilaista) alkiota voidaan valita ***erilaisia k:n alkion*** osajoukkoja eli

kombinaatioita $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ ***kpl.***

Erikoistapauksia:

- Jos n:stä alkiosta valitaan kaikki n kpl, niin valinta voidaan tietenkin tehdä vain yhdellä tavalla. Toisaalta myös binomikertoimesta

$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!\mathbf{0}!} = \frac{\cancel{n!}}{\cancel{n!}\cdot \mathbf{1}} = 1$ saadaan sama oikea tulos, kun

tarkoituksenmukaisella tavalla on sovittu 0! = 1.

- Myös $\binom{n}{0} = \frac{n!}{0!(n-0)!} = 1$ eli ”n:stä alkiosta voidaan jättää valitsematta yhtään kappaletta” vain yhdellä tavalla.

Edellisten sääntöjen avulla siis voidaan laskea erilaisten mahdollisten **otosten lukumäärä**, kun perusjoukko on äärellinen.

Otannassa käytetään yleensä merkintöinä:

äärellisen perusjoukon E koko on N ja otoskoko on n.

Kun perusjoukosta E, jossa on N tilastoyksikköä, poimitaan **palauttamatta** n:n suuruinen otos, on mahdollisia

järjestämättömiä otoksia

perusjoukon koko

↓

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

otoskoko ↑ ↑ E:hen jäävien tilastoyksiköiden määrä

ja **järjestettyjä otoksia** $\frac{N!}{(N-n)!}$ kpl.

Huom. Kombinatoriikassa vaihtoehtojen erilaisuus tarkoittaa, että

- kombinaatiot ovat erilaisia, jos niissä on edes yksi erilainen alkio, ja

- permutaatiot ovat erilaisia, jos niissä on edes yksi erilainen alkio tai edes yksi eri järjestys.

Otannassa palauttamatta otoksia voidaan tarkastella ”tilanteen mukaan” joko kombinaatioina tai permutaatioina.

Otannassa palauttaen otokseen ”arvotaan”

- ensin 1. tilastoyksikkö, joka palautetaan perusjoukkoon,

- sitten 2. samalla tavalla jne.

ja otokset ovat poimintatavasta johtuen **järjestettyjä**.

Esim. ”Hattuun” pannaan 15 henkilön nimilaput ja

poimitaan palauttaen 4 suuruinen otos palauttaen.

Siis

valitaan umpimähkään 1. nimilappu, katsotaan se ja palautetaan ”hattuun”.

2. ja 3. ja 4. henkilö arvotaan samalla tavalla.

Kertolaskuperiaatteen mukaan

sattuma voi arpoa erilaisia otoksia (jonoja, permutaatioita)

$15 \cdot 15 \cdot 15 \cdot 15 = 15^4 = 50625$ kpl.

Esim. 15 henkilöstä 9 harrastaa kuntosaliliikuntaa (H).

Tästä (hyvin pienestä) 15 henkilön perusjoukosta aiotaan poimia (myös hyvin pieni) 4:n suuruinen **otos palauttaen**.

Kuinka monta sellaista (siis järjestettyä) otosta voidaan saada, joissa on 2 kuntosalin käyttäjää?

Otokset ovat permutaatioita.

Tässä tilanteessa pystytään luettelemaan jonot, joissa on 2 kuntosaliliikunnan harrastajaa (H) ja 2 ei-harrastavaa (E):

HjaHjaEjaE tai HEHE tai HEEH tai EHHE tai **EHEH tai EEHH**

(yhteenlaskuperiaate)

↓ ↓ ↓ ↓ ↓

$9 \cdot 9 \cdot 6 \cdot 6 + 9 \cdot 6 \cdot 9 \cdot 6 + 9 \cdot 6 \cdot 6 \cdot 9 + 6 \cdot 9 \cdot 9 \cdot 6 + 6 \cdot 9 \cdot 6 \cdot 9 + 6 \cdot 6 \cdot 9 \cdot 9$

↑ ↑ ↑ ↑↑↑ ↑↑↑ ↑↑↑ ↑↑↑ ↑↑↑

(kertolaskuperiaate)

$= 2916 + 2916 + 2916 + 2916 + 2916 + 2916 = 17496$

Tulos voidaan päätellä helpomminkin:

- Sellaisia jonoja, joissa on 2 harrastaa ja 2 ei-harrastajaa **joissain juuri tietyissä kohdissa jonoa**, on 2916 (= esim. $9 \cdot 9 \cdot 6 \cdot 6$) kpl.

- 4 pituisesta jonosta ”voivat valikoitua paikat” 2 harrastajalle $\binom{4}{2} = \frac{4!}{2! \cdot 2!}$ = 6 tavalla.

2 harrastajaa ja 2 ei-harrastaja jonoon

↘ ↓ ↓ ↙

1. 2. 3. 4. paikka

- Silloin ”suotuisia” jonoja on

$\binom{4}{2} \cdot 9 \cdot 9 \cdot 6 \cdot 6 = 17496$ kpl.

Järjestettyjä otoksia on yhteensä 15·15·15·15 = 50625 kpl,

joten erilaisia otoksia, joissa kaikissa on 2 harrastajaa, on

17496 /50625 = 0.3456 = 34.56 %.

Silloin voidaan ilmeisesti järkevästi päätellä:

- Jos poimittaisiin palauttaen tällaisesta (vain) 15 henkilön perusjoukosta hyvin monta (vain) 4:n suuruista otosta,

- niin ”noin 34.6 %:ssa näistä otoksista olisi odotettavissa”
2 kuntosaliliikunnan harrastajaa.

- Silloin sanotaan, että 34.6 % :n suuruisella **todennäköisyydellä** otokseen osuu 2 harrastajaa.

- Jos tällaiset ”arpajaiset” todella järjestetään erittäin monta kertaa, empiirinen kokemus todella vahvistaa tämän päättelyn:

Niiden otosten osuus, joissa on 2 harrastajaa ja 2 ei-harrastajaa ”pyörii” 34.6 %:n tuntumassa.

- Tällaisesta äärimmäisen suppeasta perusjoukosta voidaan todella tehdä otosten poimimiskoe ”nimilappuja” umpimähkään poimimalla. Tietokoneen avulla **satunnaislukuja** käyttämällä voidaan helpommin **simuloida** (jäljitellä) otosten poimimista. Silloin myös perusjoukon koko ja otoskoko voivat olla niin suuria, kuin todellisessa otantatutkimuksessa on.

Esim. (jatkoa edelliseen)

15 henkilön perusjoukosta, jossa on 9 kuntosalin käyttäjää, aiotaan poimia 4:n suuruinen **otos palauttamatta**.

Nyt otokset voidaan tulkita joko permutaatioina tai kombinaatioina:

a) Otokset ajatellaan neljän henkilön **jonoiksi**.

- Kuinka monta erilaista otosta voidaan saada yhteensä?

Samalla tavalla kuin edellä otoksen poimiminen ajatellaan 4 henkilön arpomisena peräkkäin jonon neljään tyhjään paikkaan:

Jonojen kokonaismäärä saadaan kertolaskuperiaatteen avulla:

Valitaan jonoon 1. **ja** 2. **ja** 3. **ja** 4. henkilö aina jäljellä olevista.

↘ ↓ ↓ ↙

1.	2.	3.	4.

↑ ↑ ↑ ↑

$15 \cdot 14 \cdot 13 \cdot 12 = 32760$

erilaista 4 henkilön jonoa, jotka poikkeavat toisistaan ainakin yhden henkilön tai yhden järjestyksen osalta.

- Kuinka monta erilaista sellaista otosta voidaan saada, joissa on 2 liikunnan harrastajaa?

2 harrastajaa ja 2 ei-harrastaja jonoon

↘ ↓ ↓ ↙

1.	2.	3.	4.

- Esim. jonossa

H	H	E	E

↑ ↑ ↑ ↑

$9 \cdot 8 \cdot 6 \cdot 5 = 2160$

erilaista vaihtoehtoa, kun

ensimmäinen harrastajista voi valikoitua 6 tavalla **ja** jälkimmäinen 5 tavalla, **ja** ei-harrastajista ensimmäinen 4 **ja** jälkimmäinen 3 tavalla.

- Olivatpa kaksi harrastajaa kaksi ei-harrastajaa tällaisessa jonossa missä kohtaa tahansa, järjestyksiä on sama määrä 2160.

- Kuten edellä ”sattuma” voi valita 4:stä vapaasta paikasta paikat 2:lle harrastajalle $\binom{4}{2} = \frac{4!}{2!\cdot 2!} = 6$ tavalla.

- Yhteenlaskuperiaatteen mukaan jonoja (tällaisia otoksia) on yhteensä

(2160 + 2160 + 2160 + 2160 + 2160 + 2160 =)

$\binom{4}{2} \cdot 9\cdot 8\cdot 6\cdot 5 = 12960$ kpl.

Tällaisten jonojen suhteellinen osuus on $\frac{12960}{32760} \approx 0.3956$.

Siis jos tässä otos poimitaan palauttamatta, on todennäköisempää saada otokseen 2 harrastajaa kuin otannassa palauttaen.

b) Otokset hahmotetaan 4 henkilön (**järjestämättömiksi**) **joukoiksi**.

- 15 henkilön perusjoukosta sattuma voi valita 4:n suuruisen järjestämättömän otoksen

$\binom{15}{4} = \frac{15!}{4!\cdot(15-4)!}$ = 1365 tavalla.

- Sellaisia otoksia, joissa on 2 harrastajaa, on:

9: sta harrastajasta		6: stä ei − harrastajasta
↓		↓
voi valikoitua 2 kpl	**ja**	voi valikoitua 2 ei − harrastajaa
$\binom{9}{2}$ tavalla		$\binom{6}{2}$ tavalla

Kertolaskuperiaatteen mukaan tällaisia otoksia on

$$\binom{9}{2} \cdot \binom{6}{2} = \frac{9!}{2!\cdot(9-2)!} \cdot \frac{6!}{2!\cdot(6-2)!} = 36 \cdot 15 = 540 \text{ kpl.}$$

- Tällaisten otosten suhteellinen osuus on $\frac{540}{1365} \approx 0.3956$.

Tulos on (tietenkin) sama, mikä saatiin ajattelemalla otokset permutaatioina.

Esim. (jatkoa) Kuinka todennäköistä on saada lotossa a) kaikki 7 numeroa 40:stä oikein ja b) 5 oikein?

Tilanne on samanlainen kuin edellisessä esimerkissä:

- ”Perusjoukossa”, jossa on 40 palloa ja niistä 7 oikeaa (oikeiksi arpoutuvaa) ja 33 muuta palloa,

- poimitaan 7 suuruinen ”otos” palauttamatta.

”Otokset” (mahdolliset lottorivit) voidaan hahmottaa joko järjestettyinä tai järjestämättöminä.

a) Esim. kombinaatioina ajateltuina rivejä on yhteensä

$$\binom{40}{7} = \frac{40!}{7!\cdot(40-7)!} = \frac{40!}{7!\cdot 33!} = \frac{40\cdot 39\cdot 38\cdot 37\cdot 36\cdot 35\cdot 34}{7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1} = 18\ 643\ 560 \text{ kpl},$$

joista vain 1 yhdistelmä on ”oikea” ja täysosuman todennäköisyys on vain ”yksi noin lähes 19 miljoonasta”.

b) Aivan samalla tavalla kuin edellisessä esimerkissä päätellään,

kuinka monta sellaista 7 pallon ”otosta” on, joissa on 5 ”oikeaa” ja 2 muuta palloa:

7: stä oikeasta		33: sta muusta
↓		↓
voi valita 5 kpl	**ja**	voi valita 2 muuta
$\binom{7}{5}$ tavalla		$\binom{33}{2}$ tavalla

Kertolaskuperiaatteen mukaan tällaisia otoksia on

$$\binom{7}{5} \cdot \binom{33}{2} = \frac{7!}{5!\cdot(7-5)!} \cdot \frac{33!}{2!\cdot(33-2)!} = 21 \cdot 528 = 11088 \text{ kpl}$$

ja näitä rivejä on $\frac{11088}{18643560} \approx 0.0005947 \approx 0.06$ % kaikista mahdollisista.

- Arvonnassa (otannan järjestelyissä) pyritään erityisesti huolehtimaan, että arvonta on rehellinen, ja

- ilmeisesti voidaan olettaa, että mikä tahansa 18 643 560:stä numeroyhdistelmästä voi tulla sattuman valitsemaksi ”yhtä hyvin”.

- Silloin voitaneen järkevästi sanoa, että 5 oikein tuloksen saamisen

todennäköisyys on noin 0.06 % eli

”5 oikein on odotettavissa keskimäärin hieman alle 6 kertaa 10 000:sta”.

- Edellisessä esimerkissä ”perusjoukon” koko oli vain 40, mutta silti 7 suuruisia kombinaatioita lähes 19 miljoonaa. Permutaatioita on vielä paljon enemmän (Kuinka paljon?).

- Tavallisesti otantatutkimuksessa perusjoukon koko N sekä otoskoko n ovat hyvin paljon suurempia. Kysymykset ovat kuitenkin samanlaisia kuin edellä.

Esim. N = 100 000 kuluttajan joukosta aiotaan tehdä markkinatutkimus.

- Eräs taustatieto on, että tässä perusjoukossa on 45 % eli 45 000 naista.

- Otoskoko on n = 1000.

- Etukäteen halutaan tarkistaa, että ”riittävän suurella varmuudella” naisten osuus tulee olemaan otoksessa ”lähellä” perusjoukossa olevaa 45 prosenttia.

Esimerkiksi

”kuinka varmasti ” eli kuinka suurella todennäköisyydellä

- naisten suhteellinen osuus poimit**tavassa** otoksessa **tulee poikkeamaan** naisten todellisesta 45 %:n osuudesta perusjoukossa

korkeintaan 2 %-yksikköä eli on välillä [0.43, 0,47]?

- Siis kuinka todennäköistä on, että naisten lukumäärä 1000 suuruisessa otoksessa on välillä [430, 470]?

- Tässä voidaan periaatteessa laskea samalla tavalla kuin lottoesimerkissä:

Sellaiset kombinaatiot, joissa on täsmälleen 430 naista (ja 570 miestä), ”muodostuvat” kuten edellä:

perusjoukon 45000 naisesta		55000: sta miehestä
↓		↓
sattuma voi valita otokseen 430 naista	**ja**	sattuma voi valita 570 miestä
$\binom{45000}{430}$ tavalla		$\binom{55000}{570}$ tavalla

jne.

Tässä kuitenkin kombinaatioiden (ja permutaatioiden) määrät ovat suunnattoman suuria ja niiden käsittelyyn ja niistä seuraavien todennäköisyyksien laskemiseen tarvitaan toisenlaista tekniikkaa.

Tällainen varsin helppokäyttöinen menetelmä on binomi- ja hypergeometrisenjakauman normaaliapproksimaatio, joita tarkastellaan vähän myöhemmin.

Todennäköisyyden käsitettä tarkastellaan tässä empiirisestä näkökulmasta lähtien ja käsitellään klassista ja frekventististä todennäköisyyttä.

- Lähinnä rajoitutaan tarkastelemaan klassista todennäköisyyttä ja sen ominaisuuksia, jotka näkyvät melko helposti esimerkkien kautta.

- Myös kaikilla muilla järkevillä tavoilla määritellyillä todennäköisyyskäsitteillä ovat nämä ominaisuudet.

- Lisäksi tämä todennäköisyyskäsitteen empiirinen rakenne on taustalla ”esikuvana” aksiomaattiselle (puhtaasti matemaattiselle) todennäköisyyslaskennalle, johon tässä ei puututa.

1.3. Klassinen ja frekventistinen todennäköisyyskäsite

Klassinen todennäköisyys ja sen ominaisuuksia

Arpajaisesimerkissä edellä oli yhteensä 200 arpaa, joista 30 kpl oli voittavia.

Ennen arvan poimimista (siis ennen empiiristä kokemusta tällaisista arpajaisista) pääteltiin, että

voiton todennäköisyys on $\frac{30}{200}$ = 0.15 eli ”on 15-prosenttinen mahdollisuus, että tulee voitto”.

Tämä päättely perustuu siihen, että

satunnaiskokeessa ”Nostetaan umpimähkään arpa.” tiedetään

- kaikkien arpojen lukumäärä,

- voittavien arpojen lukumäärä ja

- ”voidaan järkevästi olettaa”, että mikään arvoista ei osu muita ”helpommin” käteen.

Tällaisessa tilanteessa sovelletaan **klassista todennäköisyyskäsitettä**:

Oletetaan, että satunnaiskokeessa $\mathcal{E}$ on

- ***äärellinen määrä*** alkeistapauksia ja

- tutkittavan tilanteen empiirisen rakenteen perusteella voidaan järkevästi olettaa, että jokainen alkeistapaus voi ”yhtä hyvin” olla tuloksena satunnaiskokeessa $\mathcal{E}$, jolloin alkeistapauksia sanotaan ***symmetrisiksi***.

Silloin satunnaiskokeeseen liittyvän tapahtuman A ***todennäköisyys*** P(A) on suhdeluku

P(A) = $\boldsymbol{\frac{k}{n}}$ **,** missä

k = A:n ”esiintymiselle suotuisten” alkeistapausten määrä ja
n = alkeistapausten kokonaismäärä.

<u>Huom.</u>

– Matemaattisesti tämä määritelmä on (täysin kelvoton) kehämääritelmä. Symmetriaoletus jo sisältää oletuksen todennäköisyydestä.

- Tässä ei kuitenkaan edes pyritä ensisijaisesti puhtaasti teoreettisen matemaattisen käsitteen määrittelemiseen.

- Klassisen todennäköisyyden avulla havainnoidaan empiiristä todellisuutta.

- Tavoitteena on, että empiiristen satunnaisilmiöiden mahdollisten eri realisaatioiden ”esiintymismahdollisuuden suuruutta” voidaan mitata empiirisesti järkevällä tavalla. Kokemus osoittaa, että (usein) tällainen ennakointi toimii hyvin ja voi ohjata maksimoimaan hyötyjä ja minimoimaan haittoja.

- Ristiriita puhtaan matematiikan kanssa voidaan kyllä hetkeksi ”lakaista maton alle” symmetrisen todennäköisyyskentän käsitteen avulla. Silloin oletetaan aksiomaattisesti alkeistapausten yhtä suuri todennäköisyys (välittämättä siitä, onko tällä mitään vastinetta empiirisessä todellisuudessa).

Sama ongelma on kuitenkin edessä heti, jos teoriaa halutaan soveltaa käytäntöön.

- Symmetriaoletuksen voimassaoloon voidaan ja myös halutaan vaikuttaa.

Erityisesti **otantatilanneessa** huolehditaan, että

tilastoyksiköt valitaan perusjoukosta otokseen ”rehellisesti arpomalla”.

Esim. (jatkoa) Aiotaan ostaa arpa.

- Arpalaatikossa 1. on 100 arpaa, joista 15 on voittavia,

- arpalaatikossa 2. on 200 arpaa, joista 30 voittaa, ja

- arpalaatikossa 3. on 467 arpaa, joista 60 on voittavia.

Mistä laatikosta arpa kannattaa valita, (jos voitot ovat joka arpajaisissa yhtä hyvät)?

- Satunnaiskokeena $\mathcal{E}$ on arvan poimiminen laatikosta.

- **Alkeistapauksina** ovat laatikossa olevat arvat.

- Jos arpa tullaan valitsemaan täysin umpimähkään, mikä tahansa arvoista ”voi yhtä hyvin tulla valituksi” eli on järkevää olettaa alkeistapaukset **symmetrisiksi**.

Tapahtuman A = ”saadaan voitto” todennäköisyys on

1. arpajaisissa P(A) = $\frac{15}{100}$ = 0.15,

2. arpajaisissa P(A) = $\frac{30}{200}$ = 0.15 ja

3. arpajaisissa P(A) = $\frac{60}{467} \approx 0.13$.

- Vaikka yksittäisen arvan ostokerran tulosta ei voida varmasti ennustaa,

- edellinen lasku kuitenkin ”ennustaa”, että

1. ja 2. arpajaisissa tulee ”keskimäärin” 15 voittoa sataa yritystä kohti ja

3. vaihtoehdossa vain noin 13 voittoa 100:sta.

Esim. (jatkoa)

15 henkilön perusjoukosta, jossa on 9 harrastaa kuntosaliliikuntaa, aiotaan poimia

4:n suuruinen otos palauttamatta.

- Tässä siis **satunnaiskoe** $\mathcal{E}$ on otoksen poimiminen palauttamatta.

- **Alkeistapauksina** ovat kaikki mahdolliset eri otokset,

ja symmetrian aikaan saamiseksi ne voidaan **hahmottaa** yhtä hyvin

joko permutaatioina tai kombinaatioina.

- Kun otokset (**alkeistapaukset**) ajateltiin edellä neljän henkilön **jonoiksi**,

jonojen kokonaismääräksi saatiin kertolaskuperiaatteen

$$15 \cdot 14 \cdot 13 \cdot 12 = 32760.$$

Erityisesti tarkasteltiin tapahtumaa

A = ”otokseen osuu 2 harrastajaa ja 2 ei-harrastajaa”.

Vastaavalla tavalla saadaan esim. tapahtumalle

B = ”otokseen sattuu tulemaan 1 harrastaja (ja 3 ei-harrastajaa)”

kertolasku- ja yhteenlaskuperiaatteen avulla ”B:lle suotuisien”

jonojen lukumääräksi H E E E

↓ ↓ ↓ ↓

$\binom{4}{1} \cdot 9 \cdot 6 \cdot 5 \cdot 4 = 4320$

- Kun otokseen osuvat tullaan arpomaan rehellisesti **”lappuja hatusta”-menetelmällä**, sattuma ei ilmeisesti ”suosi” jotain tiettyjen henkilöiden juuri tietyssä järjestyksessä tulevaa jonoa muiden kustannuksella, joten **symmetriaoletus** on tässä selvästi (empiirisesti) **järkevä**.

- Silloin todennäköisyys, että otokseen tulee osumaan 1 harrastaja on (näiden jonojen suhteellinen osuus kaikista mahdollisista realisaatioista)

$P(B) = \frac{4320}{32760} \approx 0.132$

- Kun otokset (**alkeistapaukset**) hahmotetaan 4 henkilön **kombinaatioiksi**, saatiin edellä kokonaismääräksi

$$\binom{15}{4} = \frac{15!}{4!\cdot(15-4)!} = 1365 \text{ eri otosta.}$$

Sellaisia alkeistapauksia, joissa on 1 harrastaja, on

$$\binom{9}{1} \cdot \binom{6}{3} = \frac{9!}{1!\cdot(9-1)!} \cdot \frac{6!}{3!\cdot(6-1)!} = 9 \cdot 20 = 180 \text{ kpl.}$$

- Taas voidaan järkevästi olettaa, että kaikki nämä 1365 eri järjestämätöntä otosta voivat realisoitua yhtä hyvin eli myös näin hahmotetut alkeistapaukset ovat symmetrisiä, joten

$P(B) = \frac{180}{1365} \approx 0.132.$

- Tulos on sama, mikä saatiin ajattelemalla otokset permutaatioina. Näin täytyy tietenkin ollakin.

- Todennäköisyys on satunnaiskokeeseen liittyvän tapahtuman ”ominaisuus”, ja se ei saa riippua sen ”mittaamiseen” valitusta hahmotuksesta.

- Jos **satunnaiskoe** $\mathcal{E}$ on 4:n suuruisen otoksen poimiminen palauttaen tästä 15 henkilön perusjoukosta,

$\mathcal{E}$:n **alkeistapauksina** olevat otokset ovat jonoja.

- Edellä laskettiin, että

järjestettyjä otoksia on yhteensä 15·15·15·15 = 50625 kpl ja niistä

B:lle ”suotuisia” jonoja on $\binom{4}{1} \cdot 9 \cdot 6 \cdot 6 \cdot 6 = 7776$ kpl.

- Myös nämä kaikkia 10 000 alkeistapausta ovat symmetrisiä, joten

P(B) = $\frac{7776}{50625} \approx 0.154$.

Esim. (jatkoa) Lottorivien eli 7 pallon kombinaatioiden (tai permutaatioiden) arpominen on aivan vastaavanlainen tilanne kuin edellinen otanta. Myös siinä todennäköisyydet voidaan laskea klassisen todennäköisyyden avulla, kuten edellä jo tehtiin.

Todennäköisyyslaskennan sääntöjä esitettäessä voidaan käyttää joukko-opin merkintöjä, mutta niillä on erityinen empiirisempi sisältö:

Joukko-opissa	**Todennäköisyyslaskennassa**
Perusjoukko E - kaikki mielenkiinnon kohteena olevat alkiot	**Perusjoukko (otosavaruus) Ω** - kaikki satunnaiskokeen $\mathcal{E}$ tulosvaihtoehdot eli alkeistapaukset
Joukko A - osa E:n alkioista **Tyhjä joukko ϕ** - joukko, jossa ei yhtään alkiota	**Tapahtuma A** - Satunnaiskokeen $\mathcal{E}$ tuloksena on alkeistapaus, jossa A tapahtuu. **Mahdoton tapahtuma ϕ** - Mikään alkeistapaus ei johda A:n tapahtumiseen.
Komplementti A^c - E:n alkiot, jotka eivät kuulu A:han	**Komplementtitapahtuma A^c** - A ei tapahdu. A:n vastakohta tapahtuu.
Yhdiste A∪B - alkiot, jotka kuuluvat joukkoon A tai B (tai molempiin)	**Rinnakkaiset tapahtumat A∪B** - A **tai** B tai molemmat tapahtuvat
Leikkaus A∩B - A:n ja B:n yhteiset alkiot **A∩B = ϕ** - A:lla ja B:llä ei yhteisiä alkioita	**A∩B** - A **ja** B tapahtuvat. **A∩B = ϕ** - A ja B eivät voi tapahtua yhtä aikaa eli ovat toisensa poissulkevia

Huom. Käsite ”perusjoukko” on tilastotieteen ja todennäköisyyslaskennan tarkasteluissa mukana monessa merkityksessä:

- **Tilastotieteen käsitteenä** empiirinen **perusjoukko E** tarkoittaa kaikkien

(periaatteessa) mielenkiinnon kohteena olevia tilastoyksiköiden joukkoa.

Esim. markkinatutkimuksessa jonkin alueen kotitaloudet voivat olla tilastoyksiköitä, joista koostuu tutkittava perusjoukko E.

- Näin määritelty E on perusjoukko myös **joukko-opin käsitteenä**.
Alkioina ovat kaikki tilastoyksiköt.

- Kun tähän liitetään **satunnaiskoe**

$\mathcal{E}$ = ”Perusjoukosta (esim. kotitalouksista) **arvotaan yksi** tilastoyksikkö”,

niin sama E (kotitaloudet) on myös **todennäköisyyslaskennan käsitteenä** satunnaiskokeen $\mathcal{E}$ perusjoukko Ω.

Tämä samastus on tärkeä yhteys todennäköisyyslaskennan käsitteiden ja empiiriseen perusjoukkoon liittyvien käsitteiden välillä, josta lisää myöhemmin.

Kun tutkimuksessa ei kuitenkaan tyydytä vain yhden tilastoyksikön (kotitalouden) tutkimiseen, empiirisestä perusjoukosta E **aiotaan poimia otos** (monta tilastoyksikköä).

Silloin satunnaiskokeena on

$\mathcal{E}$ = ”E:stä poimitaan (esim. palauttamatta n = 1000 suuruinen) otos”.

Nyt $\mathcal{E}$:n **alkeistapauksia** ovat kaikki mahdolliset n:n suuruiset otokset,

jotka voidaan **hahmottaa**

joko järjestämättöminä tai järjestettyinä.

Kun **satunnaiskokeena $\mathcal{E}$ on otoksen poimiminen, perusjoukko** (**otosavaruus**) Ω on siis tarkastelutavan valinnasta riippuen kaikkien mahdollisten empiirisestä perusjoukosta E:stä saatavien (esim. 1000 kotitalouden)

permutaatioiden tai **kombinaatioiden** joukko.

Otannassa palauttaen otokset voidaan hahmottaa vain järjestettyinä.

Kun otos poimitaan ”rehellisesti arpomalla” (todennäköisyysotantaa käyttäen), näin määritelty

- perusjoukko Ω on kelvollinen alusta sattuman käyttäytymisen tutkimiseen:

- Mitkä ovat säännöt, joiden avulla voidaan kuvata, kuinka **Sattuma** määrää otoksen ”sisällön”,

- kun empiirisestä perusjoukosta E hankitaan havaintoja otannan avulla.

Klassisen todennäköisyyden ominaisuuksia

Esim. Taulukossa on 2-ulotteinen frekvenssijakauma, jossa on luokiteltu 52 henkilöä sukupuolen ja verotettavien tulojen (1000 €/v) mukaan:

	Naiset (N)	Miehet (M)	Yhteensä
0 - 19 (K)	4	12	16
20 – 39 (V)	1	3	4
40 - 59 (S)	3	9	12
≥ 60 (H)	5	15	20
Yhteensä	13	39	52

(Mitä samanlaista tässä on todennäköisyyslaskennan sääntöjen havainnollistamiseen usein käytetyn korttipakan kanssa?)

Satunnaiskokeena $\mathcal{E}$ on ”Arvotaan umpimähkään yksi henkilö.”

Kaikki tähän satunnaiskokeeseen liittyvät todennäköisyydet voidaan laskea klassisen todennäköisyyden mukaan:

- Alkeistapausten kokonaismäärä n = 52.

- Jos henkilö arvotaan rehellisesti ”lappuja hatusta -menetelmällä”, symmetria-oletus on järkevä.

- Suotuisten alkeistapausten määrä k on välillä $0 \leq k \leq 52$, olipa A mikä tahansa tapahtuma. Silloin klassisen todennäköisyyden määritelmä kalibroi todennäköisyyden suuruuden välille [0, 1], ja

aina on $\mathbf{0 \leq P(A) \leq 1}$.

Esim. todennäköisyys, että arvottava henkilö tulee olemaan nainen, on

$P(N) = \frac{13}{52} = 0.25$.

Todennäköisyys, että arvottava henkilö tulee olemaan mies, voidaan tietenkin laskea samalla tavalla

P(M) = $\frac{39}{52}$ = 0.75.

Samaan tulokseen päästään myös laskemalla

Henkilöitä yht. ↘ ↙ naisia

$$\mathbf{P(M)} = \frac{52-13}{52} = \frac{52}{52} - \frac{13}{52} = 1 - P(N) = \mathbf{1 - P(M^c)}$$

Todennäköisyys, että arvottavan henkilön tulo sattuu olemaan

0 – 19 (K) **tai** vähintään 60 (1000€) (H) on

pienituloisimmat suurituloisimmat (Katso taulukkoa.)

↘ ↙

$$\mathbf{P(K \cup H)} = \frac{16+20}{52} = \frac{16}{52} + \frac{20}{52} = \frac{36}{52} = \mathbf{P(K) + P(H)}.$$

Tässä tapahtumilla K ja H ei ole yhteisiä suotuisia alkeistapauksia.
Henkilön tulot eivät voi olla yhtä aikaa pienet ja suuret.

Sen sijaan

	Naiset (N)	Miehet (M)	Yhteensä
0 - 19 (K)	**4**	12	16
20 – 39 (V)	**1**	3	4
40 - 59 (S)	**3**	9	12
≥ 60 (H)	**5**	**15**	**20**
Yhteensä	**13**	39	**52**

tapahtumilla

N = ”henkilö on nainen” ja H = ”tulo vähintään 60 000 €”

on 5 mahdollista yhteistä realisaatiota.

Tapahtuman N ∪ H = ”valittava henkilö on nainen **tai** hyvätuloinen”

toteutumiselle suotuisien alkeistapauksien määrä ja todennäköisyys on

naisia hyvät tulot 5 hyvätuloista naista kahteen kertaan

↘ ↓ ↙

$$\mathbf{P(N\cup H)} = \frac{13+20-5}{52} = \frac{13}{52} + \frac{20}{52} - \frac{5}{52} \left(= \frac{28}{52}\right) = \mathbf{P(N) + P(H) - P(N\cap H)}$$

Siis satunnaiskokeeseen $\mathcal{E}$ liittyvillä tapahtumien klassisella todennäköisyydellä on selvästikin ominaisuudet:

Kaikilla perusjoukon tapahtumien A todennäköisyydellä on rajat

$$\mathbf{0 \leq P(A) \leq 1}$$

ja erityisesti:

P(A) = 0, jos A on mahdoton tapahtuma eli perusjoukossa ei ole yhtään A:lle suotuisaa alkeistapausta.

P(A) = 1, jos A on varma tapahtuma eli kaikki alkeistapaukset ovat A:lle suotuisia.

Yhteenlaskusääntö:

$$\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

ja erikoistapauksena

P(A∪B) = P(A) + P(B), kun $A \cap B = \phi$ eli A ja B ovat toisensa poissulkevia.

Todennäköisyyslaskennan yhteenlaskusääntö on analoginen kombinatoriikan yhteenlaskuperiaatteen kanssa.

Komplementtisääntö:

$$P(A) = 1 - P(A^c)$$

Komplementtisäännön avulla voidaan joskus lyhentää laskua:

Esim. (jatkoa)

	Naiset (N)	Miehet (M)	Yhteensä
0 - 19 (K)	4	12	16
20 – 39 (V)	1	3	4
40 - 59 (S)	3	9	12
$\geq$ 60 (H)	5	15	20
Yhteensä	13	39	52

Aiotaan valita umpimähkään 5 henkilöä (viiden suuruinen otos) palauttamatta.

Kuinka todennäköistä on, että valituksi tulee **vähintään** yksi nainen?

Satunnaiskokeena on

$\mathcal{E}$ = ”Arvotaan umpimähkään 5 henkilöä (palauttamatta).”

Alkeistapauksiksi voidaan hahmottaa

kaikki mahdolliset 5 henkilön muodostamat **kombinaatiot** (palauttamatta poimittavat otokset),

ja on järkevää olettaa ne **symmetrisiksi**.

Tapahtumalla A = ” Valituksi tulee vähintään 1 nainen.” on

5 toisensa poissulkevaa ”ala-vaihtoehtoa”

”Naisia on 1 tai 2 tai 3 tai 4 tai 5.”,

joiden todennäköisyyksien summana saadaan P(A).

A^c = ”Naisia ei ole yhtään.”

= ”Kaikki 5 valittavaa sattuvat olemaan miehiä.”,

ja komplementtisäännön avulla saadaan helpommin

$P(A) = 1 - P(A^c) = 1 - \frac{\binom{39}{5}}{\binom{52}{5}} = 1 - \frac{575\,757}{2\,598\,960} \approx 0.778.$

Kertolaskusääntö

Kun todennäköisyydet voidaan laskea klassisen todennäköisyyden mukaan, myös peräkkäisten tai muuten toisiinsa kytkettyjen tapahtumien

”tapahtuu A **ja** B”

realisoitumisen todennäköisyyden laskeminen käy hyvin samalla tavalla kuin kombinatoriikan kertolaskuperiaatteen soveltaminen:

Esim. (jatkoa) Edellä tarkastelluista 52 henkilöstä aiotaan valita 2 henkilöä palauttamatta (2 suuruinen ”otos”).

Kuinka todennäköistä on, että valituksi tulee 2 naista?

Jos A = ”Ensimmäisenä valittava henkilö on nainen.”

ja B = ”jälkimmäinen valittava henkilö on nainen.”,

niin laskettavana on

P(Valituksi tulee 2 naista.) = P(1. nainen ja 2. nainen) = $P(A \cap B)$.

Satunnaiskokeena on

$\mathcal{E}$ = ”Arvotaan umpimähkään 2 henkilöä (palauttamatta).”

”Otannassa” palauttamatta alkeistapauksiksi voidaan hahmottaa

sekä kombinaatiot tai permutaatiot.

Jotta todennäköisyyslaskennan kertolaskusäännön ajatus saadaan näkyviin, ”otokset” hahmotetaan tässä järjestettyinä:

Kombinatoriikan kertolaskuperiaatteen mukaan, kun valitaan

1. henkilö ja 2. henkilö jäljelle jäävistä,

↘ ↓ ↙

jonoja on yhteensä 52 · 51 = 2652 kpl.

Samalla tavalla päätellään, että sellaisia kahden henkilön jonoja,

joissa on 1. nainen ja 2. nainen, on 13 · 12 = 156 kpl.

Symmetria-oletus on selvästi järkevä, joten

$P(A \cap B) = \frac{156}{2652} \approx 0.059 \approx 6\ \%$.

Tapahtuman A∩B mahdollisen toteutumisen todennäköisyyttä voidaan myös seurata ”vaiheittain”:

tn, että 1. henkilö on nainen **P(A)**

tn, että 2. henkilö on nainen (B) **ottaen huomioon ehto A**, että myös 1. oli nainen, josta käytetään merkintää **P(B | A)**

↘ ↙

$$\mathbf{P(A\cap B)} = \frac{13\cdot 12}{52\cdot 51} = \frac{13}{52}\cdot\frac{12}{51} = \mathbf{P(A)\cdot P(B\,|\,A)}$$

Todennäköisyyttä P(B | A) sanotaan

tapahtuman B ehdolliseksi todennäköisyydeksi ehdolla A.

Samat tulokset saadaan esille myös toisin päin:

Ehdollinen todennäköisyys

Tapahtumien todennäköisyyksiä laskettaessa on usein käytettävissä **lisäinformaatiota**, joka vaikuttaa tulokseen.

Esim. (jatkoa)

	Naiset (N)	Miehet (M)	Yhteensä
0 - 19 (K)	4	12	16
20 – 39 (V)	1	3	4
40 - 59 (S)	3	9	12
≥ 60 (H)	5	15	20
Yhteensä	13	39	52

Satunnaiskokeena on

$\mathcal{E}$ = ”Arvotaan umpimähkään 1 henkilö.”,

jota vastaavassa perusjoukossa Ω symmetrisiä alkeistapauksia ovat kaikki 52 henkilöä.

Tapahtumien

H = ”Valittavan henkilön tulo sattuu olemaan vähintään 60 000 €.”,

B = ”Valittavan henkilön tulo on korkeintaan 39 000 €.” (= K $\cup$ V) ja

N = ”Valittava henkilö on nainen.”

todennäköisyydet ovat

P(H) = $\frac{\mathbf{20}}{\mathbf{52}} \approx 0.385$ ja myös

P(B) = $\frac{20}{52} \approx 0.385$ ja

P(N) = $\frac{13}{52} = 0.25$

Seuraavaksi aineistosta jätetään pois kaikki, joiden tulo on alle 20 000 €. Silloin jäljelle jää jakauma

	Nainen (N)	Mies (M)	Yhteensä
20 – 39 (V)	1	3	4
40 - 59 (S)	3	9	12
≥ 60 (H)	5	15	20
Yhteensä	9	27	36

Tässä muutetussa tilanteessa tiedetään, että aina varmasti tapahtuu

A = ”Valittavan henkilön tulo on vähintään 20 000 €.”

Tämä vaikuttaa (mahdollisesti)

tapahtumien H, B ja N esiintymisen todennäköisyyksiin:

Esim., kun lasketaan todennäköisyys tapahtumalle

H = ”Valituksi tulee henkilö, jonka tulo on vähintään 60 000 €.”,

tiedetään, että samalla varmasti tapahtuu

A = ”Valittavan henkilön tulo on vähintään 20 000 €.”

Silloin

- sanotaan, että lasketaan **tapahtuman H todennäköisyys ehdolla A**, ja

- merkintänä käytetään **P(H | A)**.

Tässä tapahtumien H, B ja N ehdolliset todennäköisyydet voidaan laskea klassisina todennäköisyyksinä ”typistämällä” perusjoukkoa ehdon A mukaan:

- Jäljelle jää alkuperäisestä perusjoukosta osa Ω ∩ A, jossa on

36 ehtona A olevalle tapahtumalle symmetristä alkeistapausta.

- Kun nyt lasketaan

↙ **Tapahtumalle H ja ehdolle A suotuisat alkeistapaukset**

$$\mathbf{P(H \mid A) = \frac{20}{36}} \approx 0.556 > P(H) \approx 0.385.$$

↖ ehdolle A suotuisat alkeistapaukset

Siis tieto pienituloisimpien pois jättämisestä **suurentaa** (tietenkin) vähintään 60 000 € ansaitsevan valituksi tulemisen mahdollisuuden suuruuden arviota.

Taas

↙ Tapahtumalle A ja B suotuisat alkeistapaukset

$$P(B \mid A) = \frac{4}{36} \approx 0.111 < P(B) \approx 0.385$$

ja

↙ Tapahtumalle $A \cap N$ suotuisat alkeistapaukset

$$P(N \mid A) = \frac{9}{36} = 0.25 = P(N).$$

Ehdon A voimassaolo ei muuta N:n tapahtumisen todennäköisyyttä.

Silloin sanotaan, että (tässä satunnaiskokeessa) tapahtuma N on **riippumaton** tapahtumasta A.

Ehdollisesta todennäköisyydestä saadaan toinen näkökulma, kun edellisissä laskuissa lavennetaan luvulla $\frac{1}{52}$.

Silloin tarkastelu ”palautetaan” typistetystä perusjoukosta $\Omega \cap$ A alkuperäiseen perusjoukkoon Ω:

Esim. todennäköisyys sille, että valituksi tulee henkilö, jonka tulo on **korkeintaan 39 000 € (B)**, kun tulo on varmasti **vähintään 20 000 € (A)** on

$$P(B \mid A) = \frac{4}{36}$$

↙ Tapahtumalle $A \cap B$ suotuisat alkeistapaukset

↖ Ehdolle A suotuisat alkeistapaukset

↙ Tapahtuman A ∩ B todennäköisyys **alkuper**äi**sess**ä perusjoukossa Ω

$$= \frac{\frac{1}{52} \cdot 4}{\frac{1}{52} \cdot 36} = \frac{\boxed{\frac{4}{52}}}{\boxed{\frac{36}{52}}}$$

↖ Ehdon A todennäköisyys **alkuper**äi**sess**ä perusjoukossa

$$= \frac{P(A \cap B)}{P(A)}$$

Tämän mukaisesti **määritellään**:

Jos A ja B ovat satunnaiskokeeseen $\mathcal{E}$ liittyviä tapahtumia ja P(A)>0,

tapahtuman B *ehdollinen todennäköisyys* ehdolla A on

↙ Kuinka mahdollinen B on "A: n puitteissa"

$$\boldsymbol{P(B \mid A) = \frac{P(A \cap B)}{P(A)}} .$$

↖ suhteessa siihen kuinka mahdollinen A ylipäänsä on.

Huom.

– Määritelmässä ei vaadita, että todennäköisyydet lasketaan klassisen todennäköisyyden mukaan, vaan määritelmä on täysin yleinen.

On samantekevää, millä ”järkevällä” tavalla P(A) ja P(A∩B) lasketaan alkuperäisessä perusjoukossa Ω.

- Jos klassinen todennäköisyys soveltuu tilanteeseen, todennäköisyydet voidaan laskea kuten edellä typistämällä perusjoukko Ω ehdolle mahdollisten alkeistapausten joukoksi Ω ∩ A.

- Jos perusjoukko ei ole äärellinen, ei ”typistäminen” toimi ja määritelmää tarvitaan tuottamaan ehdolliset todennäköisyydet.

- Tässä hieman ulkokohtaiseksi jäävä ehdollisen todennäköisyyden määritelmä on tärkeä askel empiiristen riippuvuuksien mallintamisen suuntaan.

”Jos tapahtuu A, mitä sen perusteella voi ennakoida B:stä?”

↓

”Jos selittävä muuttuja saa arvon x, kuinka suuri arvo on keskimäärin odotettavissa selitettävälle muuttujalle y?

Todennäköisyyslaskennan kertolaskusääntö

saatiin edellä kombinatoriikan kertolaskusäännön mallin mukaisesti tilanteessa, jossa klassinen todennäköisyys soveltui.

Ehdollisen todennäköisyyden määritelmästä

$$\mathbf{P(B \mid A)} = \frac{\mathbf{P(A \cap B)}}{\mathbf{P(A)}} \quad \Big| \cdot P(A)$$

seuraa sama yleisesti, kun kerrotaan ehdon todennäköisyydellä P(A):

$$\boldsymbol{P(A \cap B) = P(A) \cdot P(B \mid A)}$$

Kertolaskusäännön yleistys:

Useammalle kuin kahdelle tapahtumalle todennäköisyydet lasketaan ”ketjussa” kuten edellä:

Esim. kolmelle peräkkäiselle tapahtumalle

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2)$$

↑ ↑

Tapahtumien todennäköisyydet lasketaan peräkkäin ottaen aina huomioon, mitä edelliset tapahtumat vaikuttavat tilanteeseen.

Esim. (jatkoa)

	Naiset (N)	Miehet (M)	Yhteensä
0 - 19 (K)	4	12	16
20 – 39 (V)	1	3	4
40 - 59 (S)	3	9	12
≥ 60 (H)	5	15	20
Yhteensä	13	39	52

Aiotaan valita umpimähkään 5 henkilöä (viiden suuruinen otos) palauttamatta.

A = ”Valituksi tulee vähintään 1 nainen.”

A^c = ”Naisia ei ole yhtään.”

= ”Kaikki 5 valittavaa sattuvat olemaan miehiä.”

= $M_1 \cap M_2 \cap M_3 \cap M_4 \cap M_5$

↑ 1. valittava henkilö on mies, jne.

Todennäköisyys saadaan tässäkin helpommin, kun ensin käytetään komplementtisääntöä ja vasta sitten kertolaskusääntöä:

$P(A) = 1 - P(A^c) = 1 - P(M_1 \cap M_2 \cap M_3 \cap M_4 \cap M_5)$

$$= 1 - \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50} \cdot \frac{36}{49} \cdot \frac{35}{48}$$

$$\approx 0.778,$$

kuten edellä saatiin kombinaatioiden avulla.

Tilastollinen (stokastinen) riippumattomuus

määritellään kertolaskusäännön avulla:

Satunnaiskokeeseen $\mathcal{E}$ liittyvät **tapahtumat A ja B ovat riippumattomia**,

jos

A: n tapahtumisesta ei saada B: hen vaikuttavaa informaatiota. Ehtoa A "ei tarvita"

↙

$$\boldsymbol{P(A \cap B) = P(A) \cdot P(B)}\ .$$

- Tapahtumien A ja B riippumattomuudesta käytetään merkintää $A \perp B$.

- Suorissa sovelluksissa riippumattomuus on usein asetelman perusteella täysin ilmeinen.

Esim. (jatkoa edelliseen) Aiotaan valita 2 henkilöä palauttaen.

Kuinka todennäköistä on, että tuloksena on

”Molemmat ovat naisia” = A∩B,

kun

A = ”1. valittava tulee olemaan nainen” ja

B = ”2. valittava tulee olemaan nainen”.

”Otos” poimitaan palauttaen, joten ensimmäisen valinnan tulos ei vaikuta toiseen ja

$P(A \cap B) = P(A) \cdot P(B) = \frac{13}{52} \cdot \frac{13}{52} = 0.0625 \approx 6\ \%.$

↑ ↑

Kummallakin kerralla sattumalla on valittavana 13 naista 52:sta.

”Otannassa” palauttamatta (Ks. edellä) $P(A \cap B) \approx 0.059 \approx 6\ \%$

- melko saman suuruinen,
- mutta kuitenkin pienempi. (Miksi?)

Tässä (empiirinen) perusjoukko on hyvin pieni (N=52).

Jos tilanne olisi muuten sama kuin edellä, mutta

henkilöitä olisi 520 000 ja heistä naisia 130 000,

olisi ”otannassa”

palauttaen

$$P(A \cap B) = P(A) \cdot P(B) = \frac{130\,000}{520\,000} \cdot \frac{130\,000}{520\,000} = 0.25 \cdot 0.25 = 0.0625$$

ja

palauttamatta

$$P(A \cap B) = P(A) \cdot P(B \mid A) = \frac{130\,000}{520\,000} \cdot \frac{129\,999}{519\,999} = 0.25 \cdot 0.2499985 \approx 0.0625.$$

Siis otannassa palauttaen ja palauttamatta tulokset ovat jokseenkin samat, jos (empiirinen) perusjoukko on ”suuri” ja otoskoko on siihen verrattuna ”kohtuullisen” kokoinen. Tähän perehdytään tarkemmin myöhemmin.

Edellä riippumattomuus tai riippuvuus näkyi ”tilanteesta suoraan”.

Toisenlaista käyttöä riippumattomuuden määritelmällä on, kun selvitetään,

millainen ”tilanteen” pitäisi olla riippumattomuuden vallitessa.

Esim. Kuntosalilla aktiivisesti harjoittelevista poimittiin 500 suuruinen otos. Siinä on

200 naista (N) ja 300 miestä (M) ja

heistä (sallittuja) voimaharjoittelua tukevia lisäravinteita

käyttää (K) 150 ja ei käytä 350 (T).

Otoksesta saatiin ristiintaulukoimalla 2-ulotteinen jakauma:

o_{ij}	Käyttää (K)	Ei (T)	Yhteensä
Nainen (N)	48	152	200
Mies (M)	102	198	300
Yhteensä	150	350	500

(o_{ij} observed frequency)

Voidaanko tämän perusteella päätellä, riippuuko käyttäminen sukupuolesta?

Tällaisten kvalitatiivisten muuttujien riippuvuutta voidaan tutkia

(mm. myöhemmin tarkemmin käsiteltävän) χ^2- riippumattomuustestin avulla.

Siinä tutkitaan aluksi,

millaiselta tämän taulukon **pitäisi** (keskimäärin) **näyttää**, eli

- kuinka suuria tämän 2-ulotteisen frekvenssijakauman solufrekvenssien **pitäisi olla**,

- jos (muuttujat x =) ”lisäravinteiden käyttö” ja (y =) ”sukupuoli” **olisivat toisistaan riippumattomia?**

	Käyttää (K)	Ei (T)	Yhteensä
Nainen (N)	**?**	?	**200**
Mies (M)	?	?	300
Yhteensä	**150**	350	**500**

Jos (muuttujien x ja y) riippumattomuus olisi voimassa, niin satunnaiskokeessa

$\mathcal{E}$ = ”Näiden 500 henkilön joukosta arvotaan 1 henkilö”

$$P(N \cap K) = P(N) \cdot P(K) = \frac{200}{500} \cdot \frac{150}{500} = 0.4 \cdot 0.3 = 0.12.$$

Siis

- (riippumattomuuden vallitessa)

“12 prosentissa tapauksista on odotettavissa”,

että henkilö on nainen ja käyttää lisäravinteita.

- Silloin 500 henkilöstä on ”keskimäärin odotettavissa” N∩K

$500 \cdot 0.12 \left(= 500 \cdot \frac{200}{500} \cdot \frac{150}{500} = \frac{200 \cdot 150}{500}\right) = 60$ tapauksessa.

Samalla tavalla riippumattomuuden voimassa ollessa on

odotettu frekvenssi

- naisille, jotka eivät käytä (N∩T) reunafrekvenssit

↙ ↘

$500 \cdot P(N \cap T) = 500 \cdot P(N) \cdot P(T) = \cancel{500} \cdot \frac{\mathbf{200}}{\cancel{500}} \cdot \frac{\mathbf{350}}{\mathbf{500}} = 0.4 \cdot 350 = 140$

↖ otoskoko

- miehille, jotka käyttävät (M∩K)

$$\boxed{\frac{\mathbf{300}}{\mathbf{500}}} \cdot 15 = 90$$

↗ ↖

miesten suhteellinen osuus käyttäjien määrä

- miehille, jotka eivät käytä (M∩T) = $\frac{300 \cdot 350}{500}$ = 210.

(Nämä kolme viimeistä odotettua frekvenssiä saadaan helpomminkin reunafrekvensseistä vähentämällä.)

Jos ravintolisien käyttäminen **olisi riippumatonta** sukupuolesta, otoksen 500 henkilön pitäisi jakautua keskimäärin edellä lasketulla tavalla:

e_{ij}	Käyttää (K)	Ei (T)	Yhteensä
Nainen (N)	**60**	**140**	200
Mies (M)	**90**	**210**	300
Yhteensä	150	350	500

(e_{ij} expected frequency)

Seuraava askel riippuvuuden tutkimisessa on luonnollisesti, että otoksesta **havaittuja frekvenssejä o_{ij}** ja riippumattomuuden vallitessa **odotettuja frekvenssejä e_{ij}.**

”verrataan toisiinsa”.

o_{ij} (e_{ij})	Käyttää (K)	Ei (T)	Yhteensä
Nainen (N)	48 (60)	152 (140)	200
Mies (M)	102 (90)	198 (210)	300
Yhteensä	150	350	500

- Havaitut frekvenssit **poikkeavat 12 verran** siitä, mitä taulukossa ”pitäisi” keskimäärin näkyä riippumattomuuden vallitessa.

- Onko ero niin suuri, että näin ei voi kohtuullisella todennäköisyydellä enää tapahtua riippumattomuuden vallitessa,

- vaan otoksen perustella on järkevää päätellä, että käyttö riippuu sukupuolesta?

Esim. (jatkoa) Yrityksen Y työntekijöistä poimittiin otos, jonka avulla selvitettiin suhtautumista tulospalkkauksen käyttöön ottoon yrityksessä. Saatiin tulokset:

o_{ij}	Kielteinen	Neutraali	Myönteinen	Yhteensä
Alle 40-v.	96	174	159	429
Yli 40-v.	117	155	122	394
Yhteensä	213	329	281	823

- Voidaanko tämän perusteella päätellä, riippuuko suhtautuminen tulospalkkaukseen työntekijän iästä?

Kuten edellä saadaan riippumattomuuden vallitessa odotetut frekvenssit jakamalla reunafrekvenssien tulo otoskoolla:

$e_{11} = \frac{429 \cdot 213}{823} = 111.03,$ $e_{12} = \frac{429 \cdot 329}{823} = 171.50$ jne.

(Loput odotetuista frekvensseistä saadaan myös reunafrekvensseistä vähentämällä, kun nämä kaksi on laskettu.)

o_{ij} (e_{ij})	Kielteinen	Neutraali	Myönteinen	Yhteensä
Alle 40-v.	96 (111.03)	174 (171.50)	159 (146.48)	429
Yli 40-v.	117(101.97)	155 (157.50)	122 (134.52)	394
Yhteensä	213	329	281	823

Ero havaittujen ja riippumattomuuden vallitessa odotettujen frekvenssien välillä otoksessa viittaa siihen suuntaan, että nuoremmat olisivat jonkin verran myönteisempiä tulospalkkaukselle.

Tästä jatketaan myöhemmin hypoteesien testaamisen yhteydessä ja selvitetään, johtuuko ero vain sattumasta vai voidaanko tämän perusteella päätellä, riippuuko suhtautuminen iästä.

Frekventistinen todennäköisyyskäsite

- Lisäravinteita käsittelevässä esimerkissä edellä lasketut todennäköisyydet ajateltiin klassisina todennäköisyyksinä satunnaiskokeessa

$\mathcal{E}$ = ”Aineiston 500 henkilön joukosta arvotaan 1 henkilö”

- Toisaalta esimerkiksi P(K) = $\frac{150}{500}$ = 0.30 on otoksesta laskettu suhteellinen frekvenssi niille, jotka käyttävät lisäravinteita.

- Jos otos on poimittu ”arpoja hatusta”- periaatteella, tämä 500 henkilön otos on edustava osa maksullisia kuntosalipalveluja käyttävien perusjoukosta.

Silloin tämä arvo 0.30 on hyvä arvio (**estimaatti**) käyttäjien todelliselle suhteelliselle osuudelle perusjoukossa ja myös

hyvä approksimaatio todennäköisyydelle, että koko perusjoukosta umpimähkään valittava henkilö on lisäravinteiden käyttäjä.

Tämä koko perusjoukkoa koskeva todennäköisyyden suuruuden arvio on frekventistinen.

Frekventistisen todennäköisyyden

ajatuksen järkevyys perustuu empiirisessä todellisuudessa havaittavaan

tilastolliseen säännönmukaisuuteen:

Esim. Kolikon heitto on äärimmäisen pelkistetty tilanne, jossa tämä ilmiö

(fysikaalisen todellisuuden lainalaisuuksien vuoksi(?)) kerta toisensa jälkeen voidaan havaita.

Kun tällaisia kokeita todella on tehty,

pitkässä heittosarjassa tapahtuman A = ”Heiton tuloksena on kruunu.”

suhteellinen frekvenssi voi vaihdella aluksi paljon, mutta heittojen määrän kasvaessa se tasaantuu lähelle arvoa 0.5.

Oletetaan, että satunnaiskoe $\mathcal{E}$ toistuu (toistetaan) samoissa olosuhteissa n kertaa ja seurataan $\mathcal{E}$:hen liittyvän tapahtuman A suhteellisen frekvenssin

$P_n(A)$ suuruuden kehittymistä toistojen lukumäärän n kasvaessa.

Jos $P_n(A)$:n vaihtelu vähenee ja **$P_n(A)$ ”näyttää lähenevän” jotain kiinteää lukua** satunnaiskokeen $\mathcal{E}$ pitkässä toistosarjassa,

tätä (kuviteltua) lukua sanotaan **tapahtuman A todennäköisyydeksi** P(A).

Tässäkään ei pyritä määrittelemään ”puhtaan matematiikan” käsitettä, vaan ajatus lähtee empiirisen todellisuuden havainnoimisesta:

1) Edellä määritelty käsite ei ole matematiikan käsittelemä lukujonon raja-arvo.

Satunnaiskoe $\mathcal{E}$ voidaan toistaa vain äärellisen monta kertaa. Mikään ei periaatteessa takaa, ettei vaikkapa kolikkoa heitettäessä kruunujen

suhteellinen osuus ”villiintyisi” esim. 1000 000 000 heiton jälkeen. Empiirinen kokemus ja terve järki kuitenkin sanovat, näin ei ilmeisesti kuitenkaan kävisi.

Tämä riittää hyvin toimivan empiirisen todennäköisyyskäsitteen määrittelemiseen.

2) Klassinen todennäköisyyden laskeminen on ”asetelman rakenteen” perusteella tehtävää järkeilyä. Pohjimmiltaan ajatus on hyvin samanlainen kuin empiiriseen kokemukseen perustuvassa frekventistisessä todennäköisyydessä.

Taustalla on tulevaisuuteen suuntautuva oletus tilastollisesta säännönmukaisuudesta.

Kun rahaa heitettäessä päätellään klassisena todennäköisyytenä P(kruunu) = $\frac{1}{2}$, odotetaan, että kolikkoa monta kertaa heitettäessä ”keskimäärin” 50 % tuloksista tulee olemaan kruunuja.

3) Käytännössä tapahtuman A todennäköisyyden arvio P(A) on käytettävissä olevien toistojen realisaatioista laskettu suhteellinen frekvenssi $P_n(A)$.

Aiemmassa esimerkissä näkyi, että poikien suhteellinen osuus syntyvistä lapsista on vuodesta toiseen noin 51 %.

Tämä useasta kymmenestä tuhannesta ”lapsen syntymän toistosta” laskettu suhteellinen frekvenssi on approksimaatio pojan syntymän todennäköisyydelle.

Usein, kuten kolikon heitossa tai syntyvien poikien suhteellisen frekvenssin seuraamisessa ei tiedetä (ainakaan tarkkaan), mistä suhteellisen frekvenssin ”lähestyminen kohti kiinteää arvoa” johtuu.

4) Otanta (ja koesuunnittelu-) tilanteet asetetaan niin, että tämä ilmiö auttaa, kun tehdään arvioita perusjoukosta (**estimoinnissa**) otoksen informaation perusteella.

Jos esim. markkinatutkimuksessa on taustatietona, että tutkittavassa perusjoukossa on 45 % naisia,

- jo klassisen todennäköisyyden mukaan P(Valituksi tulee nainen.) = 0.45,

- mutta otosta poimittaessa (kokemuksen mukaan) näkyy, että naisten suhteellinen osuus $P_n(N)$ todella asettuu ”lähelle” tätä arvoa.

Sitä lähemmäksi, mitä suurempi otoskoko n on.

Tämän pohjalta tilanne voidaan kääntää toisin päin:

Ei tiedetä hyödykkeen H käyttäjien todellista suhteellista osuutta π perusjoukossa. Kuitenkin **”suurella varmuudella”** saadaan

”lähellä oikeaa” oleva arvio käyttäjien todelliselle osuudelle perusjoukossa

otoksesta havaitun suhteellisen frekvenssin $P_n(H)$ avulla.

Myöhemmin tilastollisessa päättelyssä tarkennetaan, miten voidaan esittää lukuina

arvion **luotettavuus** (Kuinka suurella varmuudella?)

ja **tarkkuus** (Kuinka lähellä oikeaa (parametrin) arvoa?)

suhteellisen osuuden lisäksi myös eräiden muiden otoksesta laskettavien tunnuslukujen (**estimaattoreiden)** osalta.

Subjektiivista todennäköisyyttä ei käsitellä tässä tarkemmin.

Olipa tarkastelutapa mikä tahansa, empiirisellä todennäköisyyskäsitteellä ovat voimassa samat säännöt, joita ovat mm.

Kokonaistodennäköisyyden kaava ja Bayesin kaava

Todennäköisyyslaskennan yhteenlasku- ja kertolaskusääntö ovat analogisia kombinatoriikan vastaavien sääntöjen kanssa:

- Kun tapahtumat ovat rinnakkaiset ”jotain **tai** jotain tapahtuu”, todennäköisyydet (vaihtoehtojen lukumäärät) **lasketaan yhteen**.

- Kun tapahtumat ovat ketjutettuja ”jotain **ja** jotain tapahtuu”, todennäköisyydet (vaihtoehtojen lukumäärät) **kerrotaan keskenään**.

”Rinnakkaisuuden” ja ”peräkkäisyyden” tutkiminen sujuu hyvin samalla tavalla todennäköisyyksiä laskettaessa, kuin edellä tehtiin kombinatoriikassa:

Esim. (jatkoa) 52 henkilöstä 13 on naisia. Aiotaan poimia palauttamatta

2 henkilöä. Kuinka todennäköistä on, että tapahtuu

A = ”molemmat ovat samaa sukupuolta?

Tähän johtavat vaihtoehdot

”1. nainen(N_1) **ja** 2. nainen”(N_2) **tai** ”1. mies(M_1) **ja** 2. mies”(M_2).

Todennäköisyyslaskennan yhteenlasku- ja kertolaskusäännön mukaan

P(A) = P(”1. nainen **ja** 2. nainen” **tai** ”1. mies **ja** 2. mies”)

↓

= P(”1. nainen **ja** 2. nainen”) **+** P(”1. mies **ja** 2. mies”)

↓ ↓

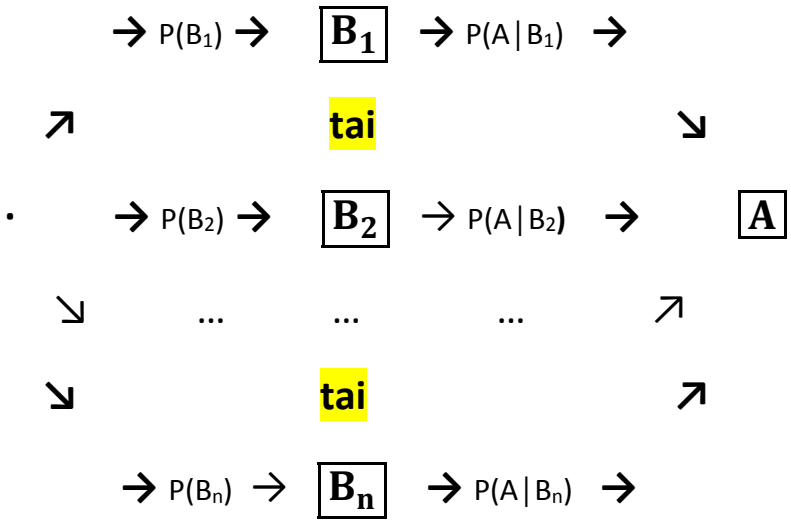
= $P(N_1) \cdot P(N_2 \mid N_1)$ + $P(M_1) \cdot P(M_2 \mid M_1)$

$$= \frac{13}{52} \cdot \frac{12}{51} + \frac{39}{52} \cdot \frac{38}{51} \approx 0.618$$

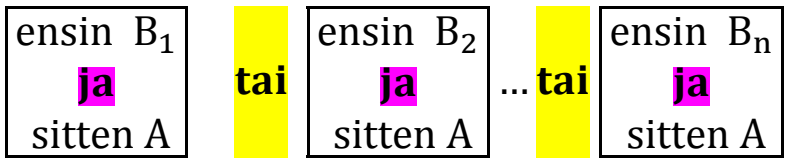
Sama tilanne yleisesti:

Tapahtumaa A ”edeltää” jokin tapahtumista B_1, B_2, …, B_n, jotka (mahdollisesti) vaikuttavat tapahtuman A todennäköisyyteen.

1. vaihe **ja** 2. vaihe



Yhteenlasku- ja kertolaskusäännön avulla saadaan



P(A) = $P((B_1 \cap A) \cup (B_2 \cap A) \cup \ldots \cup (B_n \cap A))$

= $P(B_1 \cap A) + P(B_2 \cap A) + \ldots + P(B_n \cap A)$ (yhteenlaskusääntö)

ja (kertolaskusääntö)

$$P(A) = P(B_1)\cdot P(A \mid B_1) + P(B_2)\cdot P(A \mid B_2) + \ldots + P(B_n)\cdot P(A \mid B_n),$$

jota sanotaan ***kokonaistodennäköisyyden kaavaksi***.

Esim. Tehdas tilaa komponentteja S alihankkijoilta K, L ja M.

Pieni osa tuotteista on viallisia. Taulukossa ovat tuottajien osuudet tilauksista ja viallisten tuotteiden osuudet heidän toimittamistaan tarvikkeista:

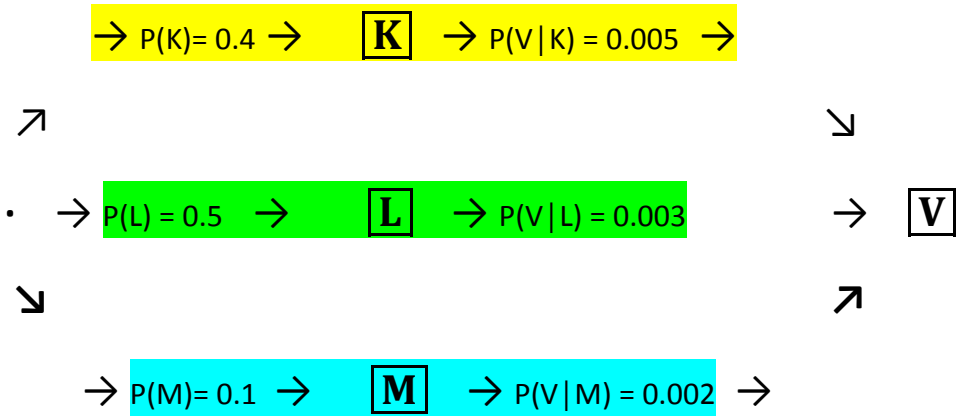
Tuottaja	K	L	M
Osuus tilauksista	40 %	50 %	10 %
Viallisten osuus	0.5 %	0.3 %	0.2 %

Kuinka todennäköistä on

V = ”umpimähkään valittava tuote on viallinen”?

Taulukossa olevat suhteelliset osuudet ovat suoraan laskussa tarvittavat todennäköisyydet:

Tuottaja	K	L	M
Osuus tilauksista	40 % P(K) = 0.4	50 % P(L) = 0.5	10 % P(M) = 0.1
Viallisten osuus	0.5 % P(V | K) = 0.005	0.3 % P(V | L) = 0.003	0.2 % P(V | M) = 0.002



$$P(V) = P(K)\cdot P(V \mid K) + P(L)\cdot P(V \mid L) + P(M)\cdot P(V \mid M)$$

$$= 0.4 \cdot 0.005 + 0.5 \cdot 0.003 + 0.1 \cdot 0.002 \cong 0.0037.$$

Jos komponentteja ostetaan (esim.) 100 000 kpl, niin

on odotettavissa, että keskimäärin käy seuraavalla tavalla:

↗	K:n tuottamia 100 000 · 0.4 = 40 000	Niistä viallisia 0.005 · 40 000 = 200
100 000 →	L:n tuottamia 100 000 · 0.5 = 50 000	Niistä viallisia 0.003 · 50 000 = 150
↘	M:n tuottamia 100 000 · 0.1 = 10 000	Niistä viallisia 0.002 · 10 000 = 20
	Yhteensä 100 000	Yhteensä 370

Siis keskimäärin odotettavissa oleva viallisten suhteellinen osuus on

$\frac{370}{100\,000} \cong 0.0037$ (= P(V)).

- Edellä arvioitiin ”ajassa eteenpäin”, kuinka todennäköistä on,

että valituksi tulee (joskus tulevaisuudessa ehkä mahdollisesti) viallinen komponentti.

- Joskus on hyödyllistä ja voidaan kääntää arvio ”ajassa taaksepäin”:

Oletetaan, että edellisessä tilanteessa

komponentti on todella valittu ja se on havaittu vialliseksi
(V on tapahtunut).

Silloin voidaan kysyä, kuinka todennäköistä on, että (esim.) alihankkija K on sen tuottanut eli

kuinka suuri on P(K|V)?

Edellisen taulukon avulla saadaan ilmeisesti tälle järkevä arvio:

100 000 komponentista on

↙ K:n tuottamia viallisia keskimäärin 200 kpl

$$P(K|V) = \frac{200}{370} \cong 0.54$$

↖ ja viallisia on yhteensä keskimäärin 370 kpl.

Kun palataan taulukossa tehtyyn laskuun, josta nämä viallisten määrät tulivat, saadaan

$$P(K \mid V) = \frac{200}{370} = \frac{40\,000 \cdot 0.005}{370} = \frac{\cancel{100\,000} \cdot 0.4 \cdot 0.005}{\cancel{100\,000} \cdot 0.00370}$$

$$= \frac{P(K) \cdot (V \mid K)}{P(V)}$$

Vastaava tilanne yleisesti:

- Jos satunnaiskokeen $\mathcal{E}$ tuloksena sattuu olemaan tapahtuma A,

niin sitä ”edeltää” jokin tapahtumista B_1, B_2, …, B_n.

1. vaiheessa siirrytään ”tilaan” B_i

(a priori-) todennäköisyydellä $\mathbf{P(B_i)}$, i = 1, 2, …, n.

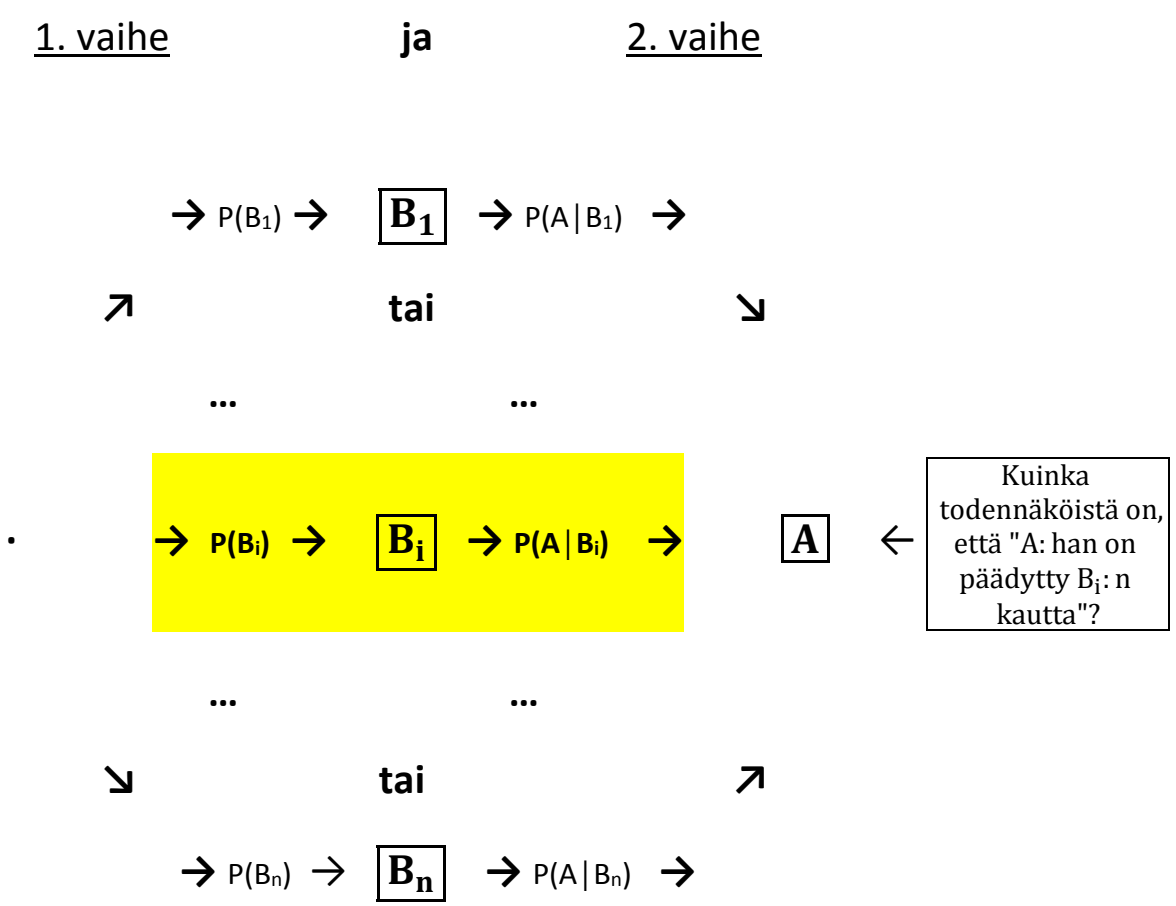
2. vaiheessa siirrytään ”tilasta” B_i ”tilaan” A

(a posteriori-) todennäköisyydellä $\mathbf{P(A \mid B_i)}$, i = 1, 2, …, n.

- Kun satunnaiskoe $\mathcal{E}$ tehdään, **tuloksena todella on tapahtuma A**.

Kuinka todennäköistä on, että sitä on **edeltänyt tapahtuma B_i**

eli kuinka suuri on $P(B_i|A)$?



Ehdollisen todennäköisyyden määritelmän mukaan on

$$P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i) \cdot P(A \mid B_i)}{P(A)}$$

ja

Tn., että päädytään A:han B_i:n kautta

↓

$$P(B_i \mid A) = \frac{P(B_i) \cdot P(A \mid B_i)}{P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + \cdots + P(B_n) \cdot P(A \mid B_n)}$$

↑

Tn., että ylipäänsä päädytään A:han

Tätä sääntöä sanotaan ***Bayesin kaavaksi***.

- **Kokonaistodennäköisyys**:

Arvioidaan (usein **ajassa eteenpäin**), kuinka todennäköistä on yhteensä eri ”reittien” B_1, B_2, …, B_n kautta havaita tapahtuma A.

- **Bayesin kaava**:

Arvioidaan (usein **ajassa taaksepäin**), kuinka todennäköistä on, että

on tultu nimenomaan vaihtoehdon B_i kautta, kun on todella havaittu

tapahtuma A.

Esim. Aiemman tutkimusaineiston perusteella arvioidaan, että

sairaudelle S altistava tekijä on 15 % väestöstä.

Alttiuden pikaseulontaa varten on kehitetty automaattinen menetelmä, ja selvitellään, onko se kelvollinen laboratorioille myytäväksi.

Havaittiin, että menetelmä

- hälyttää positiivisella tuloksella 90 prosentissa tapauksista, jos altistava tekijä todella on olemassa,

- mutta myös 2 prosentissa tapauksista, joissa tekijää ei olekaan.

Kuinka usein hälytys on väärä?

Siis testin tulos on positiivinen, mutta testattavalla ei olekaan altistusta?

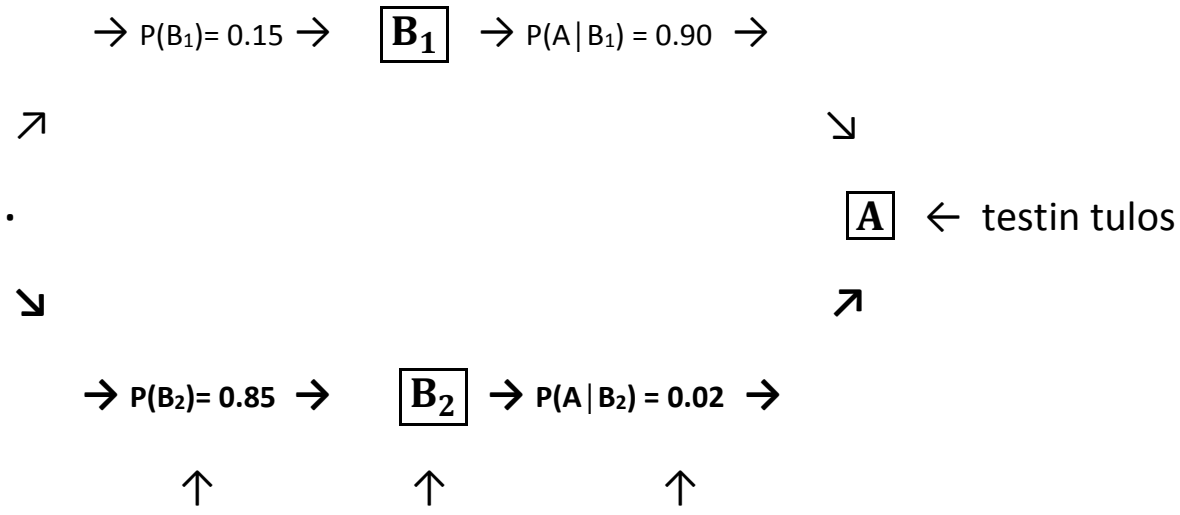
Edellisen mukaan merkinnöillä

B_1 = ”Henkilöllä on altistus.”, B_2 = ”Henkilöllä ei ole altistusta.” ja

A = ”Pikatestin tulos on positiivinen”.

Kuinka suuri on $P(B_2 \mid A)$?

Positiiviseen testitulokseen voidaan päätyä ”kahta reittiä pitkin”:



Kuinka todennäköistä on, että A:han päädyttiin ”tätä kautta”?

$$P(B_2 \mid A) = \frac{P(B_2) \cdot P(A \mid B_2)}{P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2)}$$

$$= \frac{0.85 \cdot 0.02}{0.85 \cdot 0.02 + 0.15 \cdot 0.90}$$

$$\cong 0.112$$

Siis noin 11 % positiivisista testituloksista on vääriä!

Jos myyntikelpoisuuden ehtona on, että vain 5 % positiivisista tuloksista saa olla vääriä, menetelmää ei uskalleta myydä.

Samalla tavalla kuin edellä voidaan laskea

kuinka todennäköistä on, että tuloksen ollessa negatiivinen kuitenkin ”takana” onkin altistus eli $P(B_1 \mid A^c)$.

Laskun tuloksesta (n.1.8 %) näkyy, että negatiiviset tulokset ovat niukasti riittävän usein oikeita, jos tässä rajana on 2 %.

Tässä ei jatketa pidemmälle Bayesilaisiin tilastollisiin menetelmiin, vaan ne jätetään myöhempiin tilastotieteen opintoihin.

2 Todennäköisyysjakaumista

2.1 Satunnaismuuttuja ja sen jakauma

Esim. (jatkoa)

15 henkilön perusjoukosta, jossa on 9 kuntosaliliikunnan harrastajaa, aiotaan poimia 4:n suuruinen **otos palauttamatta**.

Kun otokset hahmotetaan (esim.) kombinaatioina, on satunnaiskokeella

$\mathcal{E}$ = ”Poimitaan 4:n suuruinen otos palauttamatta.”,

$\binom{15}{4} = \frac{15!}{(15-4)!\cdot 4!}$ = 1365 (symmetristä) alkeistapausta.

Edellä laskettiin

P(”Otokseen tulee osumaan 2 harrastajaa.”) = $\frac{\binom{9}{2}\cdot\binom{6}{2}}{1365} = \frac{540}{1365} \approx 0.396$.

ja

P(”Otokseen tulee osumaan 1 harrastajaa.”) = $\frac{\binom{9}{1} \cdot \binom{6}{3}}{1365} = \frac{180}{1365} \approx 0.132.$

Kun käytetään merkintää

X = harrastajien lkm tällaisessa palauttamatta poimittavassa otoksessa,

edellisen voi esittää lyhyemmin

P(X = 2) = $\frac{540}{1365}$ ja P(X =1) = $\frac{180}{1365}$.

Näin samalla liitetään satunnaiskokeen $\mathcal{E}$ **alkeistapauksiin muuttujakäsite**.

- Sattuma määrää, mikä alkeistapaus tulee realisoitumaan.

- X kuvaa alkeistapauksen (tässä otoksen) **ominaisuutta**: Kuinka monta harrastajaa otoksessa on.

Tällaista muuttujaa, jonka arvon sattuma tulee määräämään satunnaiskokeessa, sanotaan **satunnaismuuttujaksi**.

Samalla tavalla lasketaan, kuinka todennäköistä on, että poimittavalla otoksella ($\mathcal{E}$:n alkeistapauksella)

tulee olemaan ominaisuus: Harrastajien määrä on 0, 1 tai 4 kpl.

$$P(X = 0) = \frac{\binom{9}{0} \cdot \binom{6}{4}}{\binom{15}{4}} = \frac{15}{1365} \approx 0.011$$

$$P(X = 3) = \frac{\binom{9}{3} \cdot \binom{6}{1}}{\binom{15}{4}} = \frac{504}{1365} \approx 0.369$$

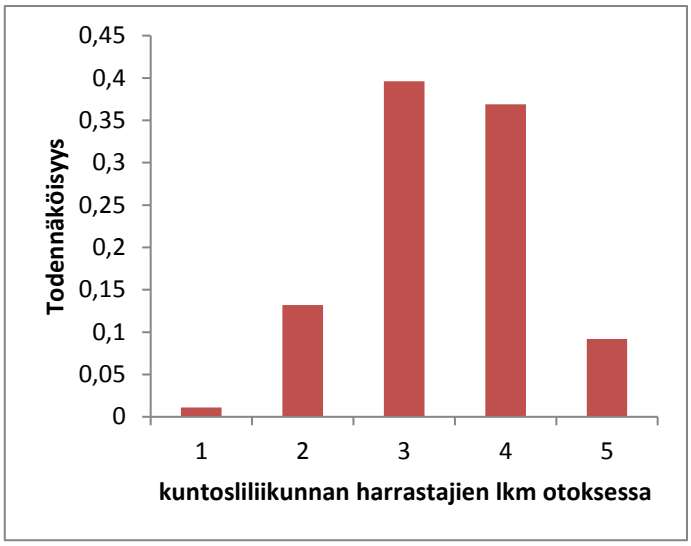
$$P(X = 4) = \frac{\binom{9}{4} \cdot \binom{6}{0}}{\binom{15}{4}} = \frac{126}{1365} \approx 0.092$$

X:n arvot ja niiden todennäköisyydet määrittelevät

satunnaismuuttujan X todennäköisyysjakauman

X:n arvo x_i	0	1	2	3	4	Σ
$p_i=P(X=x_i)$	15/1365 $\approx$ 0.011	180/1365 $\approx$ 0.132	540/1365 $\approx$ 0.396	504/1365 $\approx$ 0.369	126/1365 $\approx$ 0.092	1

Kuviona:



Todennäköisyysjakaumaa käsitellään hyvin samalla tavalla kuin diskreetin (epäjatkuvan) empiirisen muuttujan frekvenssijakaumaa. Näkökulma on kuitenkin erilainen.

- Kun käsitellään (suhteellista) **frekvenssijakaumaa**,

näkökulma on **konkreettinen**:

Muuttujan x arvot on todella mitattu (esim. otokseen osuneista) tilastoyksiköistä.

- Kun tarkastellaan **todennäköisyysjakaumaa**,

näkökulma on **spekulatiivinen**:

Pohditaan, minkälaisen arvon ominaisuus X voi saada ja kuinka suurella todennäköisyydellä eri arvot ovat odotettavissa,

jos satunnaiskoe $\mathcal{E}$ joskus tehdään.

Empiirinen muuttuja x	Satunnaismuuttuja X
Tarkasteltavina - perusjoukkoon E kuuluvat - tilastoyksiköt	Tarkasteltavana - perusjoukko Ω, johon kuuluvat - satunnaiskokeen $\mathcal{E}$ alkeistapaukset
x kuvaa tilastoyksiköiden jotain ominaisuutta	X kuvaa alkeistapausten jotain ominaisuutta
Suhteellinen frekvenssijakauma: - x:n arvot x_i ja niiden - suhteelliset frekvenssit p_i	X:n todennäköisyysjakauma: - X:n mahdolliset arvot x_i ja niiden - todennäköisyydet $p_i = P(X=x_i)$
Mitä muuta analogista konkreettisella frekvenssijakaumalla	ja todennäköisyysjakaumalla on?
↓ - summafrekvensijakauma, - keskiluvut: moodi, mediaani, aritmeettinen keskiarvo… - hajontaluvut: mm. keskihajonta, varianssi … - muut tunnusluvut	↓ ?

Satunnaismuuttuja X voi olla **diskreetti** tai **jatkuva** samoin kuin empiirinen muuttuja.

Todennäköisyyksien laskeminen ja jakauman muu käsittely on erilaista näissä tapauksissa, ja niitä käsitellään erikseen.

Matemaattisesti täsmällinen käsittelytapa johtaisi nopeasti peruskurssitason ohi satunnaisilmiöiden abstraktin perusteorian tutkimiseen, jolloin soveltamisnäkökulma jää väistämättä sivuun.

Tässä tyydytään puhtaan matematiikan näkökulmasta epätäsmällisiin määritelmiin ja esitystapoihin. Päämäärä on sattuman vaikutuksen ymmärtäminen ja analysoiminen empiiristen ilmiöiden käyttäytymisessä.

2.2. Diskreeteistä todennäköisyysjakaumista

Diskreetin satunnaismuuttujan X todennäköisyysjakauman

määrittelevät satunnaismuuttujan X

arvot x_i ja niiden todennäköisyydet $p_i = P(X=x_i)$

ja todennäköisyyksien ($p_i \geq 0$) summa $\Sigma\, p_i = 1$.

- Edellisessä esimerkissä tällainen jakauma on esitetty taulukkona.

- Tässä tapauksessa jakauma voidaan kuvata myös esittämällä sääntö, jonka avulla todennäköisyydet voidaan laskea:

k harrastajaa valikoituu 6 harrastajan joukosta		Loput $4-k$ valikoituvat 4: n muun joukosta

↘ ↙

$$P(X=k) = \frac{\binom{9}{k} \cdot \binom{6}{4-k}}{\binom{15}{4}}, \quad k = 0, 1, 2, 3, 4.$$

otoksia ↑ yhteensä

Tällaista sääntöä sanotaan **frekvenssifunktioksi**.

Diskreetillä satunnaismuuttujalla X voi olla

- äärellinen määrä arvoja, kuten edellä tai

- numeroituvasti ääretön määrä.

Esim. Optimisti täyttää joka viikko yhden lottorivin.

Satunnaismuuttuja

X = odotusaika viikkoina, kunnes tulee 1. päävoitto

X:n arvo voi olla 1, 2, 3, …

Tähän (geometriseen) jakaumaan ei syvennytä tässä tarkemmin.

Empiiristä suhteellisen summafrekvenssijakauman käsitettä vastaa todennäköisyyslaskennan käsitteistössä kertymäfunktio:

Satunnaismuuttujan X **kertymäfunktio** F kuvaa nimensä mukaisesti todennäköisyyksien kertymää ja vastaa kysymykseen:

Kuinka todennäköistä on, että satunnaismuuttuja X saa korkeintaan x:n suuruisen arvon?

Siis

F(x) = P(X≤ x) ← yleinen määritelmä kaikille satunnaismuuttujille

$= \sum_{x_i \le x} p_i$ ← Diskreetillä muuttujalla

↑ lasketaan yhteen todennäköisyydet ”x:ään asti”

Kertymäfunktion arvot kasvavat ”hyppäyksittäin” arvojen x_i kohdalla, jonka vuoksi myös sen kuvaaja on porrasfunktion kuvaaja.

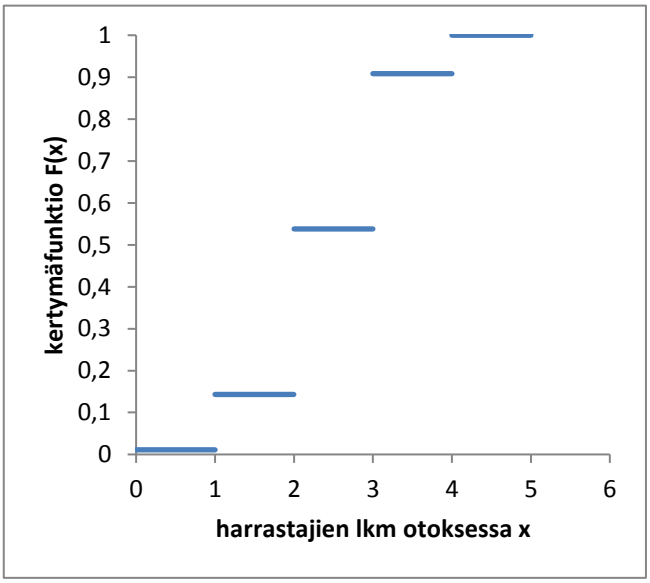
Esim. (jatkoa edelliseen)

Vain arvojen 0, 1, 2, 3, 4 kohdalla todennäköisyyksien kertymä kasvaa ja niiden välillä se ei muutu. Silloin riittää, kun kertymäfunktion arvot esitetään aikaisemman taulukon tavoin sen jatkeena:

x_i	0	1	2	3	4	Σ
p_i	15/1365 $\approx$ 0.011	180/1365 $\approx$ 0.132	540/1365 $\approx$ 0.396	504/1365 $\approx$ 0.369	126/1365 $\approx$ 0.092	1
$F(x_i)$	15/1365	195/1365	735/1365	1239/1365	1	

Esim. $P(X \leq 2) = F(2) = 735/1365 \approx 0.538$

Kuvaaja on samanlainen kuin diskreetin empiirisen muuttujan suhteellinen sunnmafrekvenssikäyrä.



Tässä X:n arvoja on vähän ja kertymäfunktio ja sen kuvaaja eivät auta paljon tämän jakauman käsittelyssä.

Sen sijaan myöhemmin jatkuvien jakaumien käsittelyssä tilanne on täysin toinen.

Esim. (jatkoa) Otokseen osuvien kuntosaliliikunnan harrastajien lukumäärän todennäköisyysjakauma on kuvitelma,

- minkälainen (empiirinen) suhteellinen frekvenssijakauma olisi **keskimäärin odotettavissa**, jos

- perusjoukosta vaivauduttaisiin todella poimimaan hyvin suuri määrä, vaikkapa 1000 000 kpl, tällaisia 4 suuruisia otoksia.

Silloin olisi siis keskimäärin odotettavissa empiirinen jakauma

Harrastajia otoksessa x_i	0	1	2	3	4	Σ
Frekvenssi f_i	11000	132000	396000	369000	92000	1000000
Suht.frekv. p_i	0.011	0.132	0.396	0.369	0.092	1.000

Jos tällainen miljoonan otoksen poimimiskoe todella tehtäisiin, tulos ei varmaankaan olisi näin siisti. Tilannetta voidaan simuloida tietokoneella satunnaislukujen avulla. Kokemus osoittaa, että todellakin ”tulos pyörii” tämän tilanteen ympärillä.

Tästä ”keskimäärin odotettavissa olevasta” jakaumasta voidaan laskea tunnuslukuja.

Aritmeettinen keskiarvo:

$$\bar{x} = \frac{1}{n} \Sigma f_i \cdot x_i$$

$$= \frac{1}{1000000} (11000 \cdot 0 + 132000 \cdot 1 + 396000 \cdot 2 + 369000 \cdot 3 + 92000 \cdot 4)$$

$$= \frac{11000}{1000000} \cdot 0 + \frac{132000}{1000000} \cdot 1 + \frac{396000}{1000000} \cdot 2 + \frac{3691000}{1000000} \cdot 3 + \frac{92000}{1000000} \cdot 4$$

satunnaismuuttujan X arvot x_i

↙ ↙ ↓ ↘ ↘

$= 0.011 \cdot 0 + 0.132 \cdot 1 + 0.396 \cdot 2 + 0.369 \cdot 3 + 0.092 \cdot 4 \approx 2.4$

↖ ↖ ↑ ↗ ↗

todennäköisyydet $p_i = P(X = x_i)$

Siis laskettiin (satunnais-)muuttujan arvojen x_i suhteellisilla frekvensseillä (todennäköisyyksillä) p_i painotettu keskiarvo.

Tällaista todennäköisyysjakaumasta laskettavaa keskiarvoa sanotaan satunnaismuuttujan X **odotusarvoksi**, josta käytetään merkintää **EX** (tai usein myös $\boldsymbol{\mu}$).

Vastaavalla tavalla kuin konkreettisessa empiirisessä jakaumassa

- aritmeettinen keskiarvo kuvaa konkreettisesti **havaittujen arvojen x_i keskimääräistä suuruutta,**

niin todennäköisyysjakaumassa

- satunnaismuuttujan X odotusarvo EX kuvaa (”arvioi ajassa eteenpäin”) X:n **keskimäärin odotettavissa olevan arvon suuruutta**, jos satunnaiskoe $\mathcal{E}$ joskus tullaan tekemään.

Varianssi ja keskihajonta:

$$\boldsymbol{\sigma^2 = \frac{1}{n} \Sigma f_i (x_i - \bar{x})^2}$$

$$= \frac{1}{1000000} (11000 \cdot (0 - 2.4)^2 + \ldots + 92000 \cdot (4 - 2.4)^2)$$

$$= \frac{11000}{1000000} \cdot (0 - 2.4)^2 + \ldots + \frac{92000}{1000000} \cdot (4 - 2.4)^2$$

satunnaismuuttujan X muunnetut arvot $(x_i - EX)^2$ $(EX = \bar{x})$

↙ ↓ ↘

$= 0.011 \cdot (0 - 2.4)^2 + \ldots + 0.092 \cdot (4 - 2.4)^2 \approx 0.75.$

↖ ↑ ↗

todennäköisyydet $p_i = P(X = x_i)$

Satunnaismuuttujan X varianssista käytetään myös merkintää Var(X) (ja myös D^2X).

Tässä on

hajonta $\sigma = \sqrt{0.75} = 0.87$.

Satunnaismuuttujan X hajonnasta käytetään myös merkintää DX.

Konkreettisessa empiirisessä jakaumassa

- keskihajonta σ kuvaa, kuinka suuri on kaikkien tarkasteltavien muuttujan **arvojen x_i ”keskimääräinen poikkeama” keskiarvosta $\bar{x}$**.

Todennäköisyysjakaumassa

- satunnaismuuttujan X hajonta σ kuvaa (”arvioi ajassa eteenpäin”) X:n **arvojen keskimäärin odotettavissa olevan vaihtelun suuruutta**, jos satunnaiskoe $\mathcal{E}$ joskus tullaan tekemään.

Siis

Diskreetin satunnaismuuttujan X jakauman

odotusarvo on X:n arvojen x_i todennäköisyyksillä p_i painotettu keskiarvo

$$\boldsymbol{EX = \sum p_i x_i}$$

ja

varianssi muunnettujen (keskistettyjen ja neliöityjen) arvojen $(x_i - EX)^2$ todennäköisyyksillä p_i painotettu keskiarvo

$$\boldsymbol{Var(X) = \sum p_i (x_i - EX)^2}$$

ja

hajonta $\boldsymbol{DX = \sqrt{Var(X)}}$.

Empiirisen jakauman ja todennäköisyysjakauman käsitteet voivat olla joskus ”hyvin lähellä” toisiaan. Tutkitaan samoja olioita ja erona on ainoastaan

- empiirisen jakauman konkreettinen ”toteava” näkökulma ja

- todennäköisyysjakauman spekulatiivinen ”ennustava” näkökulma:

Esim. Markkinatutkijalla on kuntosaliketjun jäsenien rekisteri (perusjoukko E), jossa eräs tieto on jäsenien aktiivisesti harrastamien liikuntalajien määrä. Taulukossa on

muuttujan

x = lajien lkm

frekvenssijakauma

ja suhteellinen

frekvenssijakauma:

lajeja x_i	frekvenssi f_i	suht. frekv. $p_i = f_i/n$
1	198	0.040
2	1164	0.233
3	1215	0.243
4	632	0.126
5	789	0.158
6	362	0.072
7	124	0.025
8	176	0.035
9	342	0.068
yhteensä	5002	1.00

(Joskus tulevaisuudessa ehkä tehtävä) otoksen poimiminen voidaan ajatella satunnaiskokeen

$\mathcal{E}$ = ”Arvotaan perusjoukosta E (vain) yksi tilastoyksikkö.”

toistamiseksi (toisistaan riippumatta) n kertaa.

$\mathcal{E}$:n alkeistapaukset ovat samat ihmiset kuin rekisterin **tilastoyksiköt**

ja **perusjoukko** Ω satunnaiskokeessa $\mathcal{E}$ on sama kuin empiirinen perusjoukko E.

Sattuma määrää, kuinka monta lajia valittavalla on

ja esim.

P(”Lajeja on vain yksi”) = P(X = 1) = 198/5002 = 0.04 jne.

Satunnaismuuttujan

X = Valittavan henkilön lajien määrä

todennäköisyysjakauma on taulukossa oleva suhteellinen frekvenssijakauma.

Kertymäfunktion arvot ovat vastaavalla tavalla

samat kuin suhteelliset summafrekvenssit.

lajeja x_i	$p_i = P(X=x_i)$	$F(x_i)$
1	0.040	0.040
2	0.233	0.273
3	0.243	0.516
4	0.126	0.642
5	0.158	0.800
6	0.072	0.872
7	0.025	0.897
8	0.035	0.932
9	0.068	1.000
yhteensä	1.00	

Satunnaismuuttujan X:n **todennäköisyysjakauman odotusarvo**

$EX = \sum p_i x_i$

$= 0.040 \cdot 1 + 0.233 \cdot 2 + \ldots + 0.068 \cdot 9$

$= 4.028$

on $\bar{x}$ = harrastettujen lajien määrän **keskiarvo perusjoukossa** ($\frac{1}{n} \sum f_i \cdot x_i$)

Satunnaismuuttujan X **todennäköisyysjakauman varianssi** on

$Var(X) = \sum p_i (x_i - EX)^2$

$= 0.040 \cdot (1-4.028)^2 + 0.233 \cdot (2-4.028)^2 + \ldots + 0.068 \cdot (9-4.028)^2$

$\approx 4.465,$

joka taas on sama arvo kuin lajien määrän x

empiirinen varianssi perusjoukossa E

$\sigma^2 = \frac{1}{n} \sum f_i (x_i - \bar{x})^2$

Satunnaismuuttujan X:n **hajonta**

$DX = \sqrt{4.465} \approx 2.10,$

on sama kuin

empiirisen muuttujan x **keskihajonta** perusjoukossa E.

Esimerkissä näkyy paljon **sattuman käyttäytymistä sääteleviä lainalaisuuksia:**

- Muuttujan x arvojen **suhteellinen frekvenssijakauma perusjoukossa E** on **näkökulmasta riippuen** x:n arvojen konkreettinen empiirinen jakauma mutta myös satunnaismuuttujan X todennäköisyysjakauma.

- Satunnaismuuttujan X (edellä lajien lukumäärä arvottavalla henkilöllä) **odotusarvo,** ”keskimäärin odotettavissa oleva” arvo, on ominaisuuden x **todellinen keskiarvo perusjoukossa**.

Satunnaismuuttujan arvot ”pyörivät” todellisen keskiarvon ympärillä.

- Tätä keskiarvon ympärillä ”pyörimisen” määrää voidaan mitata (esim.) keskihajonnan σ avulla ja

Satunnaismuuttujan X **hajonta DX**, ”keskimäärin odotettavissa oleva arvojen vaihtelu” satunnaiskokeessa, on ominaisuuden x **todellinen keskihajonta σ perusjoukossa**.

- Kun poimitaan otos, tehdään satunnaiskoe

$\mathcal{E}$ = ”Arvotaan perusjoukosta E yksi tilastoyksikkö.”

(toisistaan riippumatta) n kertaa.

Otoksesta havaituista x:n arvoista tehty suhteellinen frekvenssijakauma on approksimaatio perusjoukossa E olevalle todelliselle jakaumalle.

Approksimaatio on sitä parempi, mitä suurempi otoskoko n on.

Binomijakauma ja hypergeometrinen jakauma

Monissa empiirisissä tilanteissa on tutkittavalla satunnaisilmiöllä samanlainen rakenne. Näiden mallintamiseen on käytettävissä ”valmiita” jakaumia, joiden ominaisuuksia voidaan suoraan soveltaa tarkasteltavaan tilanteeseen:

Binomijakauma

Esim. (jatkoa) 15 henkilön perusjoukosta, jossa 9 harrastaa kuntosaliliikuntaa, aiotaan poimia

4:n suuruinen **otos palauttaen**.

Edellä laskettiin permutaatioiden avulla esim. todennäköisyys, että

otokseen osuu 2 harrastajaa eli P(X=2), kun satunnaismuuttuja

X = harrastajien lukumäärä tällaisessa otoksessa.

Todennäköisyydet voidaan laskea myös toisella tavalla todennäköisyyslaskennan kerto- ja yhteenlaskusäännön avulla:

- Kun otos poimitaan palauttaen, eivät henkilöiden peräkkäisten valintojen tulokset vaikuta toisiinsa ja

- jokaista henkilöä valittaessa on

P(harrastaa) = P(H) = 9/15 = 0.6 ja P(ei harrasta) = P(E) = 0.4.

- Kaikki jonot, joissa on esimerkiksi H, H, E ja E jossain järjestyksessä ovat yhtä todennäköisiä.

tapahtumat riippumattomia

↙ ↙ ↘ ↘

P(H ja H ja E ja E) = 0.6 · 0.6 · 0.4 · 0.4 = $0.6^2 \cdot 0.4^2$

- Jonoissa, joissa on 2 harrastajaa, voivat harrastajien paikat valikoitua $\binom{4}{2}$ tavalla.

- Silloin

P(X=2) = $\binom{4}{2} \cdot 0.6^2 \cdot 0.4^2 = 0.3456$.

Yleisesti: Edellisen esimerkin tavoin oleellista on:

*- Tilanne on n-kertainen **toistokoe**, jossa joka toistolla seurataan, onko tuloksena jokin tapahtuma A vai ei.*

*- Toistot ovat toisistaan **riippumattomia**.*

*- Joka toistolla A:n esiintymisen **todennäköisyys p = P(A) on sama**.*

*- Satunnaismuuttuja X = A:n esiintymisten **lukumäärä** toistosarjassa.*

Silloin sanotaan, että

satunnaismuuttuja X on ***binomijakautunut parametrein n ja p***,

mistä käytetään merkintää ***X∼Bin(n, p)***.

Edellä

- henkilön arpominen otokseen toistettiin n = 4 kertaa,
- otannassa palauttaen toistojen tulokset eivät vaikuta toisiinsa,
- p = P(H) = 0.6 jokaista henkilöä valittaessa ja
- X = kuntosaliliikunnan harrastajien lkm toistosarjassa",

joten harrastajien määrä tässä palauttaen poimittavassa "otoksessa"

X∼Bin(4, 0.6).

Kun X∼Bin(n, p), jakauman todennäköisyyksien laskemiseksi voidaan

esittää frekvenssifunktio samalla tavalla kuin edellisessä erikoistapauksessa:

- Tapahtumalle X = k suotuisia alkeistapauksia ovat jonot, joissa

tapahtuma A on realisaationa k kertaa ja A^c n-k kertaa **jossain kohtaa** jonossa.

Eräs tällainen jono ja sen todennäköisyys ovat

k kpl n-k kpl (riippumattomuus!)

↙ ↓ ↘ …↘ ↙ ↓…↘ ↘

P(A ja A ja … ja A ja A^c ja A^c ja … ja A^c) = $\boxed{p^k(1-p)^{n-k}}$

↑

- Kaikki tapahtumalle ”X = k” **suotuisat jonot ovat yhtä todennäköisiä**.

Jos jonossa on A tuloksena k kertaa ja A^c tuloksena n-k kertaa, sen todennäköisyys on tulo, jossa

tekijänä on A:n kohdalla p= P(A) k kertaa ja A^c:n kohdalla 1-p n-k kertaa.

- n:n pituisesta jonosta ? ? ? … ? ?

k paikkaa tapahtumille A ↖ ↖ … ↗ ↗

voivat valikoitua $\binom{n}{k}$ tavalla.

Siis jos *X*∼*Bin(n, p),*

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

Voidaan melko helposti osoittaa, että odotusarvo ja varianssi saadaan säännöillä

$$EX = np \quad \text{ja} \quad Var(X) = np(1-p),$$

kun X∼Bin(n, p).

Esim. (jatkoa) Harrastajien määrä otoksessa X∼Bin(4, 0.6).

EX = 4·0.6 = 2.4, Var(X) = 4·0.6·(1-0.6) = 0.96 ja DX = $\sqrt{0.96} \approx 0.98$.

Binomijakaumaa voidaan käyttää mallina myös monissa muissa tilanteissa kuin pelkästään otannassa palauttaen:

Esim. Juomien makutestissä voi saada arvonimen

Melko Suuri Maistaja (MSM *), Suuri maistaja (SM**) tai Erittäin Suuri Maistaja EMS***).

Tulos määräytyy kokeella, jossa maistaja yrittää erottaa Erityisjuoman kolmen vastaavanlaisen juoman joukosta.

Maistaminen toistetaan 13 kertaa ja tulos on oikein tunnistettujen kertojen määrä.

Maistaja on

MSM*, jos tuloksen voi saavuttaa arvaamalla alle 5 % tapauksista

SM**, jos tuloksen saa arvaamalla alle kerran sadasta

ESM***, jos tuloksen saa arvaamalla vain alle 0.1 % tapauksista.

Ekonomisti E tunnistaa 13 yrityksestä 10 kertaa juoman oikein.

Jos tunnistaminen olisi **arvaamista**, niin

↖ (Spekuloidaan tällä **hypoteesilla**!)

p = P(”Arvaus on oikein.”) = 1/3 on

sama kaikissa 13 toistossa,

joita voidaan pitää toisistaan **riippumattomina**.

Silloin

oikeiden arvausten **lukumäärä** X ∼Bin(13, 1/3).

Keskimäärin on arvaamalla odotettavissa EX = $13 \cdot \frac{1}{3} \approx 4.3$ oikeaa ja

”keskimäärin odotettavissa oleva vaihtelu” tämän arvon ympärillä on

DX = $\sqrt{13 \cdot \frac{1}{3} \cdot \frac{2}{3}} \approx 1.7$ ”onnistumista”.

Tulos 10 oikein on siis hyvä, mutta mihin se riittää?

Vähintään 10 oikeaa, joka oli tuloksena, tai vielä paremmin saadaan arvaamalla todennäköisyydellä

$P(X \geq 10) = P(X = 10) + P(X = 11) + P(X = 12) + P(X = 13)$

$$= \binom{13}{10} \cdot \left(\frac{1}{3}\right)^{10} \cdot \left(\frac{2}{3}\right)^{3} + \binom{13}{11} \cdot \left(\frac{1}{3}\right)^{11} \cdot \left(\frac{2}{3}\right)^{2}$$

$$+ \binom{13}{12} \cdot \left(\frac{1}{3}\right)^{12} \cdot \left(\frac{2}{3}\right)^{1} + \binom{13}{13} \cdot \left(\frac{1}{3}\right)^{13} \cdot \left(\frac{2}{3}\right)^{0}$$

$\approx 0.001435 + 0.00019562 + 0.0000163 + 0.0000006$

$\approx 0.00165 = 0.165\ \%$

Siis keskimäärin vain alle 2 kertaa tuhannesta näin hyvä tulos saavutetaan pelkästään arvaamalla.

Tulos riittävän **merkitsevä** SM** tittelin saamiseen, mutta ei

kuitenkaan niin **erittäin merkitsevä**, että se tukisi riittävästi

ESM*** tason myöntämistä.

Tässä aloitellaan makutestiä paljon tärkeämpää kysymystä **hypoteesien testaamisesta**, johon syvennytään tarkemmin myöhemmin.

Edellisissä esimerkeissä korrekti otanta ja koesuunnittelu takaavat toistojen riippumattomuuden ja tarkasteltavan tapahtuman esiintymistodennäköisyyden pysymisen samana joka toistossa.

Sovelluksissa joudutaan joskus pohtimaan, ovatko nämä käyttöedellytykset riittävästi voimassa.

Esim. Koripalloilija saa 80 % vapaaheitoistaan koriin.

Kuinka todennäköistä on, että hän saa ottelussa 10:stä vapaaheitosta korkeintaan 9 koriin.

Heittosarja on 10-kertainen **toistoko**e ja onnistumisten määrää kuvaa

X = korien **lkm** 10 heitossa.

Kuitenkaan **toistojen riippumattomuus** ja **onnistumistodennäköisyyden** säilyminen **samana** ei ole itsestään selvää.

Jos kuitenkin voidaan olettaa, että likimain X $\sim$Bin(10, 0.8),

saadaan

$P(X \leq 9) = 1 - P(X=10) = 1 - \binom{10}{10} \cdot 0.8^{10} \cdot 0.2^{0} \approx 0.893$

Muita diskreettejä jakaumia, joissa satunnaismuuttuja kuvaa jonkin tapahtuman A esiintymisten määrää toistokokeessa ovat:

- **Bernoulli-jakauma**, joka on binomijakauman erikoistapaus,

jossa n=1. Tämä jakauma on tärkeä apuväline teoriatarkasteluissa.

- **Poisson-jakauma**, jossa A on jokin hyvin harvinainen tapahtuma.

Näitä ei käsitellä tässä.

Otantatilannetta tutkittaessa binomijakauman lähisukulainen on

Hypergeometrinen jakauma

Esim. (jatkoa) 15 henkilön perusjoukosta, jossa 9 harrastaa kuntosaliliikuntaa, aiotaan poimia

4:n suuruinen **otos palauttamatta**.

Edellä hahmotettiin

- satunnaiskokeeksi tällaisen otoksen poimiminen ja sen

- alkeistapauksiksi kaikki mahdolliset järjestämättömät otokset (kombinaatiot).

Tarkasteltiin satunnaismuuttujaa

X = harrastajien **lukumäärä** palauttamatta poimittavassa otoksessa,

ja sen todennäköisyyksien laskemista varten saatiin frekvenssifunktio

$$P(X=k) = \frac{\binom{9}{k} \cdot \binom{6}{4-k}}{\binom{15}{4}}, \quad k = 0, 1, 2, 3, 4.$$

Oleellisia kohtia (parametreja) tässä laskussa ovat

perusjoukon koko N=15, harrastajien määrä perusjoukossa K=9 ja otoskoko n=4.

Sanotaan, että tällainen satunnaismuuttuja

X on **hypergeometrisesti jakautunut** parametrein 15, 9 ja 4, mistä käytetään

merkintää X ∼Hyperg(15, 9, 4).

Siis yleisesti:

X∼Hyperg(N, K, n), kun tilanne voidaan hahmottaa niin, että

- N:n kokoisessa perusjoukossa on

- K tilastoyksikköä, joilla on ominaisuus A,

- perusjoukosta poimitaan **palauttamatta** n:n suuruinen otos ja

- satunnaismuuttuja

X = niiden tilastoyksiköiden lukumäärä, joilla on ominaisuus A.

Samalla tavalla kuin edellä esimerkissä saadaan jakauman frekvenssifunktio

K: sta tilastoyksiköstä, joilla ominaisuus A valikoituu otokseen k kpl

N − K: sta muusta tilastoyksiköstä valikoituu otokseen n − k kpl

↘ ↙

$$P(X=k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}, \; k \leq K \text{ ja } n-k \leq N-K$$

otoksia ↑ yhteensä

Myös hypergeometrisen jakauman odotusarvon ja varianssin laskemiseksi voidaan johtaa yksinkertaiset säännöt:

Jos **X∼Hyperg(N, K, n)**, niin

odotusarvo $EX = np$, **missä** $p = \frac{K}{N}$ **(=P(A)),**

varianssi $Var(X) = np(1-p) \cdot \frac{N-n}{N-1}$ **,**

hajonta $DX = \sqrt{np(1-p) \cdot \frac{N-n}{N-1}}$

Tekijää $\frac{N-n}{N-1}$ sanotaan ***äärellisen perusjoukon korjaustekijäks*i**.

Siis odotusarvo on sama kuin otannassa palauttaen, mutta varianssi on pienempi.

Edellisen esimerkin tilanteessa

EX = $4 \cdot \frac{9}{15} = 4 \cdot 0.6 = 2.4$ sekä otannassa palauttaen että palauttamatta.

Siis kummallakin tavalla 4:n suuruiseen ”otokseen” on odotettavissa

- keskimäärin sama määrä 2.4 kuntosaliharrastajia, mikä taas on

- suhteellisena osuutena $\frac{2.4}{4} \cdot 100\ \% = 60\ \%$ eli

harrastajia on otokseen keskimäärin odotettavissa samassa suhteessa kuin perusjoukossakin on!

Otoksesta saatava harrastajien suhteellinen osuus ”vastaa hyvin” todellista perusjoukossa olevaa osuutta.

Otannassa palauttamatta varianssi on

$$\text{Var}(X) = 4 \cdot 0.6 \cdot (1-0.6) \cdot \frac{\mathbf{15-4}}{\mathbf{15-1}} = 0.96 \cdot \frac{\mathbf{11}}{\mathbf{14}} \ (\approx 0.96 \cdot \mathbf{0.786}) = 0.64$$

↑

Siis varianssi (”sattuman pelivara”) on tässä tilanteessa noin 21.4 % pienempi kuin otannassa palauttaen.

Siis jo intuitiivisesti luonnollisempi

- otantatapa palauttamatta on myös ”edullisempi” kuin otanta palauttaen,

- kun otantatutkimuksessa otoksen perusteella arvioidaan (estimoidaan) tilastoyksiköiden jonkin ominaisuuden A suhteellisen osuuden suuruutta perusjoukossa.

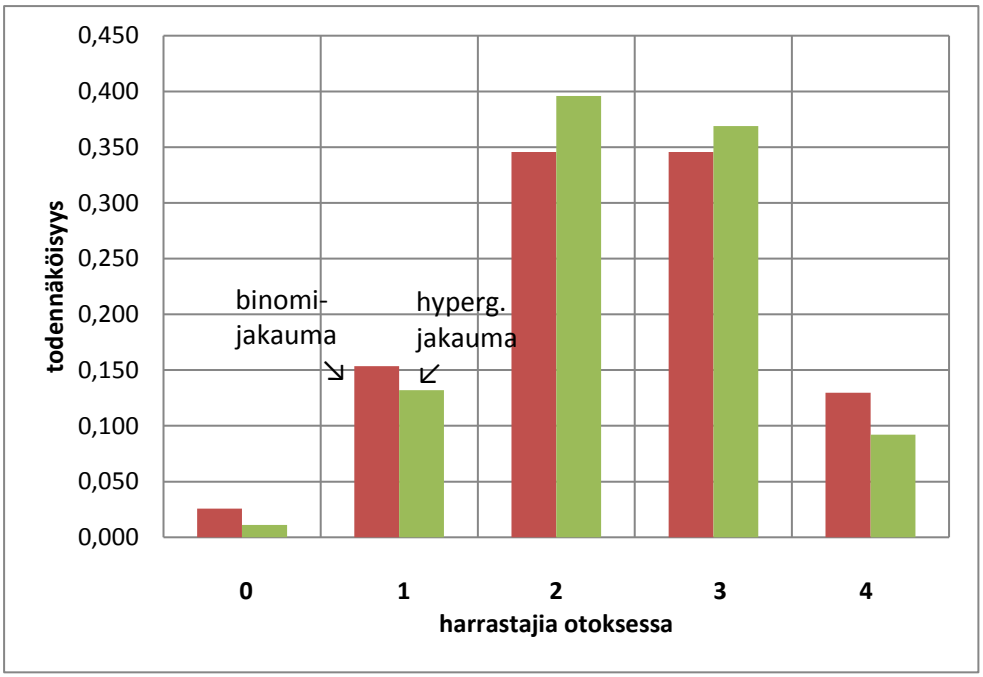
Edellinen esimerkki on melko äärimmäinen. Perusjoukon koko ja otoskoko ovat oikeassa tutkimustilanteessa paljon suurempia. Silloin myös näiden otantatapojen tuottamien tulosten tarkkuuksien ero ei ole suuri.

Binomi- ja hypergeometrisen jakauman todennäköisyydet saadaan myös helposti Excelin funktioiden avulla:

Formulas → More Functions → Statistical → BINOM.DIST
tai HYPGEOM.DIST

Edellä esimerkissä perusjoukon koko oli äärimmäisen pieni, jolloin se alkaa tyhjentyä merkittävästi pientäkin otosta palauttamatta poimittaessa ja vaihtelu pienenee. Otannassa palauttaen samat henkilöt ovat aina valittavissa. Siksi todennäköisyysjakaumat poikkeavat niin, että hypergeometrinen jakauma keskittyy tiiviimmin odotusarvonsa ympärille.

x	bin.jak	hyperg.
0	0,026	0,011
1	0,154	0,132
2	0,346	0,396
3	0,346	0,369
4	0,130	0,092



- Hypergeometrinen jakauma keskittyy tiiviimmin odotusarvon EX = 2.4 ympärille.

- Kuviosta näkyy myös, että kumpikin jakauma on lievästi **vasemmalle vino** (vasemmalle päin on pidempi ”häntä”.) Näin on aina, kun

”onnistumistodennäköisyys” $p > 0.5$ (Tässä p = P(harrastaa) = 9/15 = 0.6).

Jakauma on **symmetrinen**, jos p = 0.5 ja **oikealle vino**, jos $p < 0.5$.

Jos perusjoukko on ”suuri” (tai jopa ääretön), edellisen tilastoyksikön poimiminen ei muuta sitä paljon otannassa palauttamatta.

Silloin tilastoyksiköitä arvottaessa peräkkäisissä toistoissa

”onnistumistodennäköisyys” p = P(A) pysyy lähes samana ja toistot ovat ”lähes riippumattomia” toisistaan.

Tällaisessa tilanteessa binomi- ja hypergeometrisen jakauman ero on pieni.

Esim. Toimittaja aikoo tehdä viiden ohjelman sarjan, jossa jokaisessa ohjelmassa esiintyy kansanedustaja.

- Tasapuolisuuttaan korostaakseen hän arpoo esiintyvät kansanedustajat eli poimii 5 suuruisen otoksen 200 kansanedustajasta.

- Etukäteen hän pohtii, montakohan opposition edustajaa (ohjelmaa suunnitellessa 88:sta opposition kannattajasta) tulee valituiksi otokseen

eli minkälainen on satunnaismuuttujan

X = opposition kannattajien lukumäärä otoksessa

jakauma.

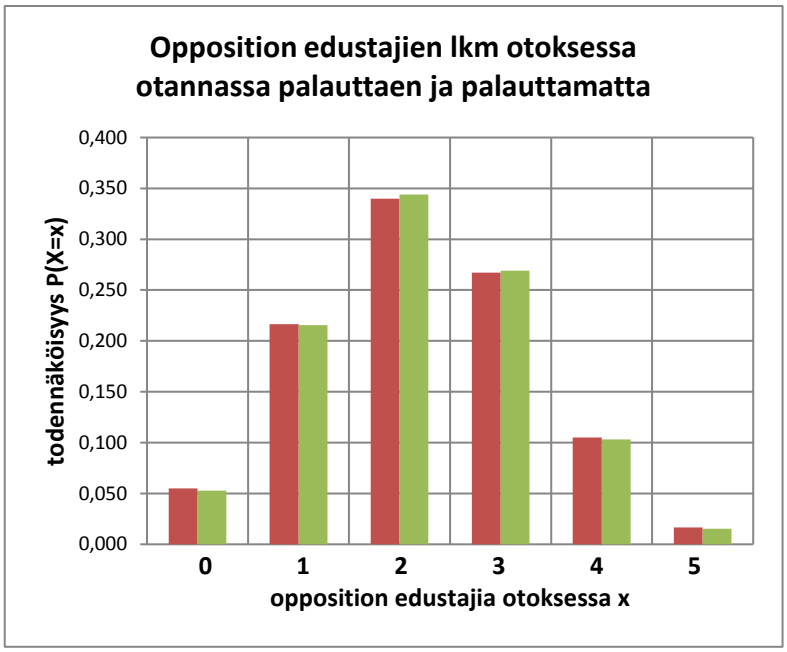
Jos otos poimitaan palauttaen, X~ Bin(5, 88/200) = Bin(5, 0.44) ja

$P(X = k) = \binom{5}{k} 0.44^k (1 - 0.44)^{5-k}$.

Jos otos poimitaan palauttamatta X~Hyperg(200, 88, 5) ja

$P(X=k) = \dfrac{\binom{88}{k} \cdot \binom{112}{5-k}}{\binom{200}{5}}$.

x	bin.jak. P(X=x)	hyperg.jak. P(X=x)
0	**0,055**	**0,053**
1	**0,216**	**0,216**
2	**0,340**	**0,344**
3	**0,267**	**0,269**
4	**0,105**	**0,103**
5	**0,016**	**0,015**



Vasemmalla puolella on binomijakauma ja oikealla hypergeometrinen.

Erot ovat hyvin pienet, vaikka N = 200 ei olekaan kovin suuri. Kuitenkin tässäkin hypergeometrinen jakauma keskittyy vähän tiiviimmin odotusarvon EX = 5 · 0.44 = 2.2 ympärille.

Jakauma on lievästi oikealle vino ($p = 0.44 < 0.5$).

Tässä laskut onnistuvat vielä käsityönäkin, mutta taulukkoon ne on saatu Excelillä. (Laske kuitenkin järjestämättömien otosten määrä $\binom{200}{5}$.)

Otantatilanteen tutkimista varten tarvitaan myös muunlaista apua binomi- ja hypergeometrisen jakauman käsittelyyn. Osoittautuu, että näiden jakaumien todennäköisyyksiä voidaan laskea kätevästi (jatkuvan) normaalijakauman avulla. Samalla saadaan normaalijakauman hyvät ominaisuudet käyttöön tilastollisen päättelyn menetelmiä varten.

2.3 Jatkuvista todennäköisyysjakaumista

Esim. Linja-auton vuoroväli on 10 minuuttia.

Satunnaiskoe ε = ”Menet aikataulusta tietämättä pysäkille.” ja

satunnaismuuttuja X = odotusaika.

- Kaikki reaaliluvut välillä [0, 10], joita on ylinumeroituva määrä, ovat mahdollisia X:n arvoja.

- Silloin ei voida ”luetella” kaikkia X:n arvoja ja niiden todennäköisyyksiä

- eikä olisi edes (empiirisesti) kovin järkevää miettiä jonkin täsmällisen **yksittäisen arvon todennäköisyyttä**, kuten esim.

P(X = π min) = P(X = 3.1414…) ?

Tällaisen (äärettömän monen desimaalin tarkkuudella ajateltavan) arvon esiintymisen todennäköisyyden arvion on oltava P(X = 3.1414…) = 0.

Siis ajatellaan, ”ettei se nyt ihan tarkalleen π:n kohdalle kyllä osu”, vaikka lähelle menisikin.

- Sen sijaan vaikkapa tapahtumalle $2.5 \leq X \leq 4.0$ min on odotettavissa realisaatioita, jos satunnaiskoe ε toistetaan monta kertaa.

- Klassisen todennäköisyyden ajatuksena on, että kaikkien alkeistapausten todennäköisyydet ovat yhtä suuret eli ”koko todennäköisyys 100 %” jakautuu tasaisesti kaikille vaihtoehdoille.

- Tässä erikoistapauksessa voidaan ilmeisesti järkevästi päätellä vastaavalla tavalla:

Kaikkia odotusajat 0:n ja 10 minuutin välillä ovat ”yhtä mahdollisia”,

välin [2.5, 4.0] pituus on 4.0 – 2.5 = 1.5 min on $\frac{1.5}{10} \cdot 100\ \%$ = 15 % koko vuorovälin pituudesta, joten

todennäköisyys P(2,5 $\leq$ X $\leq$ 4.0 min) = 0.15.

- Todella havainnoimalla odotusaikoja tai simuloimalla niitä esim. Excelin avulla voidaan kokeellisesti tarkastella samaa asiaa.

Siis jatkuvien satunnaismuuttujia käsiteltäessä lasketaan todennäköisyyksiä

että X:n arvo osuu satunnaiskokeessa jollekin **välille**.

Seuraavassa tilanteessa äskeinen ”symmetria-oletus” ei kuitenkaan toimi laskemisen lähtökohtana:

Esim. Tutkittiin Härvelitehtaan tuottamien härvelien kestoikää.

Perusjoukkona (empiirisessä mielessä) ovat kaikki samalla menetelmällä tuotetut härvelit ja lisäksi voidaan ottaa vielä tulevaisuudessa tuotettavat ja vielä nekin, jotka voitaisiin tuottaa samalla menetelmällä.

Todellisuudessa perusjoukko on tietysti äärellinen, mutta se on niin suuri, että se voidaan(?) kuvitella äärettömäksi.

Tutkittava (empiirinen) muuttuja x = härvelin kestoikä on jatkuva.

Perusjoukosta poimittiin aluksi 100 suuruinen otos, härvelit käytettiin loppuun ja kestoajoista saatiin frekvenssijakauma:

kestoaika (h)	frekv. f_i	suht. frekv. p_i
700 – 949	4	0.04
950 – 1049	25	0.25
1050 – 1099	21	0.21
1100 – 1149	20	0.20
1150 – 1249	23	0.23
1250 - 1499	7	0.07
Σ	100	1.00

1) Taulukossa olevat **suhteelliset frekvenssit ovat todennäköisyyksiä** satunnaiskokeessa

ε = ”Valitaan umpimähkään jokin **otokseen osuneista** härveleistä

(, jotka siis ovat jo loppuun käytettyinä roskiksessa!).

Kun satunnaismuuttujana on X = aika, jonka härveli **kesti**,

on esimerkiksi **todennäköisyys** (klassinen todennäköisyys!)

P(950 $\leq$ X $\leq$ 1049) = 0.25 = 2. luokan **suhteellinen frekvenssi** p_2.

Ei ole kuitenkaan kovin mielenkiintoista arvailla loppuun käytettyjen härvelien kestoikien suuruutta, vaan

otoksen informaatioon perustuvat päätelmät yleistetään koko perusjoukkoon.

Tämän välittää satunnaiskoe

ε = ”Valitaan umpimähkään härveli tehtaan tuotannosta (**perusjoukosta**)”

ja satunnaismuuttuja

X = tuotannosta valit**tavan** härvelin kestoikä.

Kun otos on poimittu korrektisti ”lappuja hatusta”-periaatteella, se on **edustava osa** perusjoukosta.

Silloin ilmeisesti edelleen esimerkiksi

suhteellinen frekvenssi $p_2 = 0.25$ on hyvä **arvio todennäköisyydelle**, että kestoikä tulee olemaan 950 ja 1049 tunnin välissä eli

$P(950 \leq X \leq 1049) \approx 0.25$.

Välillä siirrytään tarkastelussa toiseen suuntaan.

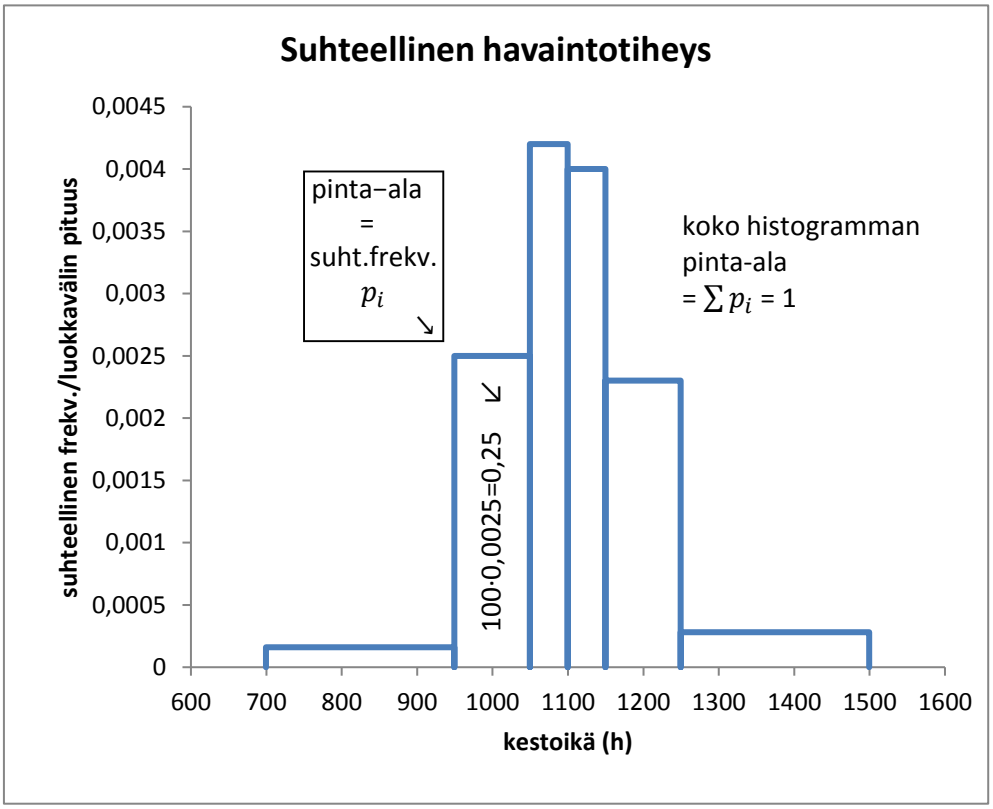
2) Otoksen suhteellinen frekvenssijakauma histogrammana:

Luokitus on epätasavälinen, ja

- pylväiden pinta-alat saadaan verrannollisiksi suhteellisiin frekvensseihin
- piirtämällä **suhteellisen havaintotiheyden =** $\frac{\mathbf{p_i}}{\mathbf{c_i}}$ korkuiset pylväät.

Kestoikä	Suhteellinen frekvenssi p_i	Luokkavälin pituus c_i	Suhteellinen havaintotiheys p_i/c_i
700 – 949	0.04	250	0.00016
950 – 1049	0.25	100	0.00250
1050 – 1099	0.21	50	0.00420
1100 – 1149	0.20	50	0.00400
1150 – 1249	0.23	100	0.00230
1250 - 1499	0.07	250	0.00028
Σ	1.00		

Härvelien kestoajan jakauma 100 suuruisessa otoksessa



Jokaisen pylvään

pinta-ala = kanta x korkeus = $c_i \cdot (p_i/c_i) = p_i$ **= suhteellinen frekvenssi**

ja

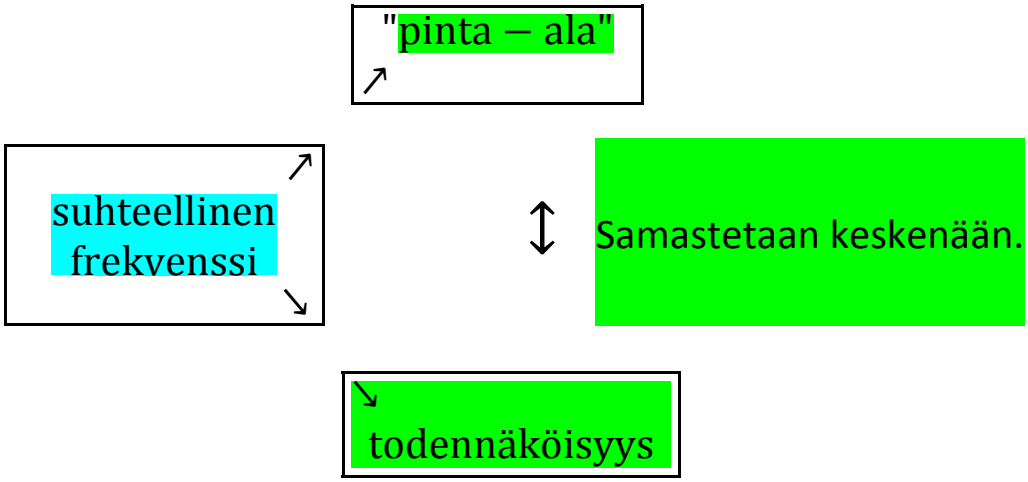
kokonaispinta-ala on $\Sigma\, p_i = 1$.

3) Silloin voidaan **samastaa** (suhteellisen frekvenssin approksimoima)

- todennäköisyys, että X:n arvo osuu satunnaiskokeessa välille [a, b]

  $P(a \leq X \leq b)$ ja

- pylväiden pinta-ala välin [a, b] kohdalla suhteellisen havaintotiheyden histogrammassa.



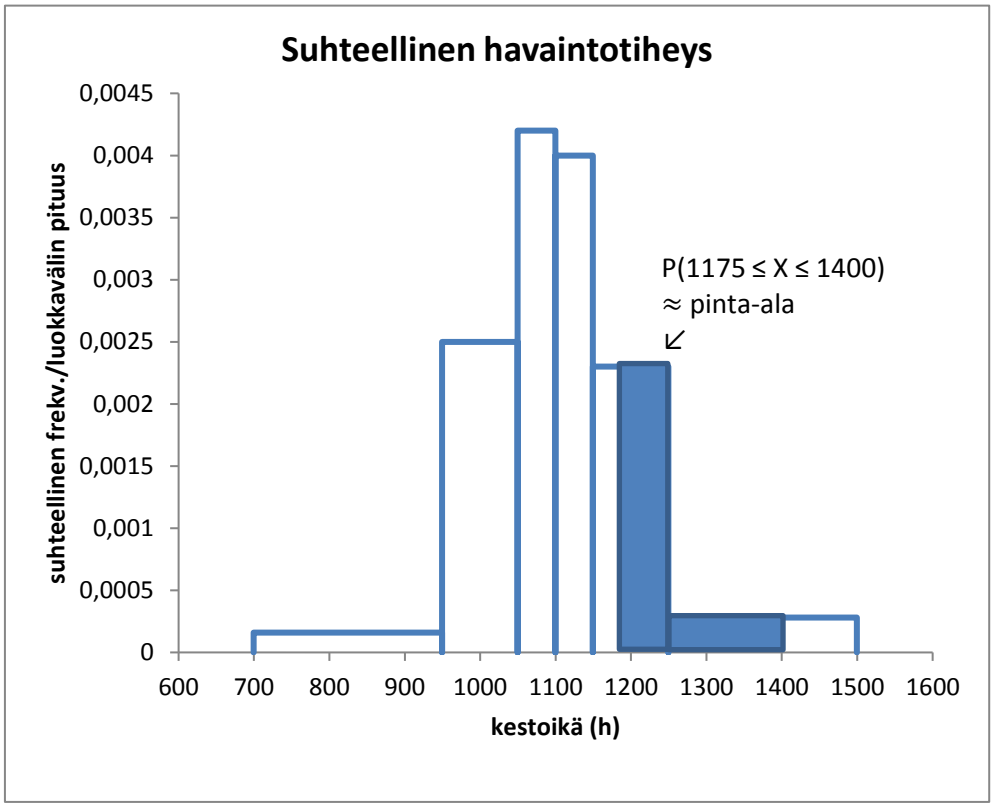
Esim.

$P(1175 \leq X \leq 1400)$

$= (1249.5 - 1175) \cdot 0.00230 + (1400 - 1249.5) \cdot 0.00028$

$= 0.17135 + 0.0414$

$\approx 0.21.$



Jatkuvan satunnaismuuttujan X jakaumalle on siis saatava sellainen esitystapa, että

todennäköisyys, että X:n arvo osuu satunnaiskokeessa välille [a, b]

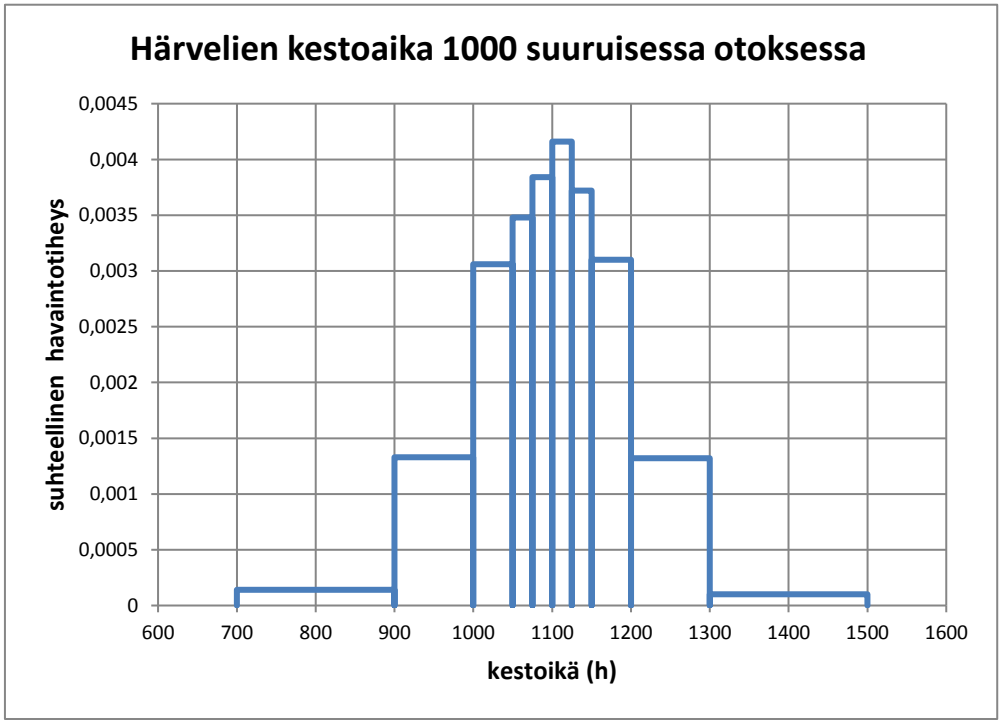
P(a≤X≤ b) voidaan esittää ”pinta-alana”.

Otantaa jatkettiin ja poimittiin 1000 härvelin otos.

Kun luokiteltavien arvoja on enemmän, luokitusta voidaan tihentää

ja

kuvio tulee säännöllisemmäksi ja arvio jakauman muodosta koko perusjoukossa tarkentuu.



Olipa otoskoko kuinka suuri tahansa,

- pylväiden kokonaispinta-ala = 1 ja

- arvio todennäköisyydelle P(a≤X≤ b) voidaan esittää ”pinta-alana”.

4) Kun otoskoko n kasvaa, myös luokkien määrää k voidaan kasvattaa esimerkiksi säännön

$\sqrt[3]{n} \leq k \leq 2 \cdot \sqrt[3]{n}$ mukaan.

- Otoskokoa ei käytännössä voida kasvattaa rajatta.

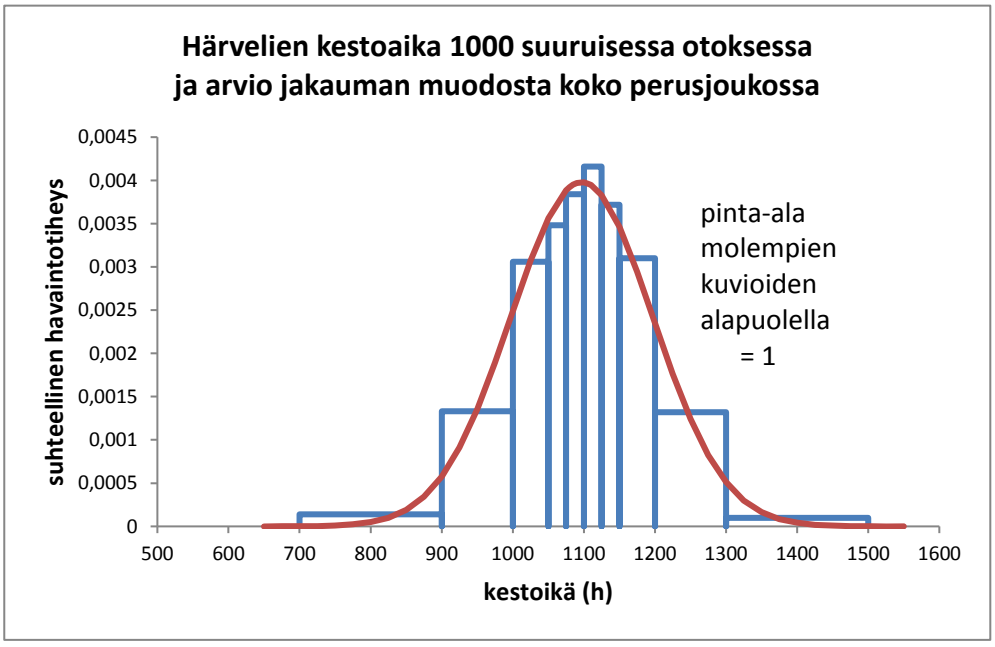
- Voidaan kuitenkin kuvitella järkevästi, mitä tapahtuisi, jos otoskokoa aina vain kasvatettaisiin:

- Luokitusta voitaisiin tihentää ja histogramman pylväiden yläreunan murtoviiva, joka rajaa alleen ykkösen suuruisen pinta-alan, lähenisi(?) sileää käyrää.

Kahdesta edellisestä kuviosta jo näkyi tästä alkua ja simulointikokeilla tästä voi jatkaa.

Satunnaislukujen avulla voidaan ”poimia” niin suuria ”otoksia” kuin halutaan.

Edellä otoskoko oli vasta vain 1000, mutta ajatuksella tätä voidaan jatkaa:



5) Jos otoskoko kasvaisi äärettömän suureksi,

- tämän voidaan ajatella vastaavan tilannetta, jossa kaikki (”äärettömän”) perusjoukon härvelit poimittaisiin otokseen.

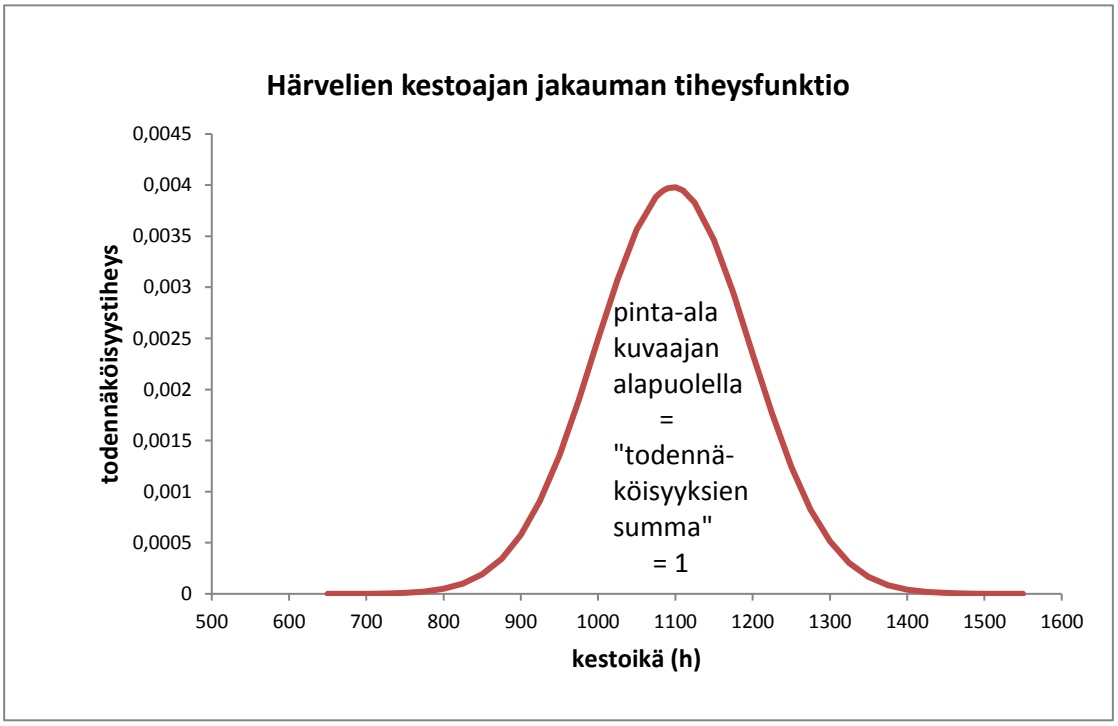
- Nämä arvot ”voitaisiin jakaa äärettömän moneen luokkaan” ja

- ”murtoviiva” olisi sileä käyrä,

- jota sanotaan satunnaismuuttujan X **tiheysfunktioksi**.

Tiheysfunktion arvot, **todennäköisyystiheydet**,

vastaavat kuten edellä suhteellisen havaintotiheyden arvoja ”äärettömän kapeissa pisteiksi kutistuneissa luokissa”.



Edelleen

todennäköisyys, että X:n arvo osuu satunnaiskokeessa välille [a, b]

$P(a \leq X \leq b)$ = ”pinta-ala”,

mutta nyt se ei enää ole otoksesta saatava arvio kuten edellä, vaan (perusjoukossa mallinnetusta) X:n jakaumasta ”laskettu” todennäköisyys.

Jatkuvan satunnaismuuttujan X todennäköisyysjakauma siis

määritellään

tiheysfunktionsa f avulla, jolla on ominaisuudet

- f on ei-negatiivinen ($f(x) \geq 0$) kaikilla $x \in \mathbf{R}$ ja

- pinta-ala f:n kuvaajan ja x-akselin välissä on 1:n suuruinen.

Nyt jatkuvien jakaumien käsittelyssä on jäljellä ”enää” kaksi käytännön ongelmaa:

- Miten saadaan selville tutkittavan muuttujan tiheysfunktio poimimatta äärettömän suurta otosta? Jos tästä selvitään, niin

- miten voidaan laskea tarkasti ja kohtuullisella vaivalla tiheysfunktion kuvaajan ja x-akselin välisiä pinta-aloja?

- Ensimmäiseen ongelmaan etsitään ratkaisua tarkastelemalla otoksesta saatua jakaumaa. Histogrammaan yritetään sovittaa funktion f kuvaaja, joka myötäilee sen muotoa ja jonka alle jää ykkösen suuruinen pinta-ala.

- Tällöin tulos ei tietenkään ole täysin varmasti juuri ”oikea” malli, vaan se on väline, jonka avulla voidaan edes jossain määrin ennakoida sattuman käyttäytymistä satunnaiskokeissa.

- Otoksesta tiivistettävän informaation käyttäytymisen ennakoinnissa apuun tulevat(?) yleiset sattuman käyttäytymistä säätelevät lainalaisuudet. Tähän palataan myöhemmin otantajakaumien käsittelyssä.

- Jos tiheysfunktio f on saatu selville, pinta-alat voidaan laskea integraalilaskennan avulla:

$P(a \leq X \leq b)$ = ”pinta-ala”

$$= \int_a^b f(x)dx$$

Tällä kurssilla ei kuitenkaan tarvita integroimistaitoja.

Jatkuvista jakaumista tarvitaan myöhemmin vain **normaalijakauma**

ja siitä johdettuja jakaumia. Integroimisongelma kierretään tässä kertymäfunktion avulla.

Ennen normaalijakaumaan siirtymistä palataan vielä aikaisempaan esimerkkiin, jossa voidaan hyvin perustein olettaa todennäköisyystiheys vakioksi:

Esim. jatkoa) Linja-auton vuoroväli on 10 minuuttia.

Satunnaiskoe ε = ”Menet aikataulusta tietämättä pysäkille.” ja

satunnaismuuttuja X = odotusaika.

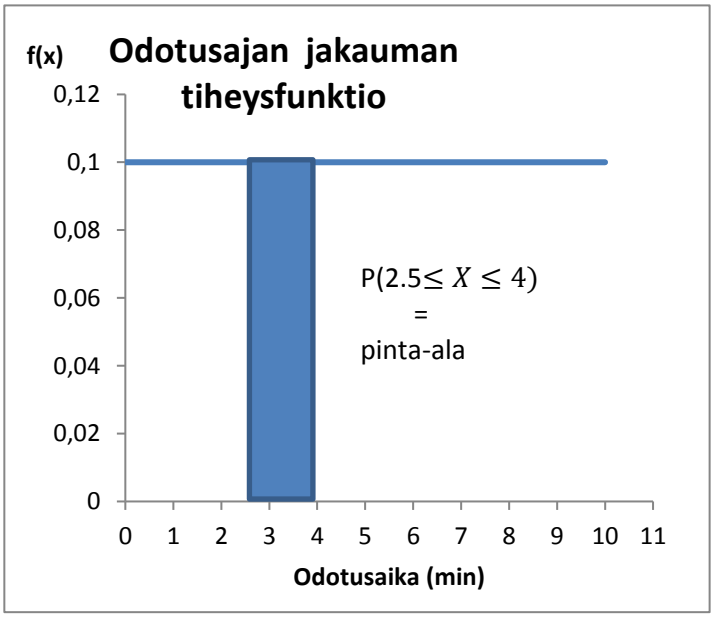
Odotusaika ”voi olla yhtä hyvin” mikä tahansa arvo välillä [0, 10] min eli todennäköisyystiheys on vakio, ja tiheysfunktio saa vakioarvon.

”Kokonaispinta-ala” = 1,

joten on oltava

f(x) = 1/10, kun x∈ [0, 10] ja

0 muualla



Koko pinta-ala on $10 \cdot 0.1 = 1$ ja

$P(2.5 \leq X \leq 4) = (4 - 2.5) \cdot 0.1 = 0.15$,

mikä laskettiin myös edellä toisella tavalla.

(Saman saa tietysti myös integroimalla.)

Normaalijakauma

on todennäköisyysjakaumista ehdottomasti tärkein.

Normaalijakauma on jatkuva jakauma, ja se määritellään tiheysfunktionsa avulla:

Satunnaismuuttujaa X sanotaan ***normaalisti jakautuneeksi parametrein μ ja σ^2,***

mistä käytetään merkintää $\boldsymbol{X \sim N(\mu, \sigma^2)}$,

jos sen tiheysfunktio f on

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} \quad \textbf{kaikilla } \boldsymbol{x \in R}$$

(e on Neperin luku ≈ 2.718… ja $\pi \approx 3.141\ldots$)

Voidaan osoittaa(?), että

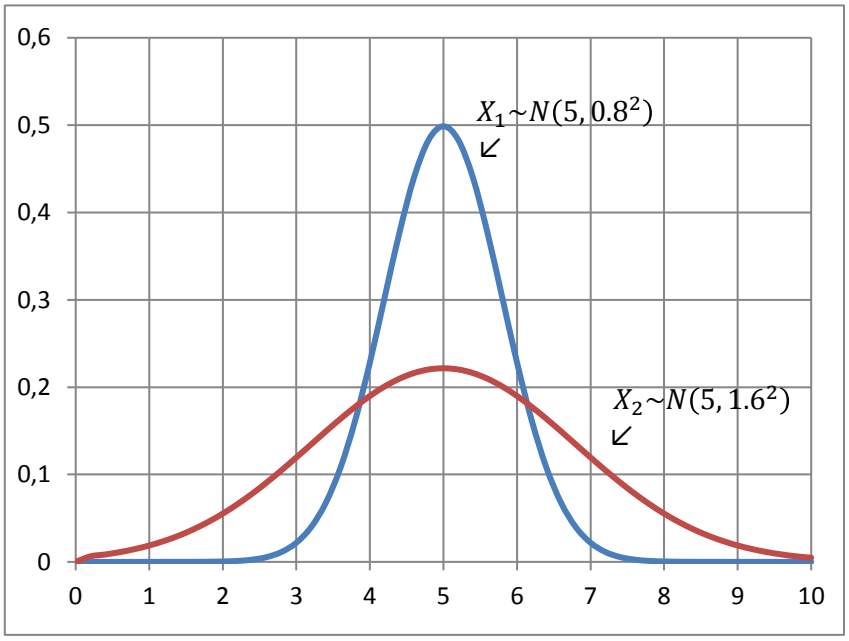
odotusarvo EX = μ ja **varianssi *Var(X) = σ²*** ja **hajonta *DX = σ*** (>0).

<u>Huom.</u> Koulukirjoissa esitetään (vastoin yleistä tapaa) yleensä

toisena parametrina hajonta σ. Tämä hyvää tarkoittava yksinkertaistus aiheuttaisi kuitenkin muita merkintäongelmia eteenpäin mentäessä.

Kuviossa ovat normaalisten satunnaismuuttujien

$X_1 \sim N(5, 0.8^2)$ ja $X_2 \sim N(5, 1.6^2)$ tiheysfunktiot:



- Jakauman **huippu** on odotusarvon μ = 5 kohdalla ja

- jakauma on **symmetrinen** sen ympärillä.

- Toinen parametri varianssi σ^2 (ja hajonta σ) säätelee, kuinka ”laakea” kuvaaja on.

Normaalijakauman tiheysfunktion kuvaajaa sanotaan myös **Gaussin käyräksi.**

Satunnaiskokeen ε alkeistapausten ominaisuutta X kuvaavalla jatkuvalla satunnaismuuttujalla

todennäköisyys, että X:n arvo osuu välille [a, b] on

P(a$\leq$X$\leq$ b) = ”pinta-ala”

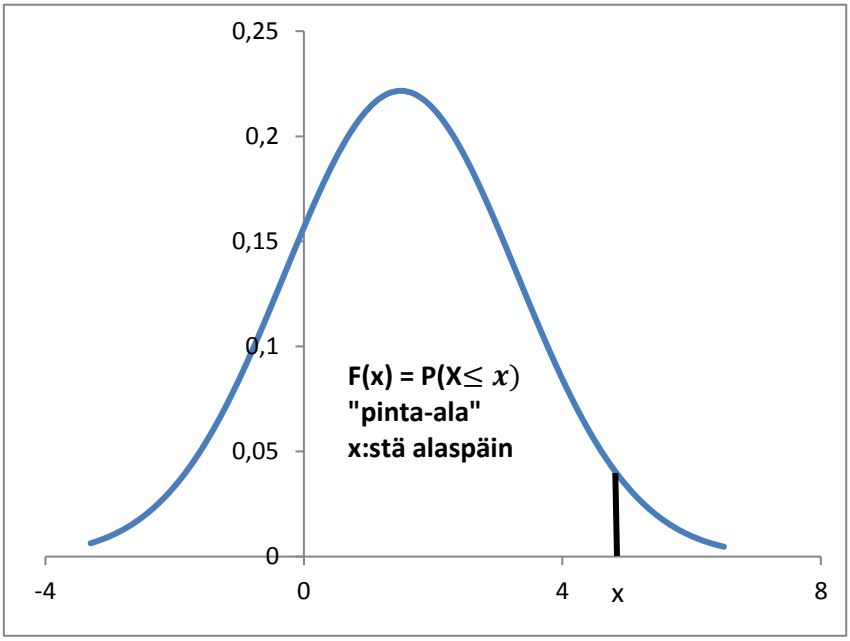
$$= \int_{\mathbf{a}}^{\mathbf{b}} \mathbf{f(x)dx}$$

Välien todennäköisyydet voidaan laskea myös kertymäfunktion avulla:

Jos f on satunnaismuuttujan X tiheysfunktio, on kertymäfunktio F sen integraalifunktio

F(x) = P(X≤ x) = ”pinta-ala”

$$= \int_{-\infty}^{x} \mathbf{f(t)dt}$$



- Joka tapauksessa todennäköisyyksiä laskettaessa on integroitava tiheysfunktio.

- Integraalifunktiota ei kuitenkaan pystytä esittämään ”suljetussa muodossa” (täsmällistä lauseketta ei saada selville).

- Ongelma voidaan ratkaista approksimoimalla normaalijakauman tiheysfunktiota ”sopivalla”(?) polynomifunktiolla, joka taas pystytään integroimaan helposti.

- Laskemisessa riittää, että käytössä on jokin (periaatteessa mikä tahansa) ”perusjakauma”. Sellaiseksi on valittu jakauma Z ∼ N(0,1), siis

normaalijakauma, jossa odotusarvo $\mu = 0$ ja varianssi $\sigma^2 = 1$.

- Kaikkiin muihin normaalisiin satunnaismuuttujiin liittyvien todennäköisyyksien laskeminen voidaan palauttaa tähän

standardoituun (normeerattuun) normaalimuuttujaan Z

perustuvaksi laskuksi.

Standardoidun normaalimuuttujan ***Z ∼ N(0,1)*** tiheysfunktiosta käytetään merkintää $\boldsymbol{\varphi}$ ja kertymäfunktiosta merkintää $\boldsymbol{\Phi}$.

Kun Z ∼ N(0,1), tiheysfunktion lauseke yksinkertaistuu vähän:

$$\boldsymbol{\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \quad \textbf{kaikilla } \mathbf{x} \in \mathbf{R}$$

- Kaikkia tarvittavia x:n arvoja vastaavat kertymäfunktion likimääräiset (mutta hyvin tarkat) arvot Φ(x) saadaan

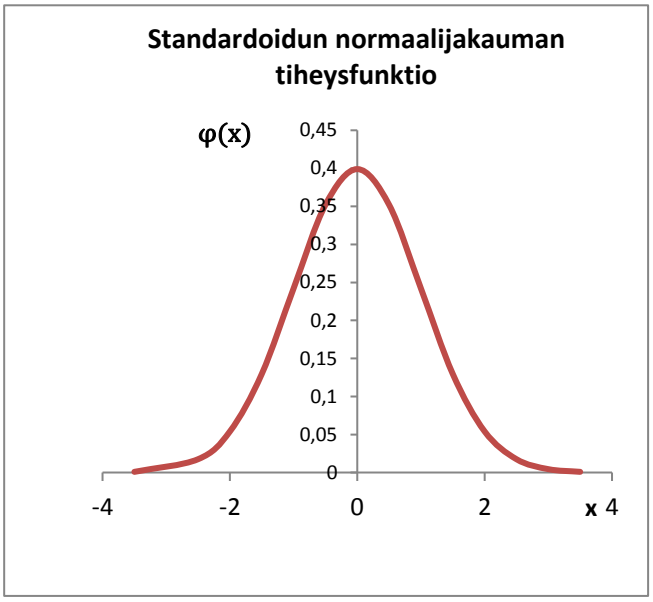
- approksimoimalla tiheysfunktiota φ ”sopivalla” polynomifunktiolla ja

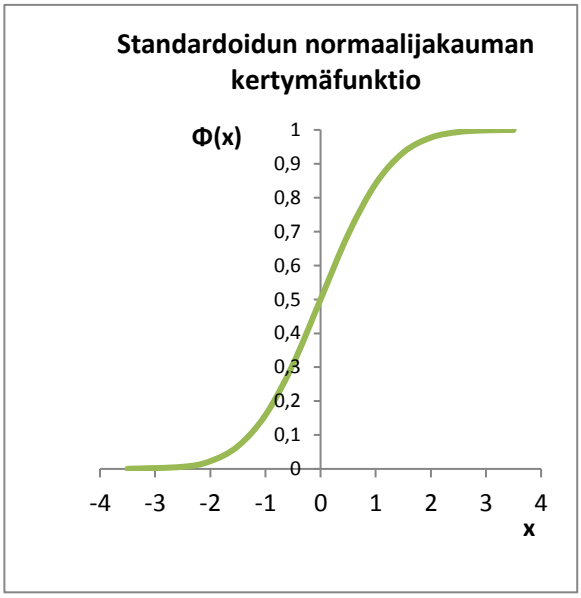
- määräämällä sen integraalifunktiosta likiarvot Φ(x):lle.

Yksityiskohtiin ei puututa tässä.

ϕ:n (ja myös muiden normaalisten muuttujien) kertymäfunktion arvot saadaan helposti Excelin avulla. Jos tällaista apuvälinettä ei ole käytössä, arvot saadaan taulukoista.

Standardoidun normaalimuuttujan tiheysfunktion ja kertymäfunktion kuvaajat ovat:





Seuraavalla sivulla ovat kertymäfunktion ϕ arvoja taulukoituina.

ϕ:n arvot kasvavat rivien suunnassa x:n kasvaessa aina 0,01:n verran.

Esim. ϕ(1,96) = 0.9750 on 1,9 alkavalla rivillä 0,06:n alla.

	x	0.00	0,01	0,02	0,03	0,04	0,05	0,06 ↓	0,07	0,08	0,09
	0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
	0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
	0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
	0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
	0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
	0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
	0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
	0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
	0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
	0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
	1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
	1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
	1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
	1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
	1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
	1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
	1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
	1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
	1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
→	1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
	2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
	2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
	2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
	2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
	2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
	2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
	2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
	2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
	2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
	2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
	3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
	3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
	3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
	3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
	3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Arvot on laskettu Excelin avulla:

Formulas → More Functions → Statistical → NORM.S.DIST

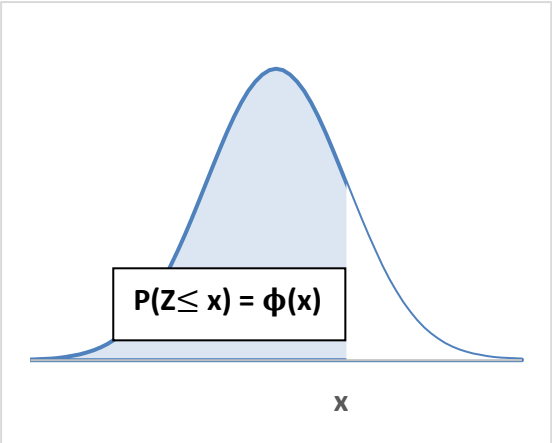
Tästä taulukossa saa vain todennäköisyydet

$P(Z \leq x) = \phi(x)$

= ”pinta-ala” tiheysfunktion

ja x-akselin välissä,

kun $x \geq 0$.



Normaalijakauma on jatkuva todennäköisyysjakauma, joten jokaisen yksittäisen arvon x todennäköisyys on $P(Z = x) = 0$.

Arvoja on ylinumeroituvasti ääretön määrä, jolloin yksittäisen täsmällisen arvon x (kuten vaikkapa π = 3.1414… … …) realisoitumisen mahdollisuus on ”häviävän pieni”.

Silloin on myös **$P(Z < x) = P(Z \leq x) = \phi(x)$.**

Oleellisia ovat tapahtumat, joissa Z saa satunnaiskokeessa arvon, joka kuuluu jollekin **välille**.

Kaikki tällaiset todennäköisyydet voidaan laskea kertymäfunktion avulla todennäköisyyslaskennan yleisten sääntöjen ja normaalijakauman erityisominaisuuksien avulla:

- Standardoitu normaalimuuttuja Z on jo sinänsä täysin korvaamaton todennäköisyyksien laskemisessa ”laskukoneena”, johon sovelluksissa tarvittavien normaalisten muuttujien laskut ”siirretään”.

- Jos kuitenkin kaivataan jotain tarinaa tähänkin, niin tilanne voisi olla:

Esim. Tarkkuusvaakaan vertaamalla saatujen havaintojen perusteella

tiedetään, että lääkeannoksen punnituksessa mittausvirheen Z

- keskimääräinen suuruus $\mu \approx 0$ mg ja hajonta $\sigma \approx 1$ mg ja

- jakauma on likimain normaalinen eli $Z \sim N(0,1)$.

Suoraan taulukosta saadaan (Piirrä myös kuvat ”pinta-aloista”.)

a) P(Mittausvirhe on korkeintaan +1.25 mg)

$= P(Z \leq 1.25 \text{ mg}) = \phi(1.25)$

$= 0.8944$

b) $P(Z < 0.75) = \phi(0.75) = 0.7734$

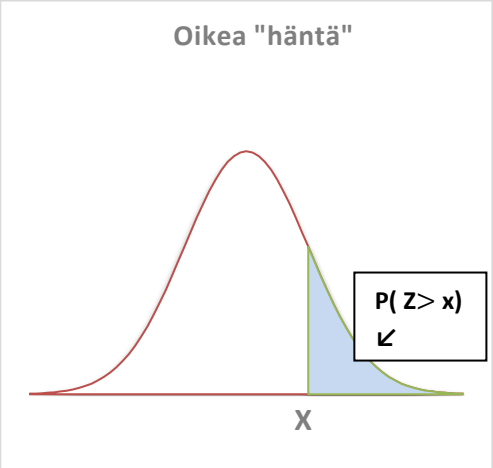
↑

”=” puuttuu, mutta jatkuvassa jakaumassa
se ei vaikuta todennäköisyyteen.

Jos kertymäfunktiota ei voi suoraan käyttää, lausekkeita muokkaamalla se onnistuu:

Jakauman ”oikea häntä”

Komplementtisäännön perusteella



$P(Z > x) = 1 - P(Z \leq x)$

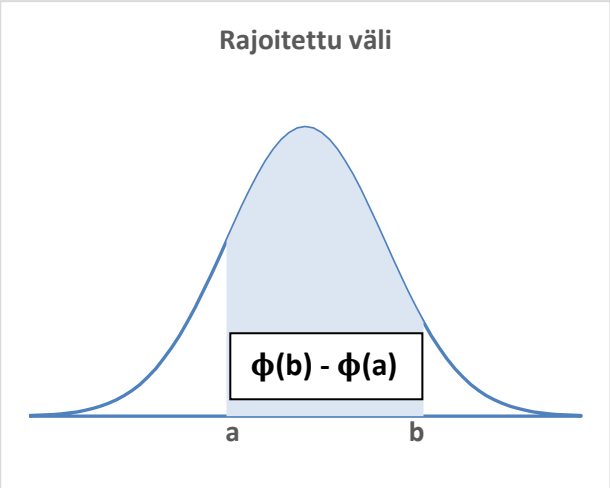
$= 1 - \phi(x)$

Esim.

$P(Z > 1.96) = \mathbf{1 - P(Z \leq 1.96)} = 1 - \phi(1.96)$

$= 1 - 0.9750 = 0.0025$

Rajoitettu väli



Kertymäfunktion yleisen ominaisuuden perusteella

$$P(a < Z \leq b) = \phi(b) - \phi(a)$$

↑ ↑

”=” voi olla tai olla olematta rajoissa (jatkuvalla jakaumalla).

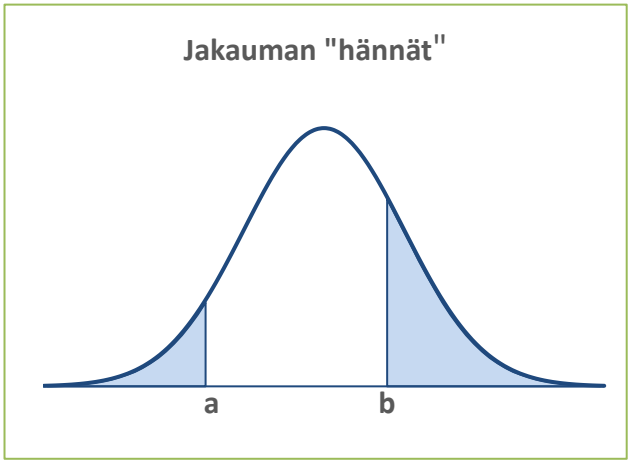
Esim.

$P(0.23 < Z \leq 1.66) = \phi(1.66) - \phi(0.23)$

$= 0.9515 - 0.5910$

$= 0.3605$

Jakauman molemmat ”hännät”



jakautuvat kahdeksi tapaukseksi yhteenlaskusäännön avulla:

$$\mathbf{P(Z \leq a \text{ tai } Z > b) = P(Z \leq a) + P(Z > b)}$$

Esim.

$P(Z \leq 0.45$ **tai** $Z > 2.33) = P(Z \leq 0.45) +$ $P(Z > 2.33)$

$= P(Z \leq 0.45) + 1 - P(Z \leq 2.33) = \phi(0.45) + 1 - \phi(2.33)$

$= 0.6736 + 1 - 0.9901$

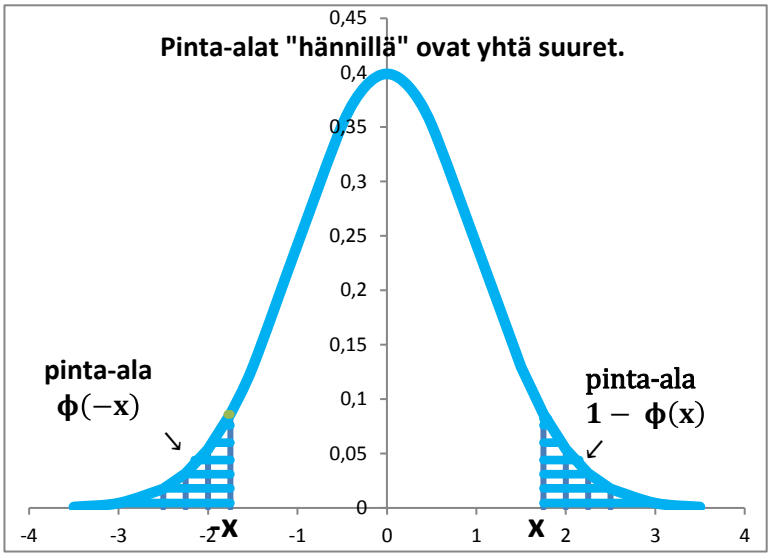
$= 0.6835$

Edellä kaikki rajat ovat olleet positiivisia. Esim. Excelin avulla saadaan ϕ:n arvot minkä tahansa arvon x kohdalla.

Jos joudutaan käyttämään taulukoituja ϕ:n arvoja, voidaan käyttää jakauman symmetrisyyteen perustuvaa tulosta:

$$\boldsymbol{\phi(-x) = 1 - \phi(x)}$$

Todistus sivuutetaan, mutta tiheysfunktion symmetrisyyden perusteella nähdään:



Tuloksen avulla laskut voidaan siirtää positiiviselle puolelle.

Esim. P(Annoksesta puuttuu yli 1mg.)

$P(Z < -1.00) = \boldsymbol{\phi(-1.00) = 1 - \phi(1.00)}$

$= 1 - 0.8413$

$= 0.1587$

P(Mittausvirhe on korkeintaan 2 mg)

$P(-2.00 < Z \leq 2.00) = \phi(2.00) - \boldsymbol{\phi(-2.00)}$

$= \phi(2.00) - \boldsymbol{(1 - \phi(2.00))}$

$= 2\,\phi(2.00) - 1 = 2 \cdot 0.9772 - 1 = 0.9544$

Tärkeitä tilastotieteen sovelluksia varten (testauksessa kriittiset rajat, luottamuskertoimet, joista myöhemmin) tarkastelu on käännettävä myös toisin päin.

Käytännössä tämä käänteisfunktio-ongelma ratkeaa helposti

- suoraan (esim.) Excelin avulla tai

- käyttämällä kertymäfunktion ϕ taulukkoa toisin päin:

Esim. $Z \sim N(0,1)$ ja on määrättävä sellainen raja, että a) $P(Z \leq x) = 0.99$.

$0.99 = P(Z \leq x) = \phi(x)$

- Excelillä saadaan $x = 2.326348 \approx 2.33$

(Formulas → More Functions → Statistical → NORM.S.INV)

- Taulukosta näkyy, että

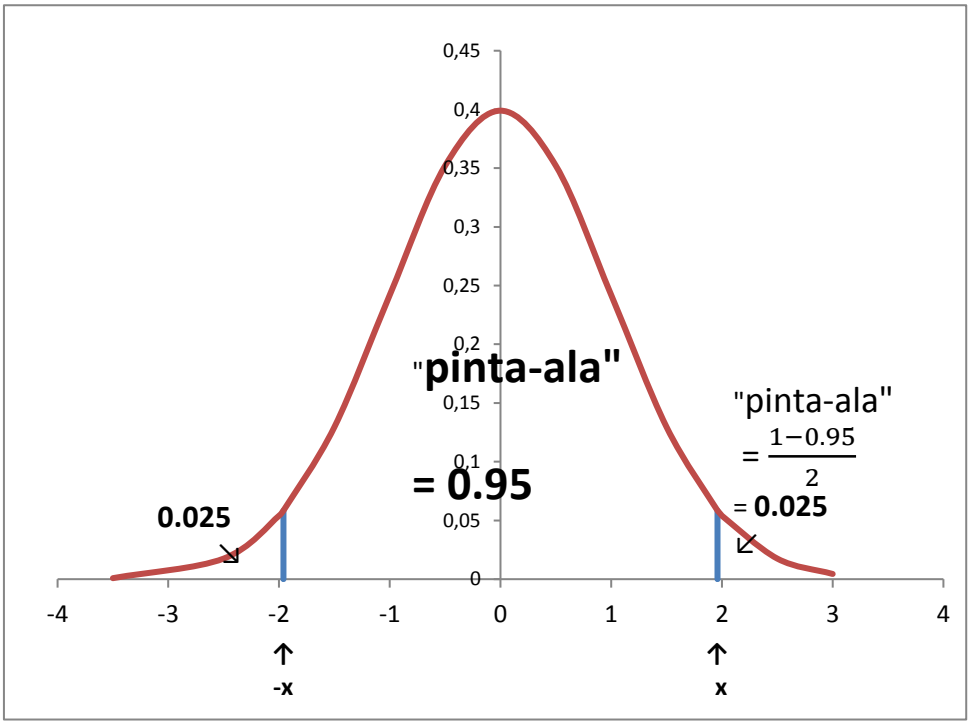
$\phi(2.33) = 0.9901.$

ϕ on aidosti kasvava, joten $x \approx 2.33$.

b) Mikä on x:n arvo, jos $P(|Z| \leq x) = P(-x < Z \leq x) = 0.95$?

Tiheysfunktion ja x-akselin välisten ”pinta-alojen” avulla nähdään:

-x:n ja x:n välissä on pinta-alaa 0.95 verran.



”Pinta-ala x:n alapuolella” eli

$\phi(x) = 0.95 + 0.025 = 0.975.$

Excelistä tai taulukosta näkyy toisaalta, että

$\phi(1.96) = 0.975$, joten $x=1.96$.

Myös suoraan laskemalla saadaan

$$\mathbf{0.95} = P(|Z| \leq x) = P(-x \leq Z \leq x)$$

$$= \phi(x) - \phi(-x) = \phi(x) - (1 - \phi(x))$$

$$\mathbf{= 2\,\phi(x) - 1,}$$

josta ratkaistaan yhtälö $\phi(x)$:n suhteen ja saadaan

$$\phi(x) = \frac{0.95+1}{2} = 0.975,$$

jne. kuten edellä.

- Standardoitu normaalimuuttuja Z, jossa ”keskimäärin odotettavissa oleva arvo” on $\mu = 0$ ja ”keskimäärin odotettavissa oleva arvojen vaihtelu” hajonta $\sigma = 1$ sopii tietenkin harvoin malliksi sovelluksissa.

- Jakauma on tärkeä ”laskukone”, johon mihinkä tahansa normaalijakaumaan liittyvät laskut voidaan siirtää:

Voidaan osoittaa:

Jos satunnaismuuttuja $\boldsymbol{X \sim N(\mu,\sigma^2)}$,

niin standardoitu (normeerattu) muuttuja $\boldsymbol{Z = \frac{X-\mu}{\sigma} \sim N(0,1)}$.

Silloin

$$\mathbf{P_X(X \leq a)} = \mathbf{P_Z}\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = P\left(Z \leq \frac{a-\mu}{\sigma}\right) = \boldsymbol{\phi}\left(\frac{\mathbf{a-\mu}}{\boldsymbol{\sigma}}\right)$$

↑ ↑

Merkinnöissä $\mathbf{P_X}$ ja $\mathbf{P_Z}$ alaviittaukset X ja Z viittaavat siihen, että todennäköisyyden laskeminen siirretään X:ään ”liittyvästä” (X:n indusoimasta) todennäköisyyskentästä Z:aan ”liittyvään” todennäköisyyskenttään. Tähän liittyvä mittateoreettinen pohdiskelu voidaan sivuuttaa. Myös laskuissa jätetään jatkossa nämä tässä turhat merkinnät pois.

Esim. Suklaatehtaassa on todettu erittäin useiden mittausten perusteella saadun aineiston perusteella, että tuotettavien suklaalevyjen paino

- X on likimain normaalisti jakautunut ja aineistosta laskettu
- keskipaino $\bar{x} \approx 100$ g ja painon keskihajonta $s \approx 4$ g.

- Tämän perusteella arvioidaan (estimoidaan), että X:n

keskimäärin odotettavissa oleva paino $\mu \approx \bar{x} \approx 100$ g ja
keskimäärin odotettavissa oleva painon hajonta $\sigma \approx s \approx 4$ g.

Siis havaintoaineiston perusteella mallina on, että tuotannosta umpimähkään poimittavan levyn paino $X \sim N(100\text{ g}, (4\text{ g})^2)$.

a) Kuinka todennäköistä on, että levy on vähintään 5 g alipainoinen?

epäyhtälön käsittelyä, joka johtaa standardoituun muuttujaan

$$P(X \leq 95) = P\left(\frac{X-100}{4} \leq \frac{95-100}{4}\right) = P(Z \leq -1.25)$$

= 1 - 0.8944 = 0.1056.

Siis on odotettavissa, että noin 10 % levyistä on yli 5 g alipainoisia.

b) $P(X > 110) = P(\frac{X-100}{4} > \frac{110-100}{4}) = P(Z > 2.50)$

$= 1 - P(Z \leq 2.50) = 1 - 0.9938 = 0.0062$

Siis keskimäärin vain noin 6 kertaa 1000:sta on odotettavissa näin painava suklaalevy.

c) $P(96 < X < 104) = P(\frac{96-100}{4} < \frac{X-100}{4} < \frac{104-100}{4})$

$= P(-1.00 < Z < 1.00) = \phi(1.00) - \phi(-1.00) = = \phi(1.00) - (1 - \phi(1.00))$

$= 2\,\phi(1.00) - 1 = 2 \cdot 0.8413 - 1 = 0.6826$

Siis (on odotettavissa, että) levyn paino on alle yhden hajontayksikön (tässä σ = 4 g) päässä keskipainosta 100 g noin 68 prosentissa levyistä.

Jos X $\sim$N(μ,σ^2), niin jakaumaan liittyvät todennäköisyydet saadaan edellä käytetyillä säännöillä:

1) **Standardointi**

Jos X $\sim$N(μ,σ^2), niin $\boldsymbol{Z = \frac{X-\mu}{\sigma} \sim N(0,1)}$.

Tätä tulosta käytetään ensimmäiseksi ja standardoidaan kaikki käsiteltävät epäyhtälöt. Tämän jälkeen kaikki laskut liittyvät muuttujaan

Z $\sim$ N(0,1).

2) Hajotetaan lausekkeet, joissa on enemmän kuin yksi epäyhtälö.

Rajoitettu väli: $\boldsymbol{P(a < Z \leq b) = \phi(b) - \phi(a)}$

Jakauman hännät: $\boldsymbol{P(Z \leq a \text{ tai } Z > b) = P(Z \leq a) + P(Z > b)}$

3) **Komplementtisäännön** avulla käännetään ”väärin päin” olevat epäyhtälöt: $\boldsymbol{P(Z > a) = 1 - P(Z \leq a)}$

4) Kertymäfunktion määritelmän mukaan on

P(Z≤ a) = ϕ(a).

5) Jos kertymäfunktion ϕ argumentti on negatiivinen, käytetään sääntöä

ϕ(-x) = 1 - ϕ(x).

(Tämä vaihe voidaan ohittaa, jos arvot katsotaan Excelistä.)

6) ϕ:n arvot saadaan Excelistä tai taulukosta.

- Kaikkia näitä sääntöjä ei tietenkään aina tarvita ja niitä voi käyttää usein eri järjestyksessä. Joskus voi myös löytää oikoteitä erityisesti jakauman symmetrisyyden perusteella.

- Jos oikoteitä ei löydy, tällä tavalla pääsee aina oikeaan ratkaisuun.

- Exceliä käytettäessä voidaan myös ohittaa standardointi (tässä) ja satunnaismuuttujan $X \sim N(\mu,\sigma^2)$ kertymäfunktion F arvot saadaan suoraan.

(**Formulas → More Functions → Statistical → NORM.DIST**)

Myös silloin, kun X∼N(μ,σ^2) joudutaan laskemaan ”toisin päin”:

Esim. On havaittu, että tuotteen kestoikä X ∼ N(μ,σ^2).

Tehdas pystyy säätelemään kestoikään vaikuttavan kemikaalin K määrän avulla keskimääräisen kestoiän μ suuruutta niin, että kuitenkin hajonta $\sigma \approx 200$ h pysyy samana.

Kuinka suureksi μ on asetettava, jotta 99 % tuotteista kestää korkeintaan 3000 h?

Vaatimus tarkoittaa, että yksittäisen tuotteen kestoiän osalta on

$$0.99 = P(X \leq 3000) = P\left(\frac{X-\mu}{100} \leq \frac{3000-\mu}{100}\right) = P\left(Z \leq \frac{3000-\mu}{100}\right)$$

$$= \phi\left(\frac{3000-\mu}{100}\right).$$

Taulukosta saadaan toisaalta $\phi(2.33) = 0.9991 \approx 0.99$.

Silloin (ϕ on aidosti kasvava) on

$$\frac{3000-\mu}{100} \approx 2.33,$$

josta saadaan $\mu \approx 3000 - 2.33 \cdot 100 = 2767$ h.

Esim. Elintarvikeannoksen lisäaineen E määrä $X \sim N(200\text{ mg}, (15\text{ mg})^2)$.

Määrää a niin, että umpimähkään valittavassa annoksessa 95 % todennäköisyydellä E:n määrä

poikkeaa keskimääräisestä arvosta 200 mg alle a:n verran.

Siis on oltava

$0.95 = P(200-a \leq X \leq 200+a)$

$$= P\left(\frac{200-a-200}{15} < \frac{X-200}{15} < \frac{200+a-200}{15}\right)$$

$$= P\left(\frac{-a}{15} < Z < \frac{a}{15}\right)$$

$$= \phi\left(\frac{a}{15}\right) - \phi\left(\frac{-a}{15}\right) = \phi\left(\frac{a}{15}\right) - (1 - \phi\left(\frac{a}{15}\right))$$

$$= 2\,\phi\left(\frac{a}{15}\right) - 1,$$

josta ratkaistaan samalla tavalla kuin aikaisemmin

$$\phi\left(\frac{a}{15}\right) = \frac{0.95+1}{2} = 0.975 = \phi(1.96) \quad (\leftarrow \text{taulukosta}).$$

Silloin $\frac{a}{15} = 1.96$ ja $a = 1.96 \cdot 15 = 29.4$

Huom. väli [200-a, 200+a] ei ole luottamusväli, kuten lukiossa on saatettu sitä virheellisesti nimittää.)

Todennäköisyyksien laskeminen on helppoa, kun tiedetään, mistä normaalijakaumasta ne lasketaan.

Paljon suurempi ongelma on, **miten empiirisessä sovelluksessa tunnistetaan** jakauman normaalisuus.

- Matemaattisen todistamisen ylivalta ei ulotu reaalimaailman ylle niin vahvasti, että vaikkapa edellisen esimerkin tilanteessa voitaisiin matemaattisen pitävästi todistaa suklaalevyn painon jakauman olevan normaalinen (tai jotain muuta).

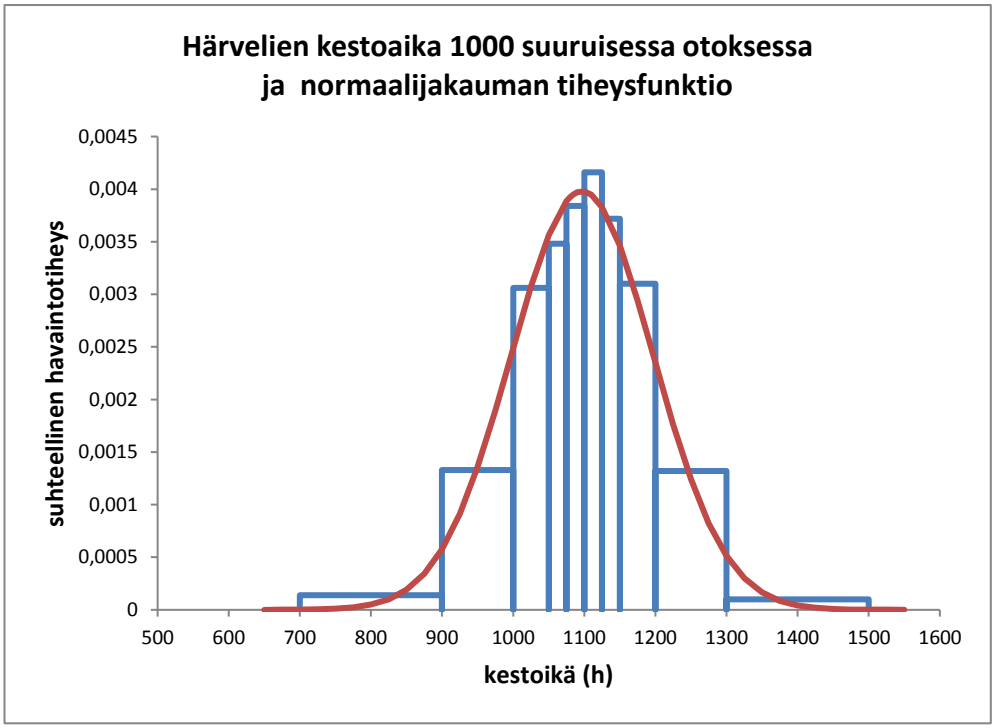
- Tunnistaminen perustuu empiiriseen havainnointiin, jossa matemaattisesta tilastotieteestä on kyllä paljon apua.

Käytännössä oikeille jäljille päästään, kun tutkitaan frekvenssijakaumaa, joka on tehty otoksesta mitatuista muuttujan arvoista.

Histogramman muodosta näkyy jo heti alustavasti, muistuttaako jakauma normaalijakaumaa.

Kuvioon voidaan myös sovittaa normaalijakauman tiheysfunktiota vastaava käyrä (odotusarvona $\mu \approx \bar{x}$ ja hajontana $\sigma \approx s$), jolloin vertaaminen on helpompaa.

Esim. Aikaisemmin tutkittiin 1000 suuruisesta otoksesta mitattujen härvelien kestoiän jakaumaa.



Vaikka yhteensopivuus on erittäin hyvä, se **ei todista** normaalisuutta.

Kuitenkin on ilmeisesti **järkevää käyttää mallina** normaalijakaumaa tähän tilanteeseen liittyvien todennäköisyyksien laskemisessa.

Muuttujan normaalisuutta voidaan myös testata, ja tässä matemaattinen tilastotiede tulee vahvasti apuun.

Tällaisia testejä on useita ja mm. (otoksesta havaitun) jakauman vinoutta ja huipukkuutta kuvaavien tunnuslukujen g_1 ja g_2 pohjalta voidaan konstruoida tällaiset testit.

Tälläkään tavalla ei voida todistaa normaalisuutta, mutta päätelmien luotettavuutta (riskiä) voidaan todennäköisyyden avulla ”mitata”.

Usein jo pelkästään aikaisemman kokemuksen perusteella voidaan arvioida kohtuullisen luotettavasti, että tutkittava muuttuja on normaalinen.

Normaalijakauma on tilastotieteessä tärkeä ainakin seuraavista syistä:

1) Empiirinen tosiasia

Useat empiirisiä ilmiöitä kuvaavat satunnaismuuttujat ovat (jostain syystä(?)) normaalisti jakautuneita.

Huom. Kuitenkaan joissain lukion oppikirjoissakin esitetty väite

”Muuttuja, joka riippuu useista tekijöistä, on normaalinen.” **ei ole totta**.

Jos näin olisi, jokseenkin kaikki empiiriset muuttujat olisivat normaalisia.

Tällainen ajatus perustuu ilmeisesti sekä teorian että sovellusten kannalta äärimmäisen tärkeän keskeisen raja-arvolauseen väärin ymmärrykseen. Aiheeseen palataan myöhemmin.

2) ”Pyrkimys normaalisuuteen” informaatiota tiivistettäessä

Kun tunnetaan riittävän hyvin sattuman käyttäytymisen lainalaisuudet, voidaan (otantatilanteessa ja kokeellisessa tutkimuksessa) ”tuottaa” satunnaismuuttujan jakauman normaalisuus.

Monien otoksesta (joskus tulevaisuudessa ehkä) laske**ttavien** tunnuslukujen (mm. otoskeskiarvo $\bar{x}$, suhteellinen osuus $\hat{p}$) otoksesta saatavien arvojen määräytymisen mallina käy (ainakin likimain) normaalijakauma. Tästä myöhemmin.

Oikealla todennäköisyysotannan sääntöjen mukaan tehdyllä otannalla (tai kokeellisessa tutkimuksessa koejärjestelyllä) saadaan sattuman käyttäytymistä säätelevät lainalaisuudet toimimaan.

3) Muita tärkeitä todennäköisyysjakaumia (t-, χ^2-, F- jakauma, …)

voidaan (matemaattisella päättelyllä) johtaa normaalijakaumasta.

Binomi- ja hypergeometrisen jakauman normaaliapproksimaatio

Esim. (jatkoa) Valtavan menestyksen saaneelle ohjelmasarjalle, jossa jokaisessa esiintyi kansanedustaja, aiotaan tehdä 20 jatko-osaa.

- Yleisön pyynnöstä seuraava esiintyvä kansanedustaja arvotaan edellisen viikon ohjelmassa kaikista 200 kansanedustajasta eli ”otos” poimitaan palauttaen.

- Etukäteen pohditaan, montakohan opposition edustajaa (ohjelmaa suunnitellessa 88:sta opposition kannattajasta) tulee valituksi otokseen.

Siis minkälainen on satunnaismuuttujan

X = opposition kannattajien lukumäärä otoksessa jakauma.

Järjestettävät arpajaiset ovat

- 20-kertainen **toistokoe**, jossa

- toistot ovat **riippumattomia**,

- P(Valituksi tulee opposition edustaja.) = 88/200 = **0.44 joka toistolla** ja

- X kuvaa **lukumäärää**,

joten **X**$\sim$ **Bin(20, 0.44)** ja

P(X = k) = $\binom{20}{\mathrm{k}}\ 0.44^{\mathrm{k}}(1-0.44)^{20-\mathrm{k}}$.

Vaikka n = 20 tässä, kaikki todennäköisyydet voidaan laskea kohtuullisella vaivalla frekvenssifunktiosta ”käsin” ja vielä helpommin (esim.) Excelillä.

Toinenkin vaihtoehto kuitenkin on:

Odotusarvo μ = 20·0.44 = **8.8**,

varianssi σ^2 = 20·0.44·(1-0.44) = 4.928 ja hajonta $\sigma \approx$ **2.22**.

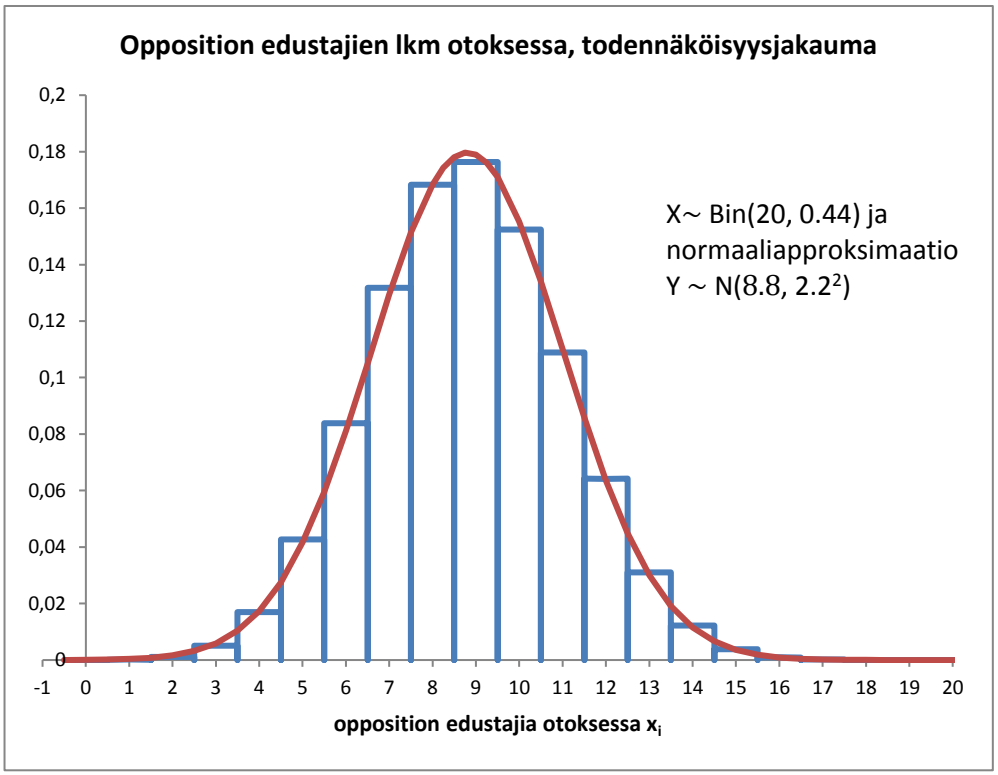
Taulukossa ”kokeillaan”, minkälaisia arvoja normaalisen satunnaismuuttujan

Y $\sim$ N($\mathbf{8.8}$, **2.22²**)

tiheysfunktio saa ja niitä verrataan X:n arvojen todennäköisyyksiin.

Kuviossa on binomijakauma ja siihen sovitettu normaaliapproksimaatio.

x_i	bin(20,0.44) $P(X = x_i)$	$N(8.8,2.22^2)$ tiheysfunktio			
0	0,0000	0,0001	11	0,1089	0,1100
1	0,0001	0,0004	12	0,0642	0,0636
2	0,0011	0,0016	13	0,0310	0,0300
3	0,0051	0,0059	14	0,0122	0,0116
4	0,0170	0,0174	15	0,0038	0,0036
5	0,0427	0,0415	16	0,0009	0,0009
6	0,0839	0,0811	17	0,0002	0,0002
7	0,1318	0,1294	18	0,0000	0,0000
8	0,1683	0,1684	19	0,0000	0,0000
9	0,1763	0,1790	20	0,0000	0,0000
10	0,1524	0,1553			



Erot ovat hyvin pienet.

Todennäköisyydet ovat ”hännillä” lähes nollia (4:n desimaalia!).

Jakauma keskittyy tiiviisti odotusarvon 8.8 ympärille.

Kuvio on piirretty niin, että

- X:n arvot 0, 1, …, 20 ovat pylväiden keskikohdissa ja

- kantojen pituus = 1 ja korkeus arvon x_i todennäköisyys $P(X = x_i)$, jolloin

- pylvään pinta-ala = $P(X = x_i)$.

- Toisaalta pinta-ala tiheysfunktion ja x-akselin välissä samalla kohtaa on likimain yhtä suuri.

Näin ollen normaalijakaumaa voidaan käyttää ”laskukoneena” binomijakauman todennäköisyyksien laskemisessa:

Oletetaan, että $\mathbf{X \sim Bin(n, p)}$.

Jos n on ”riittävän” suuri ja p on ”riittävän” lähellä arvoa 0.5, niin satunnaismuuttujan

$$\mathbf{Y \sim N(np, np(1-p))}$$

$$\uparrow \qquad \uparrow$$

$\mu = EX$ $\quad \sigma^2 = Var(X)$

jakauma ”vastaa” X:n jakaumaa niin hyvin, että

$P(x_1 \leq X \leq x_2) \approx P(x_1 - ½ \leq Y \leq x_2 + ½)$ kaikilla X:n arvoilla $x_1 \leq x_2$.

↑ ↑

Binomijakaumasta ”siirrytään” normaalijakaumaan.

Tulos on erikoistapaus tärkeästä keskeisestä raja-arvolauseesta(?). Todistaminen sivuutetaan.

- Kun parametri p = 0.5, binomijakauma on täysin symmetrinen ja approksimaatio (symmetrisellä) normaalijakaumalla toimii parhaiten.

- Jos p:n arvo ei ole 0.5, suuri toistojen määrä korvaa tätä puutetta:

- Vaikka X:n jakaumalla olisikin pitkä ”häntä” vasemmalle ($p > 0.5$) tai oikealle ($p < 0.5$), ovat ”hännällä olevien X:n arvojen todennäköisyydet merkityksettömän pieniä.
- Jakauman huipun (odotusarvon np) tienoilla jakauma on riittävän symmetrinen.

Käyttökelpoinen sääntö sille, milloin n on ”riittävän” suuri ja p ”riittävän” lähellä arvoa 0.5 on:

$np > 5$ ja $n(1-p) > 5$

- Binomijakauman odotusarvo μ= EX = np ja varianssi σ^2 = Var(X) = np(1-p) ja ne ovat luonnollisesti ”apumuuttujana” olevan normaalisen Y:n parametreina.

- Diskreetin satunnaismuuttujan X$\sim$ Bin(n, p) arvojen 0, 1, 2,…, n todennäköisyydet ovat suurempia kuin nolla.

X:ää vastaava Y $\sim$ N(np, np(1-p)) taas on jatkuva ja jokaisen yksittäisen täsmällisen arvon todennäköisyys = 0.

Jatkuvassa approksimaatiossa ajatellaan, että täsmällisiä X:n arvoja 0, 1, 2,…, n vastaavat välit, joissa luvut pyöristyvät näihin arvoihin.

Esim. P(X=8) $\approx$ P(7.5$<$Y$\leq$8.5). Vertaa edellisen kuvion pinta-aloihin.

Tätä X:n arvoista vähennettävää ja lisättävää arvoa ½ sanotaan **jatkuvuuskorjaukseksi**.

Esim. (jatkoa) Kuin todennäköistä on, että esiintymään pääsee yli 5, mutta korkeintaan 12 opposition edustajaa?

X$\sim$ Bin(20, 0.44) ja

P(5 $<$ X $\leq$12) = $\binom{20}{6} 0.44^6(1-0.44)^{20-6}$ + … +$\binom{20}{12} 0.44^{12}(1-0.44)^{20-12}$

= 0.0839 + … + 0.0642 (edellä Excel-taulukossa)

= 0.886

Tässä

$np = 20 \cdot 0.44 = 8.8 > 5$ ja $n(1-p) = 20 \cdot 0.56 = 11.2 > 5$

ja jakauma on riittävän symmetrinen.

”Apumuuttujana” on

$Y \sim N(20 \cdot 0.44, 20 \cdot 0.44 \cdot 0.56) = N(8.8, 4.928) = N(8.8, 2.220^2)$

$P(5 < X \leq 12) = P(6 \leq X \leq 12)$ ← 6, 7, 8 ,9 ,10, 11 ja 12 “mukana”

$\approx P(6-0.5 \leq Y \leq 12+0.5) = P(5.5 \leq Y \leq 12.5)$

$= P(\frac{5.5-8.8}{2.22} \leq \frac{Y-8.8}{2.22} \leq \frac{12.5-8.8}{2.22})$

$= P(-1.49 \leq Z \leq 1.67) = \phi(1.67) - \phi(-1.49) = \phi(1.67) - (1 - \phi(1.49))$

$= 0.9525 - 1 + 0.9319$

$= 0.8844$

Edellä binomijakaumasta saatiin melkein sama tulos 0.886.

Hypergeometrista jakaumaa voidaan samoin edellytyksin kuin edellä approksimoida normaalijakauman avulla samalla tavalla:

Jos **X**$\sim$ **Hyperg(N, K, n)** ja

n on ”riittävän” suuri ja p = $\frac{K}{N}$ on ”riittävän” lähellä arvoa 0.5

ja otoskoko n ei ole ”liian”(?) suuri perusjoukon kokoon N verrattuna,

niin satunnaismuuttujan

$$\mathbf{Y \sim N\left(np,\ np(1-p)\frac{N-n}{N-1}\right)}$$

jakauma vastaa X:n jakaumaa niin hyvin, että

$P(x_1 \leq \mathbf{X} \leq x_2) \approx P(x_1 - ½ \leq \mathbf{Y} \leq x_2 + ½)$ kaikilla X:n arvoilla $x_1 \leq x_2$.

Tässä voidaan käyttää samaa ”nyrkkisääntöä” kuin edellä riittävän symmetrian takaamiseen:

On oltava $\mathbf{np > 5}$ **ja** $\mathbf{n(1-p) > 5}$.

Esim. (jatkoa) Sattui käymään niin, että kansanedustaja Hannu Hanhi arvottiin kaksi kertaa ohjelmaan.

Tällaisen vääryyden välttämiseksi vaadittiin, että seuraavalla 20 ohjelman tuotantokaudella esiintyjät arvotaan palauttamatta.

Nyt ”otokseen” osuvien opposition edustajien määrä

X∼Hyperg(200, 88, 20)

ja todennäköisyydet voidaan laskea frekvenssifunktiosta

$$\mathbf{P(X=k)} = \frac{\binom{\mathbf{88}}{\mathbf{k}} \cdot \binom{\mathbf{112}}{\mathbf{20-k}}}{\binom{\mathbf{200}}{\mathbf{20}}}$$, k = 0, 1, 2, …, 20.

Myös normaaliapproksimaatio käy, kun

np = $20 \cdot \frac{88}{200} = 20 \cdot 0.44 = 8.8 > 5$ ja n(1-p) = $20 \cdot 0.56 = 11.2 > 5$ ja

20 henkilön valitseminen ei tyhjennä paljon 200 hengen perusjoukkoa.

Odotusarvo $\boldsymbol{\mu}$ **= EX = np =** **8.8** ja

hajonta $\boldsymbol{\sigma}$ **= DX =** $\sqrt{\mathbf{np(1-p)} \cdot \frac{\mathbf{N-n}}{\mathbf{N-1}}} = \sqrt{20 \cdot 0.44 \cdot 0.56 \cdot \frac{200-20}{200-1}} \approx$ **2.111**.

”Apumuuttujana” on silloin **Y** ~ **N(8.8, 2.111^2)** ja aivan samalla tavalla kuin edellä esim.

P(5< X ≤12) = P(6 ≤ X ≤12) ← 6, 7, 8 ,9 ,10, 11 ja 12 “mukana”

≈ P(6-0.5 ≤ Y ≤12+0.5) = P(5.5 ≤ Y ≤12.5)

= P($\frac{5.5-8.8}{\mathbf{2.111}} \leq \frac{Y-8.8}{\mathbf{2.111}} \leq \frac{12.5-8.8}{\mathbf{2.111}}$)

vähän pienempi hajonta ↖ kuin otannassa palauttaen

= P(-1.56 ≤ Z ≤ 1.75) = ϕ(1.75) - ϕ(-1.56) = ϕ(1.75) – (1 - ϕ(1.56))

= 0.9599 – 1 + 0.9406

= 0.9005.

Taulukossa on (Excelistä saadut) pistetodennäköisyydet P(X=k) ja kertymäfunktion arvot F(k) = P(X≤ k).

Sieltä saadaan

P(5< X ≤12) = F(12) – F(5) = 0.9602 – 0.0563 = 0.9039

- ja ero on tässäkin hyvin pieni approksimaationa saatuun tulokseen verrattuna.

k	P(X=k)	F(k)	k	P(X=k)	F(k)
0	0,0000	0,0000	11	0,1090	0,8997
1	0,0001	0,0001	**12**	0,0605	**0,9602**
2	0,0007	0,0008	13	0,0270	0,9872
3	0,0037	0,0044	14	0,0095	0,9967
4	0,0138	0,0182	15	0,0026	0,9993
5	0,0381	**0,0563**	16	0,0006	0,9999
6	0,0807	0,1370	17	0,0001	1,0000
7	0,1337	0,2706	18	0,0000	1,0000
8	0,1759	0,4466	19	0,0000	1,0000
9	0,1858	0,6324	20	0,0000	1,0000
10	0,1583	0,7907			

- Tässä todennäköisyys on vähän suurempi binomijakaumalla tai sen approksimaatiolla laskettuun tulokseen verrattuna.

Otannassa palauttamatta realisaatiot ovat suuremmalla varmuudella lähellä ”keskimäärin odotettavissa olevaa” arvoa EX (= 8.8).

- Sekä binomi- että hypergeometrisen jakauman todennäköisyydet saadaan selville helposti Excelin avulla, eikä niiden laskemiseen tarvita (nykyään) välttämättä normaaliapproksimaatiota.

- Normaaliapproksimaation suuri merkitys on siinä, että sen avulla saadaan normaalijakauman hyvät ominaisuudet avuksi tällaisen otantatilanteen tutkimiseen.

Tämä selviää kohta **suhteellisen osuuden otantajakauman** käsittelyssä ja sen kautta estimointi- ja testiteorian menetelmissä.

Normaalijakauman ominaisuuksia

Otantateoriassa useiden tärkeiden tunnuslukujen **otantajakauman** selvittäminen perustuu normaalijakaumaan ja sen hyviin ominaisuuksiin.

Esim. Tavarahissin kokonaispaino Y koostuu hissin omasta 300 kg:n painosta ja

siinä kuljetettavan kuorman painosta X, jonka suuruutta on havaittu (likimain) kuvaavan satunnaismuuttujan

$X \sim N(500\ \text{kg}, (100\text{kg})^2)$.

On intuitiivisesti luonnollista, että kokonaispainon Y = X + 300 kg jakaumassa:

- **Odotusarvo** Keskimäärin odotettavissa oleva kokonaispaino

EY = 500 kg + 300 kg

= keskimääräinen kuorman paino EX + hissin oma paino

- Kokonaispainon vaihtelu johtuu pelkästään kuorman painon vaihtelusta, joten on oltava

varianssi Var(Y) = Var(X) = $(100\ \text{kg})^2$ ja **hajonta** DX = 100 kg.

- Jokaiseen kuorman painoon X lisätään vakio hissin oma paino,

jolloin X:n jakauma vain ”siirtyy muuttuja-akselilla” 100 kg eteenpäin ja on järkevää olettaa, että myös Y:n **jakauma on normaalinen**.

Voidaankin osoittaa:

Jos satunnaismuuttuja $\boldsymbol{X \sim N(\mu, \sigma^2)}$ ja $\boldsymbol{a}$ ***vakio***, niin

$$\boldsymbol{Y = X + a \sim N(\mu + a, \sigma^2)}.$$

Yleisemminkin: Myös muille kuin normaalisille satunnaismuuttujille

E(X + a) = EX + a ja ***Var(X+a) = Var(X)***.

Esim. Tuotteen valmistamisaika minuutteina on $X \sim N(17\ \text{min}, (2\ \text{min})^2)$.

Silloin tuotteen valmistamisaika sekunteina on Y = 60 (s/min) ·X.

Tässäkin on varsin luonnollista, että Y:n jakaumassa on:

- **Odotusarvo**: Keskimäärin odotettavissa oleva valmistamisaika sekunteina

EY = (60 s/min) · 17 min = 1020 s

- Odotettavissa oleva valmistamisaikojen pituuden hajonta DX = 2 min, mikä on sekunteina (60 s/min) · 2 min = 120 s. Aikayksikön vaihtaminen ei muuta valmistamisen keston vaihtelua mitenkään. Se vain esitetään eri tavalla. Siis ”on oltava”

hajonta DY =D**(60 s/min)**·X = **(60 s/min)**·DX = **60 (s/min)**· 2 min = 120 s ja

varianssi $\text{Var}(Y) = (DY)^2 = (\mathbf{60\ (s/min)} \cdot 2\ \text{min})^2 = (\mathbf{60\ (s/min)})^{\mathbf{2}} \cdot (2\ \text{min})^2$

- Arvojen ”esiintymistiheydet” minuutteina ja sekunteina ovat ilmeisesti aivan samat ja jakauman **normaalisuus säilyy muunnoksessa.**

Voidaan osoittaa, että nämä empiirisesti järkevät ominaisuudet ovat myös matemaattisesti voimassa:

Jos satunnaismuuttuja $X \sim N(\mu, \sigma^2)$ ja $b \neq 0$ on ***vakio***, niin

$$Y = bX \sim N(b\mu, b^2\sigma^2) = N(b\mu, (b\sigma)^2).$$

Yleisemminkin: Myös muille kuin normaalisille satunnaismuuttujille

$$E(bX) = bEX \qquad \text{ja} \qquad Var(bX) = b^2 Var(X).$$

Seuraavassa normaalisten muuttujien yhteenlaskua koskevassa tuloksessa on satunnaismuuttujien oltava toisistaan riippumattomia.

Satunnaismuuttujat X ja Y ovat toisistaan ***riippumattomia*** (merkitään $X \perp Y$), jos

$$P(X \in A \text{ ja } Y \in B) = P(X \in A) \cdot P(Y \in B),$$

kaikille ”oleellisille” tapahtumille A ja B.

- Tässä ei voida ryhtyä pohtimaan, mitä ”oleellinen” määritelmässä täsmälleen matemaattisesti (mittateoreettisesti) tarkoittaa.

- Käytännössä X:n ja Y:n riippumattomuudella tarkoitetaan:

Minkä arvon X tulee saamaan satunnaiskokeessa $\mathcal{E}$, ei ”vaikuta” siihen, minkä arvon Y saa, ja päinvastoin.

Esim. Kunnan asukkaista aiotaan poimia **otos palauttaen**.

Satunnaismuuttuja

X_1 = 1. otokseen osuvan vuositulo, X_2 = 2. otokseen osuvan vuositulo, jne.

Edelliset otoksesta saatavat arvot eivät vaikuta millään tavalla seuraaviin ja X_1, X_2, … ovat toisistaan riippumattomia.

Otantatilanteessa halutaan ”rehellisillä arpajaisilla” nimenomaan **saada aikaan riippumattomuus** perusjoukosta otokseen tulevassa ”informaatiovirrassa”.

- **Käytännön näkökulmasta** tämä ilmeisesti takaa, että otos on ”edustava osa” perusjoukosta.

- **Matemaattisesti** tämän ansiosta voidaan selvittää, minkälaisia ovat tärkeiden tunnuslukujen **otantajakaumat** eli

- mitä arvoja ja millä todennäköisyyksillä -
otoksesta laskettavat tunnusluvut (ja muut otossuuret) **tulevat saamaan**,

jos otos (joskus tulevaisuudessa) poimitaan.

Riippumattomien normaalisten muuttujien yhteenlaskua koskeva tulos on ratkaisevan tärkeä tällaisessa **informaation** spekulatiivisessa **yhdistämisessä**:

Esim. Tuotettavien suklaalevyjen paino $X \sim N(100\ g, (4g)^2)$.

Levyt pakataan (väljiin) koteloihin joiden paino $Y \sim N(5\ g, (1\ g)^2)$.

Suklaa tulee pakkauskoneeseen toista liukuhihnaa pitkin ja pakkaukset toista, joten voitaneen olettaa $X \perp Y$.

Minkälainen on kokonaispainon $T = X + Y$ jakauma?

Odotusarvo: On luonnollista, että

pakatun levyn keskipaino = suklaan keskipaino + pakkauksen keskipaino.

Ilmeisesti siis on $ET = \mathbf{E(X + Y) = EX + EY}$.

Varianssi: - On luontevaa, että pakatun levyn painon T ”vaihtelun suuruus” on sitä suurempi, mitä suurempia ovat suklaan painon X ja pakkauksen painon Y ”vaihtelun suuruus”.

- Kuitenkin tässä on oleellista, että X ja Y ovat riippumattomia.

- Lisäksi ei ole mitenkään itsestään selvää,

- miten ”kokonaisvaihtelu” määräytyy ”osavaihteluista”.

- Vaihtelun suuruuden mittaamiseen käytettävien tunnuslukujen valinta on kuitenkin tarkoitushakuista:

Varianssia käytetään vaihtelun suuruuden mittarina (mm.), jotta tällaisessa tilanteessa on

(X + Y):n ”vaihtelu” = X:n ”vaihtelu” + Y:n ”vaihtelu”, siis

Var(X+Y) = Var(X) + Var(Y), jos X⊥Y. (Todistus sivuutetaan tässä.)

Jakauman muoto: Vaikka X ja Y olisivat samoin jakautuneita ja riippumattomiakin, ei X + Y:n jakauma yleensä ole saman tyyppinen. Normaalijakaumalla on kuitenkin tämä hyvä ominaisuus.

Todistukset sivuutetaan, mutta voidaan osoittaa:

Jos X $\sim N(\mu_1, \sigma_1^2)$ ***ja Y*** $\sim N(\mu_2, \sigma_2^2)$ ***ja X⊥Y, niin***

$X + Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

↗ ↑ ↖

3) 1) 2)

1) Odotusarvolla on tämä ominaisuus **kaikilla** satunnaismuuttujilla.

2) Varianssilla ominaisuus on kaikilla **riippumattomilla** muuttujilla.

3) Jakauman tyypin säilyminen (tässä normaalisena) on **erityisominaisuus**.

Näiden sääntöjen avulla pystytään **ennakoimaan, mitä otoksessa tulee tapahtumaan**, kun otos joskus tullaan poimimaan:

Esim. (jatkoa) Tehtaan tuottamien suklaalevyjen (perus-)joukossa

levyn paino $X \sim N(100\text{ g}, (4\text{ g})^2)$.

Tuotannosta aiotaan poimia 20 suuruinen otos.

Sattuma tulee määräämään jokaisen poimittavan levyn painon perusjoukossa vallitsevan normaalisen jakauman mukaisesti.

Silloin

1. otokseen osuvan levyn paino $X_1 \sim N(100\text{ g}, (4\text{ g})^2)$,

2. otokseen osuvan levyn paino $X_2 \sim N(100\text{ g}, (4\text{ g})^2)$,

…

20. otokseen osuvan levyn paino $X_{20} \sim N(100\text{ g}, (4\text{ g})^2)$.

Levyt arvotaan otokseen, ja painoja kuvaavat satunnaismuuttujat ovat riippumattomia.

Otokseen tulevan suklaan kokonaismäärän $T = X_1 + X_2 + \ldots + X_{20}$ jakauma on edellisen mukaan

$$\swarrow 20\ \text{kpl} \searrow \qquad\qquad \swarrow 20\ \text{kpl} \searrow$$

$$T = X_1 + X_2 + \ldots + X_{20} \sim N(100\ g + 100\ g + \ldots + 100\ g,\ (4\ g)^2 + (4\ g)^2 + \ldots + (4g)^2)$$

$$\uparrow \qquad\quad \uparrow \qquad\qquad\qquad \uparrow$$

$$3) \qquad 1) \qquad\qquad 2)$$

$$= N(20 \cdot 100\ g,\ 20 \cdot (4\ g)^2) = N(2000\ g,\ 320\ g^2) = N(2000\ g,\ (17.9\ g)^2).$$

Huom.: Ei 20^2 ↗

Siis tässä yhdistettiin spekulatiivisesti ennen otoksen poimimista otokseen osuvaa informaatiota ja saatiin selville otokseen osuvan suklaan kokonaismäärän otantajakauma. Tuloksen mukaan

- Keskimäärin otokseen on odotettavissa yhteensä 2000 g suklaata.

- Hajonnan avulla mitattuna ”sattuman pelivara” määrän vaihtelemiseen 2000 g:n ympärillä on keskimäärin noin 18 g.

- Tilanteeseen liittyvien todennäköisyyksien laskemisessa voidaan käyttää mallina normaalijakaumaa.

Esimerkiksi, kuinka todennäköistä on, että määrä on alle 1 950 g:

$$P(T < 1950) = P\left(\frac{T-2000}{17.9} < \frac{1950-2000}{17.9}\right) = P(Z < -2.79) = \phi(-2.79)$$

$$= 1 - \phi(2.79) = 1 - 0.9974 = 0.0026.$$

- Siis vain alle 3 kertaa tuhannesta on odotettavissa näin vähän suklaata 20 levyn joukkoon,

jos painon jakauma perusjoukossa on oletettu $X \sim N(100\ g, (4\ g)^2)$.

- Jos tällainen tulos saataisiin todella poimitusta otoksesta, silloin ilmeisesti voidaan epäillä, että keskipaino ei olekaan väitetty 100 g!

Tätä **testaus**kysymystä tutkitaan myöhemmin tarkemmin.

Kokonaispainon T jakaumasta saadaan myös helposti otokseen osuvien levyjen **keskipainon $\bar{x}$ otantajakauma**. Tästä jatketaan vähän myöhemmin.

3 Otannasta

3.1 Otanta ja tilastollinen päättely

Jos tutkitaan koko perusjoukko, tehdään **kokonaistutkimus.**

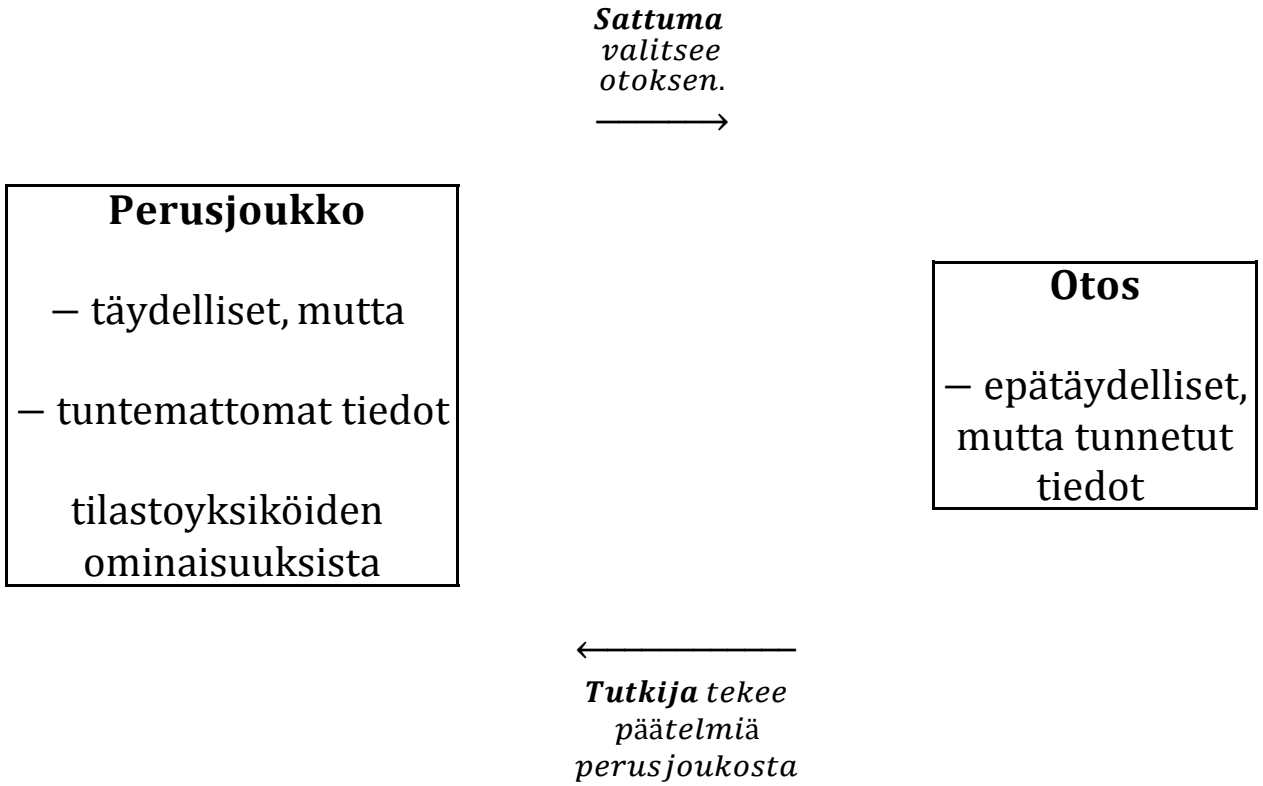
Tällainen on esimerkiksi poliittisen kannan jakauman selvittämiseksi pidettävät vaalit.

Kun perusjoukko on laaja, ei voida (kannata) tutkia kaikkia tilastoyksiköitä.

Silloin ainoa vaihtoehto on **otantatutkimus**, jonka vaiheet ovat

- otoksen poimiminen perusjoukosta ja

- perusjoukkoa koskevien päätelmien tekeminen otoksen informaation perusteella.



Otantatutkimuksessa **Sattumaa** tietoisesti hyväksi käyttäen valitaan tutkittavaksi perusjoukon osajoukko.

- **Kokonaistutkimuksesta** saatava tieto kuvaa periaatteessa täysin tarkasti, minkälaisessa tilassa perusjoukko on.

- **Otantatutkimuksen** tulokset ovat vain arvioita perusjoukosta.

Usein otantatutkimus on kuitenkin käytännössä ainoa ja parempikin vaihtoehto:

- Jos perusjoukko on ääretön, kokonaistutkimusta ei voi tehdä.

- Kokonaistutkimus vaatii usein liikaa resursseja.

- Jos mittauksia tehtäessä tilastoyksikkö tuhoutuu, tutkittavan määrän on oltava pieni.

- Kun perusjoukko on suuri, resurssit voidaan käyttää monipuolisemmin. Tiedon laatu korvaa määrän.

Eduista huolimatta vastattavaksi jää:

- Otos on yleensä hyvin suppea osa perusjoukosta.

Esimerkiksi (vain/jopa) noin 2000 suuruinen otos kaikista suomalaisista kuluttajista markkinatutkimuksessa.

Voidaanko otoksesta havaittava ”tilanne” ylipäänsä yleistää perusjoukkoon?

Ainakin tämä asettaa hyvin suuret laatuvaatimukset havaintojen valinnalle perusjoukosta. Otantateoria tutkii näitä kysymyksiä.

- Jos voidaan, kuinka ***tarkkoja*** ja ***luotettavia*** nämä yleistykset ovat?

Tilastollisen päättelyn avulla vastataan tällaisiin

- perusjoukkoa koskeviin kysymyksiin

- otoksesta tiivistetyn informaation avulla.

Sattuma ”valitsee” otoksen sisällön ja vaikuttaa sitä kautta perusjoukon ominaisuuksista tehtäviin päätelmiin.

Kuitenkin tilastollisen päättelyssä on tavoitteena, että päätelmien

- ***luotettavuuden*** ↔ kuinka suurella varmuudella (kääntäen riskillä) arviot ovat tosia (vääriä)
- ja ***tarkkuuden*** ↔ ”virhemarginaali”

suuruus voidaan **esittää lukuina**.

Tätä varten on hallittava lainalaisuudet, jotka ohjaavat sattuman käyttäytymistä otantatilanteessa.

”Hyviä” otantamenetelmiä on useita, mutta …

Oikea mielikuva: ”**rehelliset arpajaiset**”

Tällöin

- otos on ***edustava osa*** perusjoukosta, ”**perusjoukko pienoiskoossa**” ja

- tulokset voidaan **yleistää** perusjoukkoon.

Yleistysten **tarkkuus** ja **luotettavuus** voidaan (usein) selvittää ja esittää numeerisesti.

Esim. Poimittavasta otoksesta saatava

puolueen X kannattajien, hyödykkeen Y käyttäjien, alkoholin mainoskiellon kannattajien yms.

- suhteellinen osuus ”pyörii” oikean arvon tuntumassa.

Lisäksi halutaan tietää ja voidaan selvittää

- kuinka suuri on (esim. mainoskiellon kannattajien) suhteellisesta osuudesta tehdyn arvion **virhemarginaali** (esim.) **95 % varmuudella**? (95 %:n luottamusväli)

Otantateoria tutkii, mitkä lainalaisuudet säätelevät informaatiovirtaa, jonka sattuma arpoo otokseen.

Yksinkertaisin otantamenetelmä on **yksinkertainen satunnaisotanta** (YSO):

YSO:ssa kaikilla perusjoukon tilastoyksiköillä on yhtä suuri todennäköisyys tulla valituksi otokseen.

Periaatteessa:

- Jokaista perusjoukon tilastoyksikköä vastaava ”nimilappu pannaan hattuun” ja sieltä poimitaan umpimähkään otos. Tämä on hankalaa ja

käytännössä:

- perusjoukon tilastoyksiköistä tehdään **kehikko** eli luettelo, jossa kaikki tilastoyksiköt numeroidaan 1:stä alkaen.

- Otokseen tulevat tilastoyksiköt arvotaan kehikosta **satunnaislukujen** avulla. Tietokoneen avulla tehtävät (tai valmiiksi taulukkoihin arvotut) satunnaisluvut **simuloivat** eli jäljittelevät lappujen arpomista ”hatusta”.

YSO on yleensä osana myös muissa otantamenetelmissä.

Oleellista on, että hallitaan sattuman käyttäytymistä otantatilanteessa säätelevä ”mekanismi”. Silloin menetelmä on **todennäköisyysotantaa**.

Tämä takaa ***edustavuuden***, ja arvioiden ***tarkkuus*** ja ***luotettavuus*** voidaan selvittää.

Jos todennäköisyysotannan sääntöjä ei noudateta, poimittua perusjoukon osajoukkoa sanotaan **näytteeksi**, ja siitä saadut tulokset voivat olla pahasti harhaisia.

”Edustava” informaatio voi syntyä myös toisella tavalla:

Kokeellinen tutkimus

Kokeellisessa tutkimuksessa tutkitaan jonkin ilmiön lainalaisuuksia.

Esim. Miten kehiteltävä lääke vaikuttaa verenpaineeseen?

Miten hyödykkeeseen liittyvä koehenkilölle aiheutettu ärsyke vaikuttaa hänen suhtautumiseensa tuotteeseen?

Tutkija asettaa tutkimustilanteen ja kontrolloi sitä aktiivisesti.

Koeyksiköt joutuvat jonkin **käsittelyn** kohteeksi ja niiden **reaktio** mitataan.

Päämäärä on sama kuin otannassa:

Halutaan **edustavaa tietoa** ilmiöstä.

Tätä varten tutkijalla ovat välineinä ***satunnaistaminen*** ja ***toistaminen***.

Esim. Uuden verenpainelääkkeen vaikutusta tutkittaessa

Koehenkilöt arvotaan ***koe-*** ja ***vertailuryhmään***.

(Ei esim. naiset toiseen ja miehet toiseen ryhmään)

Tällä (***satunnaistamisella***) pyritään **systemaattisten virheiden** eliminoimiseen.

Toistamalla koe ”riittävän monelle” koehenkilölle pienennetään **satunnaisten virheiden** vaikutusta.

Vaikka kokeellisessa tutkimuksessa informaatio syntyy eri tavalla kuin otannassa, koejärjestelyillä pyritään saamaan ”**edustavaa tietoa**” ilmiöstä.

Tällöin voidaan käyttää samoja analyysimenetelmiä kuin otantatutkimuksessa.

Seuraavassa käsitellään ”mekanismia”, joka säätelee yksinkertaisessa satunnaisotannassa

- minkälaisia arvoja otoksesta laskett**avat** tunnusluvut (otoskeskiarvo $\bar{x}$, jonkin ominaisuuden suhteellinen osuus $\hat{p}$ jne.) tulevat saamaan ja millä todennäköisyyksillä,

kun otokseen (joskus tulevaisuudessa) osuvaa informaatiota tiivistetään.

3.2 Otantajakaumista

Otantatutkimus voidaan ajatella pelinä, jossa ovat vastakkain

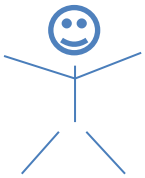
- **tutkija**, joka yrittää paljastaa perusjoukon ”salaisuudet”, ja

- **Sattuma**, joka ”tietää perusjoukosta kaiken”.

Sattuma säätelee informaatiovirtaa ***perusjoukosta otokseen*** päin.

Tutkija taas yrittää ”seurata sattuman jälkiä toiseen suuntaan” ja tekee päätelmiä ***otoksesta perusjoukkoon*** päin.

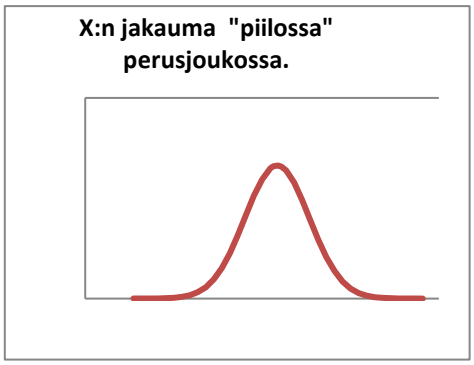
Perusjoukko



Sattuma on kunkku täällä.

”Sattuma tietää koko perusjoukon sisällön.”

μ = EX = todellinen keskiarvo
σ = DX = todellinen hajonta
….
jne. kaikista muuttujista



Sattuma *määrää otoksen sisällön.*

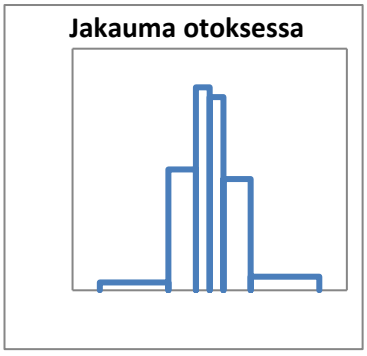
Otos ?

Tutkija

Sattuma arpoo tutkijalle muuttujan X jakaumasta arvot $x_1, x_2, \ldots, x_n$.

→ $\bar{x}$, s, …

”Rosoinen” kuva oikeasta jakaumasta:



Satunnaiskoe ε = ”**Perusjoukosta arvotaan tilastoyksikkö**”.

Satunnaismuuttuja

X = tutkittavan muuttujan arvo valittavassa tilastoyksikössä

välittää informaation perusjoukosta otokseen.

Arvot valikoituvat vallitsevan jakauman mukaisesti:

- sellaisia arvoja, joita on paljon perusjoukossa, tulee paljon myös otokseen ja

- perusjoukossa harvinaisia arvoja tulee otokseen vähän.

Näin tapahtuu suurella **varmuudella**. Toisaalta on pieni **riski**, että näin ei käykään. ”Vain ihan sattumalta” otokseen voi osua aineisto, joka poikkeaa paljon siitä, miten keskimäärin pitäisi olla.

Todennäköisyyslaskennan tehtävä on juuri sen selvittäminen,

kuinka suurella varmuudella ja kuinka suurella riskillä?

Sattuman suhde tutkijaan on kuitenkin asiallisen neutraali. Se ei pyri johtamaan harhaan, mutta ei myöskään pyri erityisesti suosimaan tutkijaa.

Se noudattaa todennäköisyyslaskennan teoriassa kirjattuja sääntöjään.

Jos otantamenetelmä ottaa huomioon nämä lainalaisuudet, voidaan luottaa, että **otoksessa nähdään ”perusjoukko pienoiskoossa”**.

Tutkija saa (tutkimusresurssiensa rajoitusten puitteissa) otokseen niukasti mutta edustavaa tietoa perusjoukosta. **Informaatio tiivistetään** erilaisiksi **tunnusluvuiksi** (otossuureiksi).

Tiivistettyä informaatiota hyväksi käyttäen **tilastollisen päättelyn** menetelmien avulla päätellään, millaisia perusjoukon tilastoyksiköiden ominaisuudet ”keskimäärin” ovat.

Jotta todella realisoituneen otoksen perusteella voidaan tehdä johtopäätöksiä taustalla olevasta perusjoukosta, on tunnettava ”mekanismi”, jonka mukaan sattuma ”yleisesti ottaen” määrää otoksen sisällön.

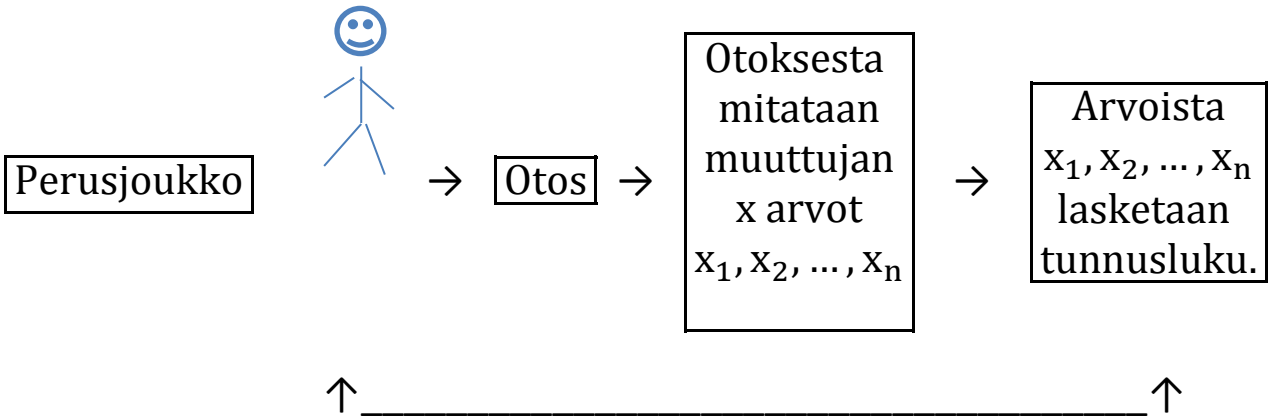
Otokseen osuvia muuttujan x arvoja ei tarkastella vain yksitellen, vaan

arvojen sisältämä informaatio tullaan tiivistämään tunnusluvuiksi, kun otos (joskus tulevaisuudessa) poimitaan.

- Sattuma määrää muuttujan arvot.

- Silloin sattuma määrää myös otoksesta laske**tavien** tunnuslukujen (otoskeskiarvo $\bar{x}$, keskihajonta s, suhteellinen osuus $\hat{p}$ jne.) arvot,

- joten **otoksesta laskettavat tunnusluvut ovat satunnaismuuttujia**.



Sattuma määrää otokseen tulevat tilastoyksiköt.

Sattuma määrää otoksesta laskettavan tunnusluvun arvon.

Näkökulma on siis tässä spekulatiivinen:

”Jos otos joskus tulevaisuudessa tullaan poimimaan, niin kuinkahan suuri esimerkiksi otoskeskiarvo $\bar{x}$ tulee olemaan?”

- Perusjoukossa vallitseva jakauma, otantatapa ja satunnaismuuttujien yleiset ominaisuudet määräävä tunnusluvun ***otantajakauman***.

Tässä käsitellään tarkemmin otoksesta laskettavan keskiarvon $\bar{X}$ ja suhteellisen osuuden $\hat{P}$ otantajakaumia:

Keskimääräisen suuruuden otantajakauma

Esim. (jatkoa) Tehtaan tuottamien suklaalevyjen (perus-)joukossa

levyn paino $X \sim N(100\text{ g}, (4\text{ g})^2)$.

Tuotannosta aiotaan poimia 20 suuruinen otos.

Sattuma tulee määräämään jokaisen poimittavan levyn painon perusjoukossa vallitsevan normaalisen jakauman mukaisesti.

Silloin

1. otokseen osuvan levyn paino $X_1 \sim N(100\text{ g}, (4\text{ g})^2)$,

2. otokseen osuvan levyn paino $X_2 \sim N(100\text{ g}, (4\text{ g})^2)$,

…

20. otokseen osuvan levyn paino $X_{20} \sim N(100\text{ g}, (4\text{ g})^2)$.

Levyt arvotaan otokseen YSO:aa käyttäen, jolloin painoja kuvaavia satunnaismuuttujia X_i, i = 1, 2,…, 20, voidaan pitää riippumattomia.

Otokseen tulevan suklaan kokonaismäärän $\mathbf{T = X_1 + X_2 + \ldots + X_{20}}$ jakauma on riippumattomien normaalisten muuttujien yhteenlaskuominaisuuden mukaan

↙ 20 kpl↘ ↙ 20 kpl↘

$T = X_1 + X_2 + \ldots + X_{20} \sim N(100\text{ g}+100\text{ g}+\ldots+100\text{ g},\ (4\text{ g})^2+(4\text{ g})^2+\ldots+(4\text{g})^2)$

↑ ↑ ↑

3) 1) 2)

$= \mathbf{N(20\cdot 100\text{ g},\ 20\cdot(4\text{ g})^2)}$

$(= N(2000\text{ g},\ (17.9\text{ g})^2).)$

Keskimääräinen levyn paino $\bar{X}$ on kokonaismäärä/otoskoko.

painon hajonta σ perusjoukossa

$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{20}}{20} = \frac{T}{20}$ ↓

$= \mathbf{\frac{1}{20}} \cdot T \sim N(\frac{1}{20} \cdot 20 \cdot 100\text{ g},\ \mathbf{(\frac{1}{20})^2} \cdot 20 \cdot (4\text{ g})^2) = N(100\text{ g},\ \frac{(4\text{ g})^2}{20})$

↗ ↑

keskipaino μ perusjoukossa otoskoko n

$= N(100\text{ g},\ 0.8\text{ g}^2) = N(100\text{ g},\ (0.8944\text{ g})^2)$

Tässä yhdistettiin spekulatiivisesti ennen otoksen poimimista otokseen (tulevaisuudessa) osuvaa informaatiota ja saatiin selville otokseen osuvien suklaalevyjen keskipainon $\bar{X}$ **otantajakauma**.

Tämän ”ohjesäännön” mukaan sattuma määrää otoksen sisällön:

- Sattuma ”pyrkii” asettamaan otokseen tulevan keskipainon $\bar{x}$ todellisen keskipainon μ = 100 g kohdalle.

- Hajonnan avulla mitattuna ”sattuman pelivara” otoskeskiarvon $\bar{x}$ suuruuden ”heiluttelemiseen” oikean keskipainon μ ympärillä on keskimäärin (vain) noin 0.9 g.

- Tilanteeseen liittyvien todennäköisyyksien laskemisessa voidaan käyttää mallina normaalijakaumaa.

Esimerkiksi, kuinka todennäköistä on, että keskipaino otoksessa tulee olemaan alle 1 g:n päässä oikeasta keskipainosta 100 g?

$$P(99 < \bar{X} < 101) = P\left(\frac{99-100}{0.8944} < \frac{\bar{X}-100}{0.8944} < \frac{101-100}{0.8944}\right)$$

$$= P(-1.12 < Z < 1.12) = \phi(1.12) - \phi(-1.12) = = \phi(1.12) – (1 - \phi(1.12))$$

$$= 2\,\phi(1.12) - 1 = 2 \cdot 0.8686 - 1 = 0.7374.$$

- Siis noin 74 % varmuudella on odotettavissa, että (vain) 20 levyn otoksessa keskipaino $\bar{X}$ tulee olemaan näin lähellä todellista keskipainoa.

Esimerkissä erikoistapauksena johdettu tulos on voimassa yleisesti:

Jos tutkittavan muuttujan jakauma on ***perusjoukossa*** $X \sim N(\mu, \sigma^2)$,

niin n:n suuruisesta **otoksesta** saatavan keskiarvon $\bar{X}$ jakauma on

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, kun

- otos poimitaan **palauttaen**

- tai perusjoukko on **ääretön** (käytännössä ”hyvin suuri”) tai sen kokoa ei tiedetä.

Jos otos poimitaan **palauttamatta** N:n suuruisesta perusjoukosta, niin jakauma on muuten sama, mutta mukaan tulee **äärellisen perusjoukon korjaustekijä** varianssia pienentämään:

Voidaan osoittaa, että silloin

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1})$.

Informaatiota otoksesta tiivistettäessä tunnuslukujen (otossuureiden) jakaumien muodostumisessa vaikuttaa voimakas ”pyrkimys normaalisuuteen”:

Vaikka edellisessä tilanteessa **jakauma ei olisikaan normaalinen** perusjoukossa,

otoskeskiarvon $\bar{X}$ otantajakaumaa koskevat edelliset tulokset ovat likimain voimassa, kun otoskoko n on ”suuri”.

Tällainen ”pyrkimys normaalisuuteen” esitetään

keskeisessä raja-arvolauseessa,

jonka oleellinen sisältö on:

Jos satunnaismuuttujat $X_1, X_2, \ldots, X_n$ ovat **samoin jakautuneita** ja **riippumattomia**, niin

satunnaismuuttujan $T = X_1 + X_2 + \ldots + X_n$ jakauma

”lähestyy” normaalijakaumaa, kun n kasvaa.

Tämä konvergenssi toteutuu ”hyvin yleisesti voimassa olevilla ehdoilla”.

(**Ei kuitenkaan**: ”Satunnaismuuttujan X jakauma riippuu useista osatekijöistä, joten se on normaalinen, kuten joissain lukion oppikirjoissa on väitetty.)

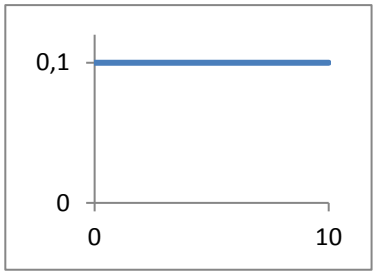
Esim. jatkoa) Linja-auton vuoroväli on 10 minuuttia.

Satunnaiskoe ε = ”Menet aikataulusta tietämättä pysäkille.” ja

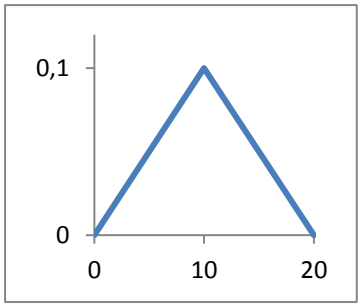
X_1 = odotusaika 1. kerralla, X_2 = odotusaika 2. kerralla, …

Kokonaisodotusajan $T = X_1 + X_2 + \ldots$ jakauma muuttuu kertojen kasvaessa:

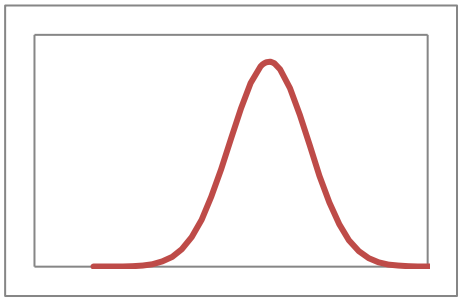
$X_1 \sim Tas(0,10)$ $\qquad\qquad$ $X_1 + X_2$



→



Hyvin nopeasti $T = X_1 + X_2 + \ldots$ →



- Käytännössä otoskokoa n pidetään ”riittävän suurena”, kun **n** $>$ **30**.

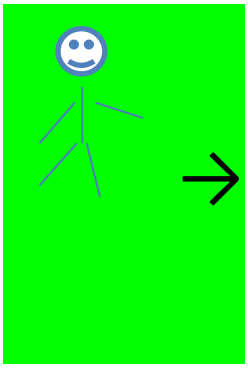
- Näin siis hallitaan ”mekanismi”, jonka mukaan sattuma määrää otoksessa realisoituvan keskiarvon $\bar{X}$ suuruuden.

- Ainoastaan ei-normaalinen muuttuja hyvin pienessä otoksessa jää käsittelemättä.

Sattuma toimii ohjesääntönsä mukaisesti.

Perusjoukko

muuttujan **X** todellinen keskiarvo $\boldsymbol{\mu}$ ja hajonta $\boldsymbol{\sigma}$



Otos

otoskeskiarvo $\bar{\mathbf{x}}$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ tai}$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}\right).$$

Huom.

1) $\bar{X}$:n otantajakaumasta näkyy, että $\mathbf{E\bar{X} = \mu}$,

joten otoskeskiarvo $\bar{X}$ ”pyrkii asettumaan” todellisen keskiarvon μ kohdalle.

Tästä seuraa, että **tarkastelun suunta voidaan kääntää**:

Otos todella poimitaan ja siitä laskettu keskiarvo $\bar{x}$ on ”keskimäärin” oikea arvio (**estimaatti**) tutkittavan muuttujan arvojen todelliselle (tuntemattomalle) keskimääräiselle suuruudelle μ perusjoukossa.

2) Otantajakauman varianssi

$$\frac{\sigma^2}{n} \text{ tai } \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

ja hajonta eli **keskiarvon keskivirhe** $\sigma_{\bar{x}}$ (←merkintä)

$$\frac{\sigma}{\sqrt{n}} \text{ tai } \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

mittaavat, kuinka paljon eri n:n suuruisista otoksista saatavat $\bar{X}$:n arvot voivat ”keskimäärin” vaihdella μ:n ympärillä.

- Odotettavissa oleva vaihtelu on sitä suurempaa, mitä suurempi hajonta σ on.

Tämä on **empiirisesti järkevää**:

Mitä erilaisempia muuttujan x arvot ovat perusjoukossa, sitä enemmän sattumalla on ”pelivaraa” tuottaa erisuuruisia keskiarvoja otokseen.

- Otoskoko n on keskivirheen $\boldsymbol{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$ nimittäjässä, joten

suuresta otoksesta laskettava $\bar{X}$:n arvo poikkeaa todellisesta keskiarvosta μ ”keskimäärin” vähemmän kuin pienestä otoksesta laskettu.

Myös tämä on **empiirisesti järkevää**:

Jos keskiarvon laskemiseen on käytettävissä paljon informaatiota perusjoukosta, $\bar{x}$ ”vastaa paremmin” todellista keskiarvoa.

- Kun otos poimitaan palauttamatta äärellisestä perusjoukosta, $\bar{X}$:n otantajakauman vaihtelu on pienempää, minkä kuvaa

äärellisen perusjoukon korjaustekijä $\boldsymbol{\frac{N-n}{N-1}} < 1$.

Tämäkin on **empiirisesti järkevää**:

Otannan edistyessä muuttujan arvojen vaihtelu perusjoukossa pienenee koko ajan, kun muuttujan arvojen määrä vähenee. Silloin sattumalla on vähemmän pelivaraa otoskeskiarvon $\bar{X}$ ”heiluttelemiseen” todellisen keskiarvon μ ympärillä.

Siis otanta palauttamatta on (hieman) edullisempi kuin otoksen poimiminen palauttaen.

Vastaavat ominaisuudet ovat voimassa yleisemminkin otoksesta laskettaville tunnusluvuille.

Esim. (jatkoa) Tehtaan tuottamien suklaalevyjen (perus-)joukossa

levyn paino $X \sim N(100\ g, (4\ g)^2)$.

Tuotannosta aiotaan poimia 80 suuruinen otos.

Perusjoukon kokoa ei tunneta, jolloin joudutaan toimimaan, kuin se olisi ääretön. Äärellisyyskorjausta ei voida käyttää hyväksi (!) otantajakauman varianssissa.

$$\bar{X} \sim N\left(100\ g, \frac{(4\ g)^2}{80}\right) = N(100\ g, 0.2\ g^2) = N(100\ g, (0.4472\ g)^2)$$

↑

Otoskoko on kasvatettava 4-kertaiseksi, jotta keskivirhe (sattuman ”pelivara”) pienenee puoleen. (Ks. aikaisemmasta, kun n = 20.)

Nyt todennäköisyys, että keskipaino otoksessa tulee olemaan alle 1 g:n päässä oikeasta keskipainosta 100 g, on

$$P(99 < \bar{X} < 101) = P\left(\frac{99-100}{0.4472} < \frac{\bar{X}-100}{0.4472} < \frac{101-100}{0.4472}\right)$$

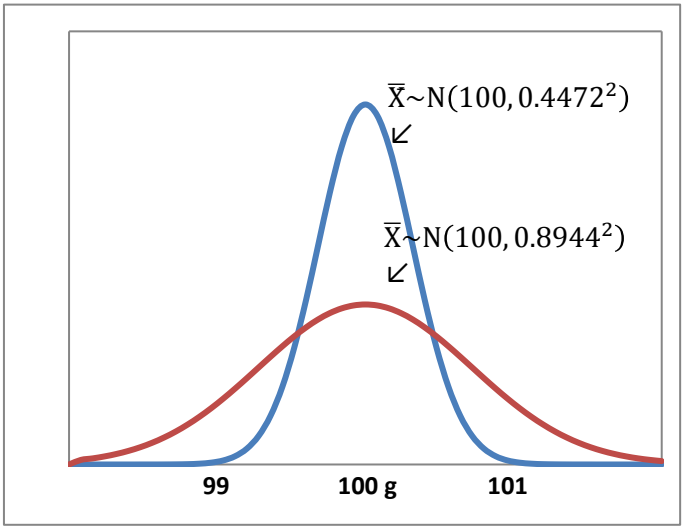
$$= P(-2.24 < Z < 2.24) = \phi(2.24) - \phi(-2.24) = \phi(2.24) - (1 - \phi(2.24))$$

$$= 2\,\phi(2.24) - 1 = 2 \cdot 0.9875 - 1 = 0.975$$

- Siis otoskeskiarvo $\bar{X}$ ”asettuu” peräti 97.5 % varmuudella näin lähelle todellista keskipainoa 100 g, jos otoskoko n = 80.

(Vrt. ”vain” 74 % varmuudella, jos otoksesta saatava informaatiomäärä on neljäsosa n = 20 tästä.)

Kuvassa on otoskeskiarvon otantajakauma, kun otoskoko n=20 ja n=80:



Esim. (jatkoa) Suklaatehdas väittää, että

levyn keskipaino μ = 100 g ja painon hajonta σ = 4 g.

50 suuruisessa otoksessa saatiin keskipainoksi vain $\bar{x}$ = 98.5 g.

Uskallatko väittää suklaatehtailijaa tämän perusteella huijariksi?

- Jos tehtaan väite (**nollahypoteesi H_0**) μ = 100 g

- olisi tosi,

- niin otoksesta laskettavan keskiarvon otantajakauma, jonka mukaan sattuma tällaisessa otantatilanteessa silloin toimisi, olisi ollut

$\bar{X} \sim N(100\text{ g}, \frac{(4\text{ g})^2}{50}) = N(100\text{ g}, 0.32\text{ g}^2) = N(100\text{ g}, (0.5657\text{ g})^2)$.

- Kuinka todennäköistä on, että tällainen sattuman toimintaa säätelevä mekanismi tuottaisi ”vain sattumalta”

otoksesta havaitun 1.5 g tai vielä enemmän todellisesta keskipainosta poikkeavan otoskeskiarvon?

- Suklaatehtaan tuotteiden laatua ei ole koskaan epäilty.

Nyt kuitenkin joudutaan pohtimaan, onko otoksessa havaittu 1.5 gramman keskimääräinen alipaino vain sattuman leikkiä vai ei.

Silloin on kohtuullista pitää tarkastelussa mukana mahdollisuus, että **yhtä hyvin** ”vain sattumalta” keskipaino $\bar{X}$ **olisi voinut** ”heilahtaa” yhtä paljon suurempaan suuntaan.

Silloin todennäköisyys, että vain sattumalta otokseen osuu tehtaan väitteen

(H_0:) μ = 100 g

epäilyksen alaiseksi saattava havaitun suuruinen poikkeama 1.5 g keskipainossa, on (ehdollinen todennäköisyys)

Spekuloidaan hypoteesilla, että tehtaan väite ↙ olisi totta

$$p = P(\bar{X} \leq 98.5 \text{ g tai } \bar{X} \geq 101.5 \text{ g} \mid \mu = 100 \text{ g})$$

↑ Näin kävi otoksessa.

↑ Näin olisi voinut yhtä hyvin käydä.

$$= P\left(\frac{\bar{X}-100}{0.5657} \leq \frac{98.5-100}{0.5657} \text{ tai } \frac{\bar{X}-100}{0.5657} \geq \frac{101.5-100}{0.5657}\right)$$

$= P(Z \leq -2.65 \text{ tai } Z \geq 2.65) = \phi(-2.65) + 1 - P(Z < 2.65)$

$= 1- \phi(2.65) + 1- \phi(2.65)$

$= 2 \cdot (1 - 0.9960)$

$= 0.008$

Siis keskimäärin vain 8 kertaa 1000:sta näin käy vain sattumalta.

- Varsin hyvin perustein voidaan päätellä, että tehtaan väite ei ole totta.

- Kuitenkin silloin hyväksytään 0.8 %:n **riski** sille, että

- keskipaino onkin väitetty 100 g ja

- tehdas pystyy sen todistamaan (vaikkapa punnitsemalla 10 000 … levyä)

- ja sattuma on ”vain sattumalta” aivan otantajakauman puitteissa generoinut tällaisen otoksen

- ja tehdas haastaa herjaajan oikeuteen ja vaatii korvausta aiheutetusta vahingosta …

Siis **päätöksenteko**:

uskalletaanko lähtökohtana olevasta (nolla-)hypoteesista luopua vai ei,

on suhteutettava siihen, kuinka vakavia käytännön seurauksia tällaisesta (**hylkäämis-**)**virheestä** seuraa.

Tässä tarkasteltiin tilannetta **2-suuntaisesti**. Ajateltiin, että poikkeama väitettyyn 100 gramman keskipainoon voisi tulla sattumalta yhtä hyvin ylöspäin.

Jos tässä voitaisiin jostain syystä ennen otoksen poimimista olla ”täysin varmoja”,

”että keskipaino ei nyt ainakaan 100 grammaa suurempi voi olla”,

niin (hylkäämisvirheen) riski on puolta pienempi

$p = P(\bar{X} \leq 98.5 \text{ g}$ ~~tai $\bar{X} \geq 101.5$ g~~$\mid \mu = 100 \text{ g}) = 0.004$.

Tällaiseen **1-suuntaiseen** päättelyyn (**testaamiseen**) on oltava todella vankat otoksen ulkopuolelta tulevat perusteet.

Tätä aihetta jatketaan myöhemmin.

$\bar{X}$:n otantajakauman avulla voidaan tutkia, mille välille otoskeskiarvon $\bar{x}$ ”pitäisi asettua”, kun otos poimitaan. Tällaisten **kontrollirajojen** määrääminen voi olla hyödyllistä mm. laadunvalvonnassa, kun verrataan todella realisoitavaa otoskeskiarvoa tällaiseen väliin.

Esim. Elintarvikeannoksen lisäaineen E määrä $X \sim N(200\text{ mg}, (15\text{ mg})^2)$.

Annoksista aiotaan poimia 100 suuruinen otos.

Määrää a niin, että 95 %:n varmuudella otoskeskiarvo $\bar{x}$ tulee poikkeamaan todellisesta keskiarvosta μ = 200 mg korkeintaan a:n verran.

Silloin sattuman käyttäytymistä säätelee otantajakauma

$$\bar{X} \sim N(200\text{ mg}, \frac{(15\text{ mg})^2}{100}) = N(200\text{ mg}, 2.25\text{ mg}^2) = N(100\text{ g}, (1{,}5\text{ mg})^2).$$

ja on oltava

$$0.95 = P(200-a \leq \bar{X} \leq 200+a)$$

$$= P(\frac{200-a-200}{1.5} < \frac{\bar{X}-200}{1.5} < \frac{200+a-200}{1.5}) = P(\frac{-a}{1.5} < Z < \frac{a}{1.5})$$

$$= \phi(\frac{a}{1.5}) - \phi(\frac{-a}{1.5}) = \phi(\frac{a}{1.5}) - (1 - \phi(\frac{a}{1.5}))$$

$$= 2\,\phi(\frac{a}{15}) - 1,$$

josta $\phi(\frac{a}{1.5}) = \frac{0.95+1}{2} = 0.975 = \phi(1.96)$ (← taulukosta).

Silloin $\frac{a}{1.5} = 1.96$ ja $a = 1.96 \cdot 1.5 = 2.94 \approx 3$ mg.

Huom. väli [200-a, 200+a] = [200-3, 200+3] = [197, 203] **ei ole luottamusväli**, kuten lukiossa on saatettu sitä virheellisesti nimittää.

Suhteellisen osuuden otantajakauma

seuraa varsin suoraan binomi- ja hypergeometrisen jakauman normaaliapproksimaatiosta:

Esim. (jatkoa) Aiotaan tehdä markkinatutkimus, jossa

- eräs taustatieto on, että tässä perusjoukossa on 45 % naisia.

- Otoskoko on n = 1000.

- Etukäteen halutaan tarkistaa, että ”riittävän suurella varmuudella” (mm.) naisten osuus tulee olemaan otoksessa ”lähellä” perusjoukossa olevaa 45 prosenttia.

Esimerkiksi

”kuinka varmasti ” eli kuinka suurella todennäköisyydellä

- naisten suhteellinen osuus poimit**tavassa** otoksessa **tulee poikkeamaan** naisten todellisesta 45 %:n osuudesta perusjoukossa

korkeintaan 2 % - yksikköä eli on välillä [0.43, 0,47]?

Perusjoukko on suuri, jolloin naisten lukumäärä otoksessa

$X \sim \text{Bin}(1000, 0.45)$

ainakin likimain poimittiinpa otos palauttaen tai palauttamatta.

Jakaumasta voidaan laskea joko (Excelillä) suoraan tai normaaliapproksimaation avulla $P(430 \leq X \leq 470)$.

Toisena vaihtoehtona on, että siirrytään suoraan suhteellisiin osuuksiin:

$np = 1000 \cdot 0.45 = 450 > 5$ ja $n(1-p) = 1000 \cdot (1 - 0.45) = 550 > 5$, joten

X:n jakaumaa voidaan approksimoida normaalisella apumuuttujalla

$Y \sim N(1000 \cdot 0.45, 1000 \cdot 0.45 \cdot (1 - 0.45)) = N(450, 247.5) = N(450, 15.732^2)$.

$\hat{P}$ = naisten suhteellinen osuus = $\frac{\text{naisten lkm otoksessa}}{\text{otoskoko}} = \frac{X}{1000}$.

Koska normaaliapproksimaatio käy, on $\hat{P}$:n jakauma likimain

$\hat{P} = \frac{X}{1000} \approx \frac{Y}{1000} \sim N(\frac{1}{1000} \cdot 450, (\frac{1}{1000})^2 \cdot 247.5) = N(0.45, 0.0002475)$

$= N(0.45, (0.015732)^2)$.

$P(0.43 \leq \hat{P} \leq 0.47) = P(\frac{0.43-0.45}{0.015732} \leq \frac{\hat{P}-0.45}{0.015732} \leq \frac{0.47-0.45}{0.015732})$

$= P(-1.27 \leq Z \leq 1.27) = \phi(1.27) - \phi(-1.27) = \phi(1.27) - (1 - \phi(1.27))$

$= 2\,\phi(1.27) - 1 = 2 \cdot 0.8980 - 1$

$= 0.796.$

- Siis noin 80 % **varmuudella** naisten osuus otoksessa tulee osumaan alle 2 % - yksikön päähän oikeasta arvosta 45 %.

- Toisaalta on noin 20 % suuruinen **riski**, että naisten osuus sattuukin olemaan kauempana 45 prosentista. Jos tällainen riski on liian suuri, sitä voi pienentää otoskokoa kasvattamalla. Silloin sattuman ”pelivara” pienenee.

Tämä näkyy selvästi(?) $\hat{P}$:n otantajakauman varianssista.

- Oikeastaan laskussa olisi pitänyt käyttää jatkuvuuskorjausta, mutta sen merkitys on tässä mitätön:

$P(430 \leq X \leq 470) \approx P(430 - ½ \leq Y \leq 470 + ½) = P(429.5 \leq Y \leq 470.5)$

$= P(\frac{429.5}{1000} \leq \hat{P} \leq \frac{470.5}{1000}) = P(0.4295 \leq \hat{P} \leq 0.4705) \approx P(0.43 \leq \hat{P} \leq 0.47)$

Edellä laskettiin

$EX = np = 1000 \cdot 0.45 \quad Var(X) = np(1-p) = 1000 \cdot 0.45 \cdot (1-0.45)$

$\downarrow \qquad\qquad \downarrow$

$\hat{P} = \frac{X}{1000} \approx \frac{Y}{1000} \sim N(\frac{1}{1000} \cdot \mathbf{450}, (\frac{1}{1000})^2 \cdot \mathbf{247.5})$

$= N(\frac{1}{1000} \cdot 1000 \cdot 0.45, (\frac{1}{1000})^2 \cdot 1000 \cdot 0.45 \cdot (1-0.45))$

naisten suhteellinen osuus perusjoukossa p $\qquad$ $1 - p$

$\downarrow \quad \searrow \qquad\qquad \swarrow$

$= N(\boxed{\mathbf{0.45}}, \frac{\boxed{\mathbf{0.45}} \cdot \boxed{(\mathbf{1-0.45})}}{\boxed{\mathbf{1000}}})$

$\uparrow$
otoskoko n

Suhteellisen osuuden otantajakaumalla on sama rakenne myös yleisesti:

Oletetaan, että niiden tilastoyksiköiden suhteellinen osuus, joilla on ominaisuus A, on **perusjoukossa** p = P(A).

Jos perusjoukosta poimittavan otoksen koko n on ”suuri”

eli np $>$ 5 ja n(1-p) $>$ 5,

niin otoksesta laskettavan suhteellisen osuuden $\hat{P}$ otantajakauma on likimain normaalinen.

- Jos otos poimitaan **palauttaen** tai

palauttamatta äärettömän (käytännössä hyvin) suuresta perusjoukosta tai perusjoukon kokoa ei tiedetä, on

suhteellisen osuuden $\hat{P}$ otantajakauma

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- Jos otos poimitaan **palauttamatta N:n suuruisesta perusjoukosta**,

niin

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}\right).$$

Rakenne on hyvin samanlainen kuin otoskeskiarvon jakaumalla.

Sattuma toimii ohjesääntönsä mukaisesti.

Perusjoukko		**Otos**
Ominaisuuden A suhteellinen osuus on p	 →	A: n suhteellinen osuus on $\hat{p}$

↑

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ tai}$$

2) ↓ ↙ 3)

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}\right)$$

1) ↗

1) Otoksesta saatava suhteellinen osuus $\hat{P}$ ”pyrkii” asettumaan todellisen suhteellisen osuuden p kohdalle. (Vrt. $\bar{X} \leftrightarrow \mu$)

2) Otantajakauman varianssi (, josta käytetään merkintää) $\sigma^2_{\hat{P}}$ ja **keskivirhe** $\sigma_{\hat{P}}$ mittaavat, kuinka suuri on sattuman ”pelivara” otoksesta saatavan suhteellisen osuuden $\hat{P}$ ”heiluttelemiseen” oikean suhteellisen osuuden p ympärillä.

Ne rakentuvat kahdesta osasta:

X:n hajonta DX

↙

$$\sigma^2_{\hat{P}} = \frac{p(1-p)}{n} \text{ ja } \sigma_{\hat{P}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

↖ otoskoko

- Mitä lähempänä p on arvoa 0.5, sitä suurempi on(?) ”sattuman pelivara”. (Piirrä f(p) = p(1-p):n kuvaaja.)

- Jos otoskokoa n kasvatetaan, keskivirhe pienenee.

3) Otannassa palauttamatta N:n kokoisesta perusjoukosta saadaan mukaan äärellisen perusjoukon korjaustekijä.

Vaihtelu pienenee otannan edistyessä ja tämä pienentää keskivirhettä. Otantajakauma on tiiviimmin todellisen suhteellisen osuuden p ympärillä.

Esim. Kunnassa M puoluetta Ö kannatti edellisissä vaaleissa 35 % äänestäjistä.

Puolueen epäillään sotkeutuneen törkyiseen lahjusskandaaliin. Uudet vaalit ovat tulossa, ja paikallislehti teki otantatutkimuksen. Siinä selvitettiin (mm.), onko puolueen Ö kannatus pienentynyt.

Lehden kesätoimittaja päättää etukäteen väittää artikkelissaan lehdelle tärkeän Ö:n kannatuksen **merkitsevästi** (**) pienentyneen, jos otoksesta havaittava vaaleja pienempi kannatus voi tulla vain sattumalta alle 1 % todennäköisyydellä.

Kunnan 40000 äänestysikäisestä poimittiin 1500 suuruinen otos palauttamatta.

Otoksessa Ö:tä kannatti 32 % vastaajista.

On ”täysin selvää”, että Ö:N kannatus ei ainakaan ole kasvanut edellisistä vaaleista. Silloin toimittajan riski:n suuruus paikkansa pitämättömän väitteen esittämiseen on

1-suuntaisen satunnaisen kannatuksen ”heilahduksen” todennäköisyys 32 prosenttiin, vaikka kannatus edelleen olisikin 35 %.

Jos ((?)nollahypoteesi H_0:) kannattajien osuus olisi edelleen p = 0.35,

niin sattuma olisi generoinut otoksen otantajakauman

(Korjaustekijä pienentää varianssia noin 3.7 %.) ↓

$$\hat{P} \sim N(0.35, \frac{0.35(1-0.35)}{1500} \cdot \frac{40000-1500}{40000-1}) = N(0.35, 0.0001516 \cdot \mathbf{0.9625})$$

$= N(0.35, 0.0121^2)$ mukaan.

Silloin on todennäköisyys, että otoksessa ”vain sattumalta” korkeintaan 32 % kannattaa Ö:tä, vaikka todellisuudessa kannatus koko äänestäjäkunnassa olisikin 35 %:

$p = P(\hat{P} \leq 0.32) = P(\frac{\hat{P} - 0.35}{0.0121} \leq \frac{0.32 - 0.35}{0.0121})$

$= P(Z \leq -2.48) = \phi(-2.48) = 1 - \phi(2.48) = 1 - 0.9934$

$= 0.0066$

$=$ **0.66 % $<$ 1 %,**

joten päätössääntönsä mukaan toimittaja käy jutun tekoon.

Tässäkin on jo otantajakaumien tutkimisen ohella samalla ennakolta käsitelty **hypoteesien testaamista**. Aiheeseen syvennytään tarkemmin myöhemmin, kun ensin on käsitelty toista **tilastollisen päättelyn** osa-aluetta **estimointiteoriaa**.

Muiden otoksesta laskettavien tunnuslukujen (mediaani Md, otosvarianssi s^2, korrelaatiokerroin r, jne.) otantajakaumien tarkka käsittely sivuutetaan tässä.