5 Hypoteesien testaamisesta

Hypoteesien testaaminen on myös otoksesta perusjoukkoon päin tehtävää päättelyä.

Estimoinnissa yritetään selvittää parhaalla mahdollisella tavalla jonkin perusjoukossa ”piilossa” olevan parametrin suuruus.

Testaamisessa näkökulma on hieman erilainen ja myös rajoitetumpi:

- Lähtökohtana on jokin perusjoukkoa koskeva väite (**hypoteesi**).

- Otoksen perusteella tutkitaan, onko tämä hypoteesi **tosi vai ei**.

Esim. (jatkoa) Juomien makutestissä tulos määräytyy kokeella, jossa maistaja yrittää erottaa oman suosikkijuomamerkkinsä kolmen vastaavanlaisen juoman joukosta.

Maistaminen toistetaan 13 kertaa

ja tulos on oikein tunnistettujen kertojen määrä.

Arvostettu juoma-asiantuntija ekonomisti E osallistuu juomien makutestiin ja tunnistaa 13 yrityksestä 10 kertaa juoman oikein.

Saiko ekonomisti E tuloksen vai arvaamalla?

Jos tunnistaminen olisi arvaamista, niin **nollahypoteesi**

H_0: π = P(Vastaus on oikein.) = 1/3

on tosi.

Jos ekonomisti E pystyy tunnistamaan juoman arvaamista paremmin, niin **vastahypoteesi**

H_1: $\pi > 1/3$

on totta.

Tässä kokeen järjestäjät ovat ekonomisti E:n aikaisempien suoritusten perusteella ”täysin varmoja”, että kokelaan tunnistuskyky ei ainakaan ole arvaamista heikompi ($\pi < 1/3$).

Silloin ongelmaa voidaan tarkastella **1-suuntaisesti** arvausta parempaan suuntaan.

Arvosteluraadin on valittava:

- **Jääkö** maistajakokelaille asetettu lähtökohta $\mathbf{H_0}$ (arvaa vain) **voimaan**,

- vai **hylätäänkö $\mathbf{H_0}$** ja uskotaan, että H_1 (kyllä se tunnistaa) on totta.

Maistajaseuran raati luopuu nollahypoteesista H_0 vain, jos näyttö tätä arvaamis-oletusta vastaan on todella vahva.

Seuran jäsenyys on hyvin haluttu, ja pääsyehdoksi on asetettu **päätössääntö**:

Jos kokelas voisi saada pelkästään **arvaamalla** saavuttamansa tuloksen **alle** α = **5 % todennäköisyydellä**,

niin **$\mathbf{H_0}$ hylätään** ja kokelas otetaan jäseneksi.

On tapana sanoa, että tulos on **melkein merkitsevä**, ja kokelas saa arvonimen MSM*.

Tätä päätössääntöä käyttäessään seura hyväksyy **riskin**, että 100:sta pelkästään arvaamalla maistavasta kokelaasta keskimäärin 5 onnekasta tulee **vain sattumalta** valituiksi, eli **H_0 hylätään, vaikka se on totta**. Tällaista valitettavaa, mutta sattumalta mahdollista, tapahtumaa sanotaan (H_0:n) **hylkäämisvirheeksi.**

Oleellista on, että seura pystyy sietämään $\alpha = 0.05$ suuruisen (H_0:n) **hylkäämisvirheen** tekemisen **riskin** MSM* arvonimeä antaessaan.

Jos H_0: $\pi = 1/3$, olisi tosi,

niin sattuma määrää kokeessa realisoituvan oikeiden arvausten lukumäärän binomijakauman X ~Bin(13, 1/3) mukaan.

Frekvenssifunktiosta saadaan todennäköisyys, että maistaja saavuttaa

vähintään **havaitun** tuloksen 10 oikein ↓	arvaamalla eli ehdolla, että **H_0 olisi tosi** ↙

$$\mathbf{p = P(X \geq 10 \mid \pi = 1/3)} = P(X = 10) + P(X = 11) + P(X = 12) + P(X = 13)$$

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

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

$\approx 0.00165 = 0.165$ %.

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

Koska $\mathbf{p = 0.00165 < 0.05 = \alpha}$,

asetetun päätössäännön mukaan $\mathbf{H_0}$ **hylätään**.

Hylkäämisvirheen riski(, jota myös lyhyesti sanotaan **p-arvoksi**)

$p = 0.00165$ alittaa jopa 1 %:n rajan.

Silloin sanotaan, että tulos on (seuran SM** arvonimeen oikeuttava)

(tilastollisesti) **merkitsevä**.

Kun arvaamishypoteesin H_0 hylkäämisen päätössäännössä $\alpha = 0.01$, keskimäärin vain yksi ”Hannu Hanhi” sadasta saa arvaamalla SM** tittelin.

Jos H_0:sta luopumisen päätössäännöksi on asetettu ankara (ESM*** tasoon oikeuttava) arvo $\alpha = 0.001$,

p-arvo 0.00165 ylittää sen ja H_0 jää voimaan.

Tulos **ei ole erittäin merkitsevä**.

Huom. Aikaisemmassa on, kuten on tavallista, binomijakauman toisen parametrin, ”onnistumistodennäköisyyden” symboli ollut p. Samaa merkintää on myös käytetty jonkin ominaisuuden A suhteellisesta osuudesta perusjoukossa.

Sama merkintä p on vakiintunut hylkäämisvirheen tekemisen riskin symboliksi.

Jotta asiat eivät sekoitu keskenään, nyt käytetään testaamisessa suhteellisesta osuudesta symbolia π.

Hypoteesien asettelusta

Ongelma esitetään kahden perusjoukkoa koskevan hypoteesin

- **nollahypoteesin H_0** ja

- **vastahypoteesin H_1**

avulla.

Nollahypoteesi on usein ”varovainen perustilanne”:

Esim. a) Juomien tunnistustestissä

H_0: π = P(Tunnistaa juoman oikein.) = 1/3, siis pelkästään arvaa.

b) Sanomalehden suunnittelemassa äänestäjien poliittisen kannan tutkimuksessa (ks. edellä)

H_0: π = 0.35 eli puolueen Ö on pysynyt samana kuin edellisissä vaaleissa

(, vaikka ehkä vahvastikin epäiltäisiin sen pienentyneen).

c) Tutkitaan, onko suklaalevyn todellinen keskipaino μ tehtaan ilmoittama 100 g.

H_0: μ = 100 g eli lähtökohtana on väitetty arvo(, vaikka sitä epäiltäisiinkin).

d) Alueen kotitalouksista tehdyssä markkinatutkimuksessa

- 1500 suuruisessa otoksessa otokseen osuneista kotitalouksista 30 % käyttää hyödykettä Ö.

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?

Vaikka uskottaisiin, että 4 % -yksikön ero ei ole ”sattuman leikkiä”, niin kuitenkin nollahypoteesi asetetaan tylsän varovaisesti:

H_0: $\pi_1 = \pi_2$ eli käyttäjien suhteellinen osuus koko perusjoukossa

ennen (π_1) ja jälkeen (π_2) kampanjan on yhtä suuri.

e) 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/l ja keskihajonta $s_1 = 1.2$ mmol/l ja

vertailuryhmän 150:n lumelääkettä saaneen

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

Vaikka lääkkeen kehittäjät tietenkin toivovat tuotteensa olevan hyvä, nollahypoteesi asetekaan kuitenkin varovaisesti

H_0: $\mu_1 = \mu_2$ eli todellinen keskimääräinen kolesterolimäärä on yhtä suuri.

Esim. Kuntosalilla aktiivisesti harjoittelevista poimittiin 500 suuruinen otos, jonka avulla tutkittiin (sallittujen) voimaharjoittelua tukevien lisäravinteiden käyttöä.

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

Voidaanko tämän perusteella päätellä, **riippuuko** käyttäminen sukupuolesta?

Nollahypoteesina on tässä (kuten aina riippuvuutta tutkittaessa)

H_0: Muuttujat sukupuoli ja käyttö ovat riippumattomia.

Vastahypoteesissa H_1 kerrotaan, missä tilassa perusjoukko on, jos H_0 ei ole tosi.

Jos H_1:ssä ei ilmoiteta (ei uskalleta, pystytä) ”poikkeaman suuntaa” H_0:n mukaiseen tilanteeseen verrattuna, on testaus **2-suuntainen**.

Jos poikkeaman suunta ilmoitetaan, testi on **1-suuntainen**.

1-suuntaisessa testauksessa tarvitaan havaintoaineiston ulkopuolelta empiiristä lisätietoa, joka oikeuttaa sulkemaan pois poikkeaman toiseen suuntaan H_0:n mukaisesta tilanteesta.

Esim. (jatkoa edelliseen) a) Juomien tunnistustestissä

vastahypoteesiksi on asetettava

H_1: $\pi \neq 1/3$ eli oikein osumisen todennäköisyys poikkeaa arvaamisesta,

jos kokelas on ”tavallinen tallaaja”, jonka kyvyistä ei ole mitään koetilanteen ulkopuolelta saatua lisätietoa.

Silloin on mahdollista, että ero H_0:n mukaiseen tilanteeseen on kahteen suuntaan:

- Maistaja voi tunnistaa juoman arvaamista paremmin,

- mutta mahdollista on myös, että hän tunnistaakin juoman arvaamista huonommin.

Ekonomisti E on yleisesti tunnettu hienona juoma-asiantuntijana, ja hänet on vihdoin saatu hakemaan juomaseuran jäsenyyttä.

Aikaisemman perusteella pidetään ”täysin varmana”, että tässä tapauksessa tunnistuskyky ”ei voi olla” arvaamista huonompi. Nyt vastahypoteesi voidaan tehdä 1-suuntaisesti.

H_1: π = P(Vastaa oikein.) $>$ 1/3 eli tunnistaa arvaamista paremmin.

b) Puolueen skandaaleihin sekaantumisesta johtuva ”poliittinen ilmasto” oikeuttanee sulkemaan pois mahdollisuuden, että puolueen kannatus olisi ainakaan suurentunut. Vastahypoteesiksi asetetaan 1-suuntaisesti

H_1: $\pi < 0.35$ eli kannatus on pienentynyt edellisistä vaaleista.

Vastahypoteesin H_0 asettelussa ajatuksena on:

- 1-suuntaista testiä **saa käyttää**, jos ”varmaa” lisätietoa on käytettävän havaintoaineisto ulkopuolelta.

- 2-suuntaista testiä **joudutaan käyttämään**, jos tällaista lisätietoa ei ole ja joudutaan olemaan varovaisempia.

c) Jos ilman mitään erityistä ennakkotietoa vain rutiininomaisesti tutkitaan suklaalevyjen painoa, on vastahypoteesi 2-suuntainen

H_1: $\mu \neq 100$ g eli paino poikkeaa ilmoitetusta jompaankumpaan suuntaan.

Käytettävissä voi olla jotain lisätietoa koneiden virheellisestä toiminnasta tms., mikä selvästi voi vaikuttaa tuotteen laatuun. Silloin saattaa olla riittävästi perusteita 1-suuntaiseen testaamiseen.

d) Jos ollaan varovaisia, asetetaan 2-suuntaisesti

H_1: $\pi_1 \neq \pi_2$ eli käyttäjien suhteellinen osuus on muuttunut.

Jos vastaavanlainen kampanja on aikaisemmin toiminut hyvin eikä käyttäjien osuus ole koskaan ainakaan pienentynyt, voi olla riittävästi perusteita 1-suuntaiseenkin vastahypoteesiin

H_1: $\pi_1 < \pi_2$ eli käyttäjien osuus oli pienempi ennen kampanjaa.

e) Uutta lääkettä tutkittaessa on ehkä parasta olla varovainen ja tehdä vastahypoteesi 1-suuntaisesti

H_1: $\mu_1 \neq \mu_2$ eli keskimääräiset kolesteroliarvot ovat eri suuria koe- ja vertailuryhmällä.

f) Otoksessa näyttäisi olevan poikkeamaa siihen suuntaan, että miehet käyttäisivät lisäravinteita enemmän kuin naiset.

Vastahypoteesin asettelussa kuitenkaan **ei saa ”vilkuilla” otokseen** ja sulkea sen perusteella pois toista suuntaa. Jos poikkeama H_0:n mukaisesta tilanteesta on vain ”sattuman leikkiä”, se olisi voinut **yhtä hyvin** sattua toiseen suuntaan.

Tässä ilman muita lisätietoja oletetaan tylsästi

H_1: Lisäravinteiden käyttö **riippuu** sukupuolesta.

Testaamisen lähtökohtana on nollahypoteesi H_0.

- H_0:aa pidetään totena (H_0 **hyväksytään**), jos otoksen ”perusjoukon tilasta” antama informaatio ei poikkea ”kohtuuttoman paljon” tilanteesta, joka otoksessa pitäisi keskimäärin olla H_0:n ollessa voimassa.

- H_0:aan ei uskota (H_0 **hylätään**), jos otoksessa havaittu tilanne on vahvasti ristiriidassa sen kanssa.

Siis H_0:an ei enää uskota,

- kun sattuman käyttäytymistä säätelevä ”mekanismi” eli otoksesta laskettavien tunnuslukujen otantajakaumat

- eivät pysty tuottamaan otoksessa todella realisoitunutta tilannetta kohtuullisen suurella todennäköisyydellä,

- jos perusjoukko olisi H_0:n väittämässä tilassa.

Silloin päätellään, että otoksessa havaittu asetelma on lähtöisin toisenlaisesta, H_1:n mukaisessa tilassa olevasta perusjoukosta.

Otos on vain suppea osa perusjoukosta, ja nollahypoteesin hylkäämis- tai hyväksymispäätöstä tehtäessä ei koskaan tiedetä varmasti, onko tehty päätös oikea.

- Jos H_0 hylätään, vaikka se todellisuudessa onkin totta, tehdään **hylkäämisvirhe** eli **1. lajin virhe**.

- Jos H_0 hyväksytään, vaikka se todellisuudessa onkin väärä, tehdään **hyväksymisvirhe** eli **2. lajin virhe**.

Tilastollinen analyysi on **päättelyä epävarmuuden vallitessa**, ja nämä virheet ovat aina mahdollisia.

- Vaikka H_0 olisi tosi, ”vain sattumalta” otokseen voivat realisoitua sellaiset havainnot, jotka ”puhuvat H_0:a vastaan”.

- Samoin sattuma voi valita otokseen sellaiset havainnot, jotka tukevat H_0:a, vaikka tämä olisikin väärä.

Siis testaamisessa voidaan päätyä neljään tulokseen:

	Todellinen tilanne	perusjoukossa:
Päätelmä:	**H_0 on tosi.**	**H_0 on väärä.**
H_0 hyväksytään	Päätös on **oikea.**	Tehdään **hyväksymisvirhe.**
H_0 hylätään.	Tehdään **hylkäämisvirhe**.	Päätös on **oikea.**

- Kun valitaan sopivaa testiä tutkittavana olevien hypoteesien testaamista varten, valinnassa on tärkeää **testin voimakkuus**.

Tämä tarkoittaa (väljästi määriteltynä), kuinka hyvin testi pystyy hylkäämään väärän nollahypoteesin eli välttämään hyväksymisvirheen.

Alkeita opeteltaessa valittavia vaihtoehtoja on vähän ja ne ovat kaikki oikein hyviä.

Testien voimakkuuksien tutkimiseen ei syvennytä tässä tarkemmin.

- Kun testausmenetelmä on valittu jotakin yksittäistä testaustilannetta varten, keskitytään hylkäämisvirheen tekemisen todennäköisyyden laskemiseen.

Testaaminen on ”väittelyä” perusjoukon tilasta.

1. ”Oletetaan nyt sitten, että H_0 olisi voimassa, ja katsotaan … ”

<u>2.”Katsominen”</u>

- Sattuma määrää sääntöjensä (otantajakaumien) mukaan otoksen sisällön. Otoksessa havaittava tilanne voi olla (kuitenkin vain pienellä todennäköisyydellä) hyvinkin erikoinen H_0:n kannalta, vaikka se olisi tosi.

- Kuitenkin on loogista toimia niin, että H_0 hylätään, jos havaittu tilanne on hyvin epätodennäköinen H_0:n ollessa voimassa.

- Otantajakaumien avulla lasketaan (tai ainakin arvioidaan), kuinka suuri on

hylkäämisvirheen tekemisen riski, jota sanotaan myös **merkitsevyystasoksi**. Se on ehdollinen todennäköisyys p = P(H_0 hylätään.│H_0 onkin tosi.)

↖ hylkäämisvirhe ↗

3. Jos hylkäämisvirheen riski, **p-arvo,** on ”pieni”, H_0 **uskalletaan** hylätä.

- Riippuu täysin tutkittavan ongelman luonteesta, kuinka ”pieni” merkitsevyystason p on oltava, ennen kuin H_0 uskalletaan hylätä.

- Usein päätetään etukäteen, kuinka suuri hylkäämisvirheen riski voidaan enimmillään sietää, jos H_0 päätetään hylätä.

Tällaisina kynnysarvoina käytetään yleisesti

rajoja (α =) 0.05, 0.01 ja 0.001.

Jos hylkäämisvirheen riskin p ylärajaksi on valittu $\alpha = 0.05$ ja H_0 hylätään ($p<0.05$), niin tulosta sanotaan **melkein merkitseväksi** ja sanotaan,

että H_0 hylätään 5 %:n merkitsevyystasolla.

Jos H_0 hylätään 1 %:n merkitsevyystasolla ($p<0.01$), tulos on **tilastollisesti merkitsevä** ja jos H_0 hylätään 0.1 %:n merkitsevyystasolla ($p < 0.001$), tulos on **erittäin merkitsevä**.

Seuraavassa on joitain tärkeitä erikoistapauksia hypoteesien testaamisesta.

Suhteellisen osuuden testi yhden perusjoukon tapauksessa

Esim. Kunnassa M puoluetta Ö kannatti edellisissä vaaleissa 35 % äänestäjistä.

Paikallislehti tekee otantatutkimuksen, jossa selvitettiin (mm.), onko puolueen Ö kannatus pienentynyt.

Puolueen epäillään sotkeutuneen törkyiseen lahjusskandaaliin ja ”poliittisen ilmaston” perusteella voidaan testata 1-suuntaisesti

H_0: $\pi = 0.35$ ($=\pi_0$) eli Ö:n kannatus ei ole muuttunut.

H_1: $\pi < 0.35$ eli Ö:n kannatus on pienentynyt.

Etukäteen päätetään, että

testaamisessa voidaan sietää korkeintaan α = 1 %:n suuruinen hylkäämisvirheen tekemisen riski.

Kunnan N = 40000 äänestysikäisestä poimittiin n = 1500 suuruinen otos palauttamatta.

Otoksessa Ö:tä kannatti $\hat{p}$ = 32 % vastaajista.

Jos nollahypoteesi H_0: π = 0.35 pitäisi paikkansa,

niin sattuma olisi generoinut otoksen otantajakauman

H_0:ssa oletettu π_0 N n

$\swarrow \quad \downarrow \quad \searrow \qquad \downarrow \quad \downarrow$

$\hat{P} \sim N(0.35, \frac{0.35(1-0.35)}{1500} \cdot \frac{40000-1500}{40000-1}) = N(0.35, 0.0121^2)$ mukaan.

$\uparrow \qquad \uparrow$

n N

Silloin on todennäköisyys, että otoksessa ”vain sattumalta” korkeintaan 32 % kannattaa Ö:tä eli **hylkäämisvirheen tekemisen riski** on

p = P($\hat{P} \leq 0.32 \mid \pi = 0.35$) = P($\frac{\hat{P} - 0.35}{0.0121} \leq \frac{0.32 - 0.35}{0.0121}$)

= P(Z $\leq$ - 2.48) = ϕ(- 2.48) = 1 - ϕ(2.48) = 1- 0.9934

= 0.0066

p = 0.66 % $<$ **1 % = $\boldsymbol{\alpha}$**, joten H_0 uskalletaan hylätä.

Päätellään, että Ö:n kannatus on pienentynyt.

<u>Esim.</u> (jatkoa)

Otantatutkimuksessa kysyttiin myös, aikovatko haastateltavat ensi vuonna tilata tutkimuksen teettäneen lehden.

- Kuluvana vuonna kunnan äänestysikäisistä 60 % on tilannut lehden ja

- otoksen 1500:sta haastatellusta 62.0 % arveli, että lehti tilataan.

Voidaanko tämän perusteella päätellä, että tilaajien määrä on **muuttunut**?

Lehdessä ei ole ennen tutkimusta selvää käsitystä, onko lehden suosio muuttunut ja, jos on, niin mihin suuntaan.

Silloin on testattava 2-suuntaisesti ja hypoteesit ovat

H_0: π = 0.60 (=π_0) eli tilaajien osuus on edelleen <u>sama</u> kuin ennen.

H_1: $\pi \neq$ 0.60 eli tilaajien osuus on <u>muuttunut</u> (kasvanut tai pienentynyt.)

Jos H_0 tulisi hylätyksi, vaikka mitään muutosta ei olisikaan tapahtunut, seuraukset eivät ole mitenkään vakavat. Päätetään testata **5 %:n merkitsevyystasolla**.

Jos H_0 on tosi eli tilaavia olisi edelleen 60 %, niin sattuma on tuottanut otoksen sisällön otantajakauman

$$\hat{P} \sim N(\mathbf{0.60}, \frac{\mathbf{0.60}(1-\mathbf{0.60})}{1500} \cdot \frac{40000-1500}{40000-1}) = N(0.60, 0.0124^2) \text{ mukaan.}$$

Jos otoksessa havaittu 62 % on vain ”sattuman leikkiä”, olisi yhtä hyvin voinut käydä niin, että tilaajia olisi ollut 2 % -yksikköä 60 %:ia vähemmän.

Hylkäämisvirheen tekemisen riski on silloin

Otoksessa sattui käymään näin.

↓

p = P($\hat{P} \leq 0.58$ tai $\boldsymbol{\hat{P} \geq 0.62}$ | $\pi = 0.60$)

↑

Yhtä hyvin olisi voinut käydä näin.

$$= P\left(\frac{\hat{P} - 0.60}{0.0124} \leq \frac{0.58 - 0.60}{0.0124}\right) + \boldsymbol{P\left(\frac{\hat{P} - 0.60}{0.0124} \geq \frac{0.62 - 0.60}{0.0124}\right)}$$

$$= P(Z \leq -1.61) + \boldsymbol{P(Z \geq 1.61)} = \phi(-1.61) + \boldsymbol{1 - P(Z < 1.61)}$$

$$= 1 - \phi(1.61) + \boldsymbol{1 - \phi(1.61)} = 1 - 0.9463 + \boldsymbol{1 - 0.9463}$$

$$= 0.0537 + \boldsymbol{0.0537} = 2 \cdot \boldsymbol{0.0537} = 0.1074.$$

↑ ↑

2 − suuntaisessa testissä hylkäämisvirheen riski p on **aina** 2 − kertainen 1 − suuntaiseen testiin verrattuna

p = 0.1074 $>$ 0.05 = α ja H_0:aa ”ei uskalleta” hylätä.

Huom. Tulos ei suinkaan tarkoita, että olisi saatu lopullinen totuus asiasta.

- Otoksen informaation ”todistusvoima H_0:aa vastaan” vain ei ole riittävä

- ja edelleen on pidettävä mahdollisena, että tilanne ei ole muuttunut.

Kuitenkin on voitu tehdä **hyväksymisvirhe**.

- Todellinen suhteellinen osuus π on voinut muuttua ”vähän”, mutta

- **testin voimakkuus** ei riitä paljastamaan muutosta

- näin **pienen otoksen** informaation avulla.

Yleensäkin testauksessa H_0:n voimaan jääminen ei ole välttämättä lopullinen totuus. Otoksen informaatio ei vain riitä ”kaatamaan” H_0:aa.

Jos epäillään, että H_0 ei ole sittenkään tosi, se paljastuu kyllä (lisää resursseja käyttämällä) suuremmasta otoksesta.

Yleisestikin

suhteellisen osuuden testaamisen vaiheet ovat samat kuin esimerkeissä:

- Perusjoukossa ominaisuuden A suhteellinen osuus tilastoyksiköissä on π

- ja n:n suuruisesta otoksesta on saatu π:n estimaatiksi $\hat{p}$.

1. Asetetaan hypoteesit:

Nollahypoteesi **H_0: $\boldsymbol{\pi = \pi_0}$,**

jossa π_0 on jokin oletettu, väitetty, aikaisempi, tms. arvo.

Vastahypoteesi **H_1: $\boldsymbol{\pi \neq \pi_0}$,**

jos joudutaan käyttämään 2-suuntaista testiä,

ja **H_1: $\boldsymbol{\pi < \pi_0}$** tai **H_1: $\boldsymbol{\pi > \pi_0}$,**

jos otoksen ulkopuolelta saadun tiedon perusteella ”tiedetään” mahdollisen poikkeaman suunta H_0:n mukaisesta tilanteesta.

2. Pohditaan, kuinka suuri on suurin siedettävissä oleva hylkäämisvirheen riski α.

3. Oletetaan, että H_0 olisi tosi, ja selvitetään, minkä otantajakauman

mukaan sattuma silloin olisi generoinut otoksen.

Jakauma olisi

<u>kaikissa muissa tapauksissa</u>: paitsi

$$\hat{\mathbf{P}} \sim \mathbf{N}\left(\boldsymbol{\pi_0}, \frac{\boldsymbol{\pi_0(1-\pi_0)}}{\mathbf{n}}\right)$$

$$\uparrow$$

<u>kun otos poimitaan palauttamatta N:n suuruisesta perusjoukosta</u>:

$$\hat{\mathbf{P}} \sim \mathbf{N}\left(\boldsymbol{\pi_0}, \frac{\boldsymbol{\pi_0(1-\pi_0)}}{\mathbf{n}} \cdot \frac{\mathbf{N-n}}{\mathbf{N-1}}\right)$$

$$\uparrow$$

”Spekuloidaan” sillä, että $\boldsymbol{\pi_0}$ olisi oikea suhteellinen osuus.

Otantajakauman avulla lasketaan

otoksessa havaitun ja vielä sitäkin poikkeavamman tilanteen (”hännän”) todennäköisyys

$\mathbf{P(\hat{P} \leq \hat{p} \mid \boldsymbol{\pi} = \boldsymbol{\pi_0})}$ (, jos $\hat{p} < \pi_0$,) tai $\mathbf{P(\hat{P} \geq \hat{p} \mid \boldsymbol{\pi} = \boldsymbol{\pi_0})}$ (, jos $\hat{p} > \pi_0$,)

4. Hylkäämisvirheen tekemisen riski p on

- edellä laskettu todennäköisyys, jos voidaan testata 1–suuntaisesti, ja

- ja p = **2**$\cdot$ P($\hat{P} \leq \hat{p} \mid \pi = \pi_0$) (tai **2**$\cdot$P($\hat{P} \geq \hat{p} \mid \pi = \pi_0$)),

jos joudutaan testaamaan 2-suuntaisesti.

5. Jos laskettu näin laskettu

hylkäämisvirheen riski $\mathbf{p} < \boldsymbol{\alpha}$ = riskille asetettu yläraja,

niin **H_0 uskalletaan hylätä**.

Muuten H_0 jää voimaan.

Otantajakaumat perustuvat normaaliapproksimaatioon. Tässä vaaditaan vastaavasti kuin edellä,

että $\mathbf{n} \cdot \boldsymbol{\pi_0} > \mathbf{5}$ ja $\mathbf{n(1-} \boldsymbol{\pi_0}) > 5$.

Ekonomisti E:n juomien tunnistuskykyä tutkittaessa testausasetelma oli aivan vastaava.

Siinä approksimointiehto ei täyty, kun 13$\cdot$(1/3) $\approx$ 4.3.

Otantajakaumana käytettiin H_0:n mukaista binomijakaumaa.

Minkä tahansa parametrin arvon testaamisessa voidaan myös käyttää luottamusväliä apuna:

2-suuntaisen testin ja luottamusvälin yhteys

<u>Esim.</u> (jatkoa) Lehden otantatutkimuksessa kysyttiin, aikovatko haastateltavat ensi vuonna tilata tutkimuksen teettäneen lehden.

- Kuluvana vuonna kunnan lehden on tilannut 60 % äänestysikäisistä ja
- otoksen 1500:sta haastatellusta 62.0 % arveli, että lehti tilataan.

Voidaanko tämän perusteella päätellä, että tilaajien määrä on **muuttunut**?

Hypoteesit ovat samat kuin edelläkin. Siis tilaajien todellinen suhteellinen osuuden arvo π on

H_0: π = 0.60 kuten aikaisemminkin tai

H_1: $\pi \neq 0.60$ muuttunut.

Hypoteesit jätetään hetkeksi sivuun ja tutkitaan, mitä otoksen perusteella voidaan sanoa π:n suuruudesta:

Kun tässä aiotaan testata α = 5 %:n merkitsevyystasolla,

- hyväksytään 5 %:n **risk**i hylkäämisvirheen tekemiselle.
- Silloin toisaalta päättely on oikea 95 %:n **varmuudella**.

<u>95 %:n luottamusväli oikealle π:n arvolle:</u>

Otoskoko n = 1500 ja sitä kautta vapausasteluku f = 1500 - 1 = 1499 ovat tässä niin suuria, että luottamuskertoimena voidaan käyttää normaalijakaumasta saatua

$z_{0.95}$ = 1.96 (Vrt. $t_{0.95}(1499)$ = 1.961548 Excelistä.)

Päätepisteet ovat

$$0.620 \pm 1.96\cdot \sqrt{\frac{0.620\cdot(1-0.620)}{1500}\cdot\frac{40000-1500}{40000-1}} = 0.620 \pm 0.024$$

ja 95 %:n luottamusväli tilaajien todelliselle suhteelliselle osuudelle kunnassa on

(0.620 – 0.024, 0.620 + 0.024) = (0.596, 0.644).

Silloin päätellään:

1) Oikea π: n arvo
on jossain tällä välillä
95 %: n suuruisella **varmuudella**.

↙ ↙ ↓ ↘ ↘

|__|________________|__________________|

↑ 0.596 ↖ 0.620 0.644 ↑

3) ↖ H_0: n mukainen
arvo $\boldsymbol{\pi_0 = 0.600}$

↑ ↑

↑ ↑

↖ ↗

↖ ↗

2) ↖ Kääntäen: On 5 %: n **riski**, että ↗
oikea π: n arvo
onkin välin ulkopuolella

4) H_0:n mukainen arvo π_0 = 0.60 on 95 %:n luottamusvälillä eli

”95 %:n varmuuden puitteisiin mahtuu vielä mahdollisuus, että π = 0.600.

5) Silloin on hylkäämisvirheen riski $p > 0.05 = \alpha$, jos H_0 hylättäisiin,

ja H_0 jää voimaan 5 %:n merkitsevyystasolla.

Menetelmä toimii myös yleisesti, kun testataan minkä tahansa parametrin arvon suuruutta 2-suuntaisesti:

- Jos suurin siedettävissä oleva hylkäämisvirheen tekemisen **riski on α**,

- tehdään parametrille luottamusväli **luottamustasolla c = 1 - α**.

- Jos H_0:n mukainen parametrin arvo on luottamusvälillä, H_0 jää voimaan.

Muuten H_0 hylätään.

Keskimääräisen suuruuden μ testi yhden perusjoukon tapauksessa

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 ja otoksesta laskettiin (tietenkin) myös keskihajonta, joka oli s = 3.8 g, ja painon jakauma näyttää normaaliselta.

Uskallatko väittää suklaatehtailijaa tämän perusteella huijariksi?

Suklaatehtaan tuotteiden laatua ei ole koskaan epäilty ja 1-suuntaiseen testaukseen ei ole perusteita. Hypoteesit ovat

H_0: $\mu = 100$ g

H_1: $\mu \neq 100$ g

Jos testaamisessa sattuu tulemaan hylkäämisvirhe ja se paljastuu myöhemmin suuremmasta havaintoaineistosta, seuraukset ovat epämiellyttävät.

Testaajat päättävät, että heidän suurin siedettävissä oleva hylkäämisvirheen riskinsä on $\alpha = 0.01$ eli he testaavat 1 %:n merkitsevyystasolla.

Kuten edellä jo laskettiin, hylkäämisvirheeseen päätymisen riski on

p = $P(\bar{X} \leq 98.5 \text{ g tai } \bar{X} \geq 101.5 \text{ g} \mid \mu = 100 \text{ g})$.

Lasku jatkui

H_0:n mukaisen otantajakauman

σ ?

$\bar{X} \sim N(100\text{ g}, \frac{(\mathbf{4\ g})^2}{50})$ = N(100 g, 0.32 g^2) = N(100 g, (0.5657 g)^2) avulla.

- Kuitenkaan ei tiedetä, onko oikea hajonnan arvo todellakin σ = 4 g, vaikka otos (s=3.8 g) tätä jonkin verran tukee.

- Paras tieto asiasta on kuitenkin otoksesta laskettu s = 3.8 g. Laskussa on käytettävä sitä ja standardoinnissa

keskivirhe $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4\text{ g}}{\sqrt{50}}$ korvataan estimaatillaan $\frac{s}{\sqrt{n}} = \frac{3.8\text{ g}}{\sqrt{50}}$

Siis

p = P($\bar{X} \leq$ 98.5 g tai $\bar{X} \geq$ 101.5 g | μ = 100 g)

$$= P\left(\boxed{\frac{\bar{\mathbf{X}}-\boldsymbol{\mu}}{\frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}}} \leq \frac{98.5-100}{\frac{3.8}{\sqrt{50}}} \text{ tai } \frac{\bar{X}-100}{\frac{3.8}{\sqrt{50}}} \geq \frac{101.5-100}{\frac{3.8}{\sqrt{50}}}\right)$$

↑

Voidaan osoittaa, että(?) tämä standardoitu muuttuja, jota sanotaan **testisuureeksi**, on

t-jakautunut vapausastein f = n - 1. (Vrt. luottamusväli edellä.)

Muuten lasku jatkuu samalla tavalla kuin edellä

$= P(t(50\text{-}1) \leq -2.791$ tai $t(49) \geq 2.791)$ ja lyhyemmin merkittynä

$= P(t \leq -2.791$ tai $t \geq 2.791)$

$= \mathbf{F_{t(49)}}(-2.791) + 1 - P(t < 2.791)$

↑

t-jakauman kertymäfunktio

$= 1- F_{t(49)}(2.791) + 1- F_{t(49)}(2.791)$

$= 2 \cdot (1 - \mathbf{F_{t(49)}(2.791)})$

↖ Excelistä **0.996266** (Vrt. $\phi(2.79) = 0.9974$.)

$\approx 2 \cdot 0.004$

$= 0.008$

Koska

- hylkäämisvirheen tekemisen riski $\mathbf{p \approx 0.008 < 0.01 = \alpha}$,
- niin H_0 uskalletaan hylätä
- ja päätellään, että keskipaino μ poikkeaa väitetystä 100 grammasta.

- Alun perin testattiin 2-suuntaisesti. Kun nyt päädyttiin H_0:n hylkäämiseen, voidaan toki nyt tarkemmin päätellä, että poikkeama on nimenomaan pienempään suuntaan.

Otoskoko n=50 on niin suuri ($>$ 30), että ei tehdä suurta virhettä,

jos p-arvo lasketaan vastaavalla tavalla normaalijakauman avulla.

Silloin tulos on p $\approx$ 0.005 ($<$ 0.008 oikea arvo) eli silloin **aliarvioidaan** hylkäämisvirheen tekemisen **riski**.

Tällaista testausmenettelyä sanotaan **radikaaliksi**. Jos jokin menetelmä taas yliarvioi riskin, sitä sanotaan **konservatiiviseksi**.

Jos t-jakauman kertymäfunktion arvoja ei voida katsoa Excelistä, on arvioitava **hylkäämisvirheen riski p** taulukoista saatavien **kriittisten rajojen** avulla:

Edellisessä laskussa oli

$$p = P(\bar{X} \leq 98.5 \text{ g tai } \bar{X} \geq 101.5 \text{ g} \mid \mu = 100 \text{ g})$$

$$= P\left(\boxed{\frac{\bar{\mathbf{X}}-\boldsymbol{\mu}}{\frac{s}{\sqrt{n}}}} \leq \frac{98.5-100}{\frac{3.8}{\sqrt{50}}} \text{ tai } \frac{\bar{X}-100}{\frac{3.8}{\sqrt{50}}} \geq \frac{101.5-100}{\frac{3.8}{\sqrt{50}}}\right)$$

$\uparrow$

$\downarrow$

= P($t \leq$ **-2.791** tai $t \geq 2.791$)

....

= 0.008 $<$ 0.01 = α

Koska testisuureen t = $\frac{\bar{\mathbf{X}}-\boldsymbol{\mu}}{\frac{s}{\sqrt{n}}}$

arvo t = $\frac{98.5-100}{\frac{3.8}{\sqrt{50}}}$ = -2.791 poikkeaa (tässä alaspäin) paljon nollasta,

”tilanne on **kriittinen H_0:n kannalta**” ja

lasku johtaa siihen, että p = 0.008 $<$ 0.01 = α.

t-jakauman taulukoissa on valmiiksi laskettuja **kriittisiä rajoja** testisuureelle t.

Ne kertovat,

kuinka paljon testisuureen arvon pitää poiketa nollasta, jotta hylkäämisvirheen riski p on valitun sietorajan α suuruinen.

Edellisessä t = $\frac{9\bar{8}.5-100}{\frac{3.8}{\sqrt{50}}}$ = -2.791.

Vapausasteluku f = 50 – 1 = 49 $\approx$ 50 ja α = 0.01 ja testi on 2-suuntainen.

t-jakauman taulukossa

- vapausasteluku f $\approx$ 50 määrää rivin ja

- otsikon ”2 – suuntaisen testin kriittisiä rajoja” alta α = 0.01 määrää sarakkeen.

- Keskeltä saadaan kriittinen raja $t_{0.01}^{(2)}(50) = 2.678$.

Siis,

- jos 2-suuntaisessa testissä

- testisuureen t arvo olisi tasan – 2.678 tai +2.678,

- olisi hylkäämisvirheen riski tasan p = 0.01 = α.

Nyt testisuureen arvo on kauempana t-jakauman ”hännällä”

t = -2.791 $<$ – 2.678 ja hylkäämisvirheen riski p $<$ 0.01

ja H_0 voidaan hylätä 1 %:n merkitsevyystasolla.

Tässäkin voidaan **testata luottamusvälin avulla:**

Otoskoko oli n = 50, keskipaino $\bar{x}$ = 98.5 g ja hajonta s = 3.8 g.

Maksimiriski saa olla α = 0.01, jolloin luottamustaso c = 1- 0.01 = 0.99.

Vapausasteluku f = 50 – 1 = 49 ≈ 50.

t-jakauman taulukosta saadaan (samasta kohdasta kuin äsken)

luottamuskerroin $t_{0.99}(50) = 2.678$.

99 %:n luottamusväli päätepisteet ovat

$98.5 \pm 2.678 \cdot \frac{3.8}{\sqrt{50}} = 98.5 \pm 1.44$, josta saadaan

(98.5 – 1.4, 98.5 + 1.4) = (97.1, 99.9).

Samalla tavalla kuin aikaisemmin päätellään:

1) Oikea μ: n arvo
on jossain tällä välillä
99 %: n suuruisella **varmuudella**.

↙ ↙ ↓ ↘ ↘

97.1 98.5 99.9

↑ ↑

↑ 3) H$_0$: n mukainen ↗
arvo $\boldsymbol{\mu_0} = \mathbf{100\ g}$ ↑

2) ↖ Kääntäen: On 1 %: n **riski**, että ↗
oikea μ: n arvo
onkin välin ulkopuolella

4) H_0:n mukainen arvo μ_0 = 100 g ei ole 99 %:n luottamusvälillä eli

”99 %:n varmuuden puitteisiin” ei mahdu enää mahdollisuus μ = 100 g.

5) Silloin on hylkäämisvirheen riski $p < 0.01 = \alpha$

ja H_0 hylätään 1 %:n merkitsevyystasolla.

Esim. Tehtaan jätevesien ympäristöluvassa on ehtona, että lievästi myrkyllistä kemikaalia K saa olla jätevedessä keskimäärin 10 mg/l.

Tehtaassa havaitaan toimintahäiriö puhdistusprosessissa ja jätealtaasta ”poimitaan” satunnaisesti n = 20 yhden litran suuruista jätevesinäytettä asian tutkimiseksi.

Keskimääräinen K:n määrä oli $\bar{x}$ = 12.4 mg/l ja hajonta s = 4.2 mg/l ja jakauma näytti likimain normaaliselta.

Ylittääkö keskimääräinen K:n määrä μ luvan rajan 10 mg/l?

- Häiriön vuoksi ympäristöviranomaiset pitävät selvänä, että μ ei ole ainakaan alle sallitun rajan.

- Tehtaan johto taas vaatii, että testauksessa voidaan sietää vain 0.1 %:n hylkäämisvirheen riski.

1) Hypoteesit asetetaan 1-suuntaisesti

H_0: μ = 10 mg/l

H_1: $\mu > 10$ mg/l

ja

2) testataan α = 0.1 %:n merkitsevyystasolla.

3) Testisuureen arvo on t = $\frac{12.4 - 10.0}{\frac{4.2}{\sqrt{20}}}$ = 2.555.

4) t-jakauman taulukosta saadaan

vapausastelukua f = 20 -1 = 19 ja merkitsevyystasoa α = 0.001

vastaava 1- suuntaisen testi kriittinen raja $t_{0.001}^{(1)}(19) = 3.579$.

5.a) Koska testisuureen arvo $t = 2.555 < 3.579 = t_{0.001}^{(1)}(19)$,

niin hylkäämisvirheen riski

p = (= $P(\bar{X} \geq 12.4$ mg/l | $\mu = 10$ mg/l) = … = $P(\mathbf{t} \geq 2.555)) > 0.001$,

ja H_0:aa jää voimaan.

5.b) t-jakauman kertymäfunktion arvot saadaan Excelistä ja hylkäämisvirheen tekemisen riski voidaan laskea tarkkaan:

p = (= $P(\bar{X} \geq 12.4$ mg/l | $\mu = 10$ mg/l) = …)

$= P(\mathbf{t} \geq 2.555) = 1 - P(\mathbf{t} < 2.555) = 1 - F_{t(19)}(2.555) = 1 – 0.990324 = 0.0097$

$p = 0.0097 > 0.001$ ja H_0 jää voimaan.

Näyttö H_0:aa vastaan ei ole riittävä.

<u>Huom.</u> Kriittisistä rajoista käytettävät merkinnät eivät ole valitettavasti vakiintuneet täysin samoiksi eri esityksissä.

Tässä käytetään ”lyhennettyä puhetta”:

”kriittinen raja

2-suuntaisessa ↓ testissä 1-suuntaisessa ↓ testissä

t-jakaumasta → $\mathbf{t}_{\boldsymbol{\alpha}}^{(2)}$**(f)** ja vastaavasti $\mathbf{t}_{\boldsymbol{\alpha}}^{(1)}$**(f)**

merkitsevyystasolla ↑ ↖vapausastein”

Jos **X** $\sim$**N(**$\boldsymbol{\mu}$**,** $\boldsymbol{\sigma}^2$**) perusjoukossa** ja **H**$_0$**:** $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ olisi tosi,

otantajakauma, jonka mukaan sattuma silloin olisi generoinut otoksen

olisi

<u>kaikissa muissa tapauksissa</u>: paitsi

<u>kun otos poimitaan palauttamatta N:n suuruisesta perusjoukosta</u>:

$$\bar{\mathbf{X}} \sim \mathbf{N}\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\sigma}^2}{\mathbf{n}}\right)$$

$$\bar{\mathbf{X}} \sim \mathbf{N}\left(\boldsymbol{\mu}_0, \frac{\boldsymbol{\sigma}^2}{\mathbf{n}} \cdot \frac{\mathbf{N}-\mathbf{n}}{\mathbf{N}-\mathbf{1}}\right)$$

↑ ↑

”Spekuloidaan” sillä, että $\boldsymbol{\mu}_0$ olisi oikea keskimääräinen suuruus.

Kun tuntematon todellinen hajonta σ korvataan otoksesta lasketulla estimaatillaan s, on otantajakaumasta standardoimalla saatava

H_0:n mukainen odotusarvo

↙ ↘

testisuure $\mathbf{t} = \dfrac{\bar{\mathbf{X}} - \mu_0}{\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}}$ tai $\mathbf{t} = \dfrac{\bar{\mathbf{X}} - \mu_0}{\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}\sqrt{\frac{\mathbf{N}-\mathbf{n}}{\mathbf{N}-\mathbf{1}}}}$,

↗ ↖

otoksesta estimoitu keskivirhe

t-jakautunut vapausastein f = n-1.

Esimerkeissä nähtiin, että tämän testisuureen arvoon tiivistyy testauksessa tarvittava otoksesta saatava informaatio.

<u>Todellisen keskimääräisen suuruuden μ testaamisen vaiheet</u>

ovat samat kuin esimerkeissä:

- Tutkittavan muuttujan arvojen jakauma perusjoukossa on **X** $\sim$**N(**$\boldsymbol{\mu}$**,** $\boldsymbol{\sigma}^2$**)**

- ja n:n suuruisesta otoksesta lasketaan

(μ:n estimaatti) otoskeskiarvo x̄ ja (σ:n estimaatti) otoshajonta s.

1) Asetetaan hypoteesit:

Nollahypoteesi **H_0: $\boldsymbol{\mu} = \boldsymbol{\mu_0}$,**

jossa $\boldsymbol{\mu_0}$ on jokin oletettu, väitetty, aikaisempi, tms. arvo.

Vastahypoteesi **H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu_0}$,**

jos joudutaan käyttämään 2-suuntaista testiä,

ja **H_1: $\boldsymbol{\mu} < \boldsymbol{\mu}_0$** tai **$H_1$: $\boldsymbol{\mu} > \boldsymbol{\mu_0}$,**

jos otoksen ulkopuolelta saadun tiedon perusteella ”tiedetään” mahdollisen poikkeaman suunta H_0:n mukaisesta tilanteesta.

2) Pohditaan, kuinka suuri on suurin siedettävissä oleva hylkäämisvirheen riski α.

3) Lasketaan testisuureen arvo

$\mathbf{t} = \frac{\bar{\mathbf{x}}-\boldsymbol{\mu}_0}{\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}}$ tai ($\mathbf{t} = \frac{\bar{\mathbf{X}}-\boldsymbol{\mu}_0}{\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}\sqrt{\frac{\mathbf{N}-\mathbf{n}}{\mathbf{N}-\mathbf{1}}}}$, kun korjaustekijä tarvitaan.)

4. Vapausasteluku f = n-1 ja merkitsevyystasona α.

<u>2-suuntaisessa testissä</u> t-jakauman taulukosta etsitään

kriittinen raja $t_\alpha^{(2)}(n-1)$ ja

<u>1-suuntaisessa testissä</u> t-jakauman taulukosta etsitään

kriittinen raja $t_\alpha^{(1)}(n-1)$.

5.a) <u>2-suuntaisessa testissä:</u>

Jos laskettu testisuureen arvo

$\mathbf{t} = \frac{\bar{\mathbf{x}}-\boldsymbol{\mu}_0}{\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}} < -\mathbf{t}_{\boldsymbol{\alpha}}^{(2)}(\mathbf{n-1})$ **tai** $\mathbf{t} > \mathbf{t}_{\boldsymbol{\alpha}}^{(2)}(\mathbf{n-1})$ (t ”joutuu jakauman hännälle”),

niin hylkäämisvirheen riski $p < \alpha$,

ja H_0 hylätään.

Muuten H_0 jää voimaan.

1-suuntaisessa testissä

- Kun vastahypoteesina on **H_1:** $\boldsymbol{\mu} < \boldsymbol{\mu}_0$ (testataan jakauman ”vasemmalla hännällä”.)

ja laskettu testisuureen arvo

$\mathbf{t} = \dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} < -\mathbf{t}_{\alpha}^{(1)}\mathbf{(n\text{-}1)}$, niin H_0 hylätään.

- Kun vastahypoteesina on **H_1:** $\boldsymbol{\mu} > \boldsymbol{\mu}_0$ (testataan jakauman ”oikealla hännällä”.)

ja laskettu testisuureen arvo

$\mathbf{t} = \dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} > \mathbf{t}_{\alpha}^{(1)}\mathbf{(n\text{-}1)}$, niin H_0 hylätään.

Tai

5.b) Excelin avulla lasketaan 3):n jälkeen t-jakauman kertymäfunktiosta

- otoksessa havaitun ja vielä sitäkin poikkeavamman tilanteen (”hännän”) todennäköisyys eli hylkäämisvirheen tekemisen riski p täsmällisesti:

$\mathbf{P(\bar{X} \leq \bar{x} \mid \mu = \mu_0) = P\left(t \leq \dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right) = F_{t(n\text{-}1)}\left(\dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right)}$, jos $\bar{x} < \mu_0$,

tai

$$\mathbf{P(\bar{X} \geq \bar{x} \mid \mu = \mu_0) = 1 - P\left(t < \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right) = 1 - F_{t(n-1)}\left(\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right)}, \text{ jos } \bar{x} > \mu_0.$$

Hylkäämisvirheen tekemisen riski p on

- edellä laskettu todennäköisyys, jos voidaan testata 1–suuntaisesti, ja

- 2-suuntaisessa testissä

p = **2**$\cdot$ $P(\bar{X} \geq \bar{x} \mid \mu = \mu_0)$ (tai **2**$\cdot P(\bar{X} \geq \bar{x} \mid \mu = \mu_0)$),

Jos

hylkäämisvirheen riski $\mathbf{p} < \boldsymbol{\alpha}$ = riskille asetettu yläraja,

niin **H_0 uskalletaan hylätä**. Muuten H_0 jää voimaan.

Menetelmää saa käyttää,

- aina, jos perusjoukossa on **X** $\sim\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ ja

- ”aralla mielellä”, kun otoskoko on ”suuri” ($n > 30$), vaikka jakauma perusjoukossa ei olisikaan normaalinen.

Jos otoskoko n on ”suuri” ($n > 30$), voi laskuissa korvata t-jakauman normaalijakaumalla.

Keskimääräistä suuruutta koskevien hypoteesien testaamisesta, kun tutkittavana on kaksi perusjoukkoa

Esim. (jatkoa) 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/l ja keskihajonta $s_1 = 1.2$ mmol/l ja

vertailuryhmän 150:n lumelääkettä saaneen

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

Testattavat hypoteesit ovat

H_0: $\mu_1 = \mu_2$ eli todellinen keskimääräinen kolesterolimäärä on yhtä suuri.

H_1: $\mu_1 \neq \mu_2$ eli keskimääräiset kolesteroliarvot ovat eri suuria

(1- suuntaiseenkin testiin voisi ehkä olla perusteita.)

Tässä ”otosten” informaatio on tuotettu kokeellisen tutkimuksen avulla.

Kun koejärjestely on tehty hyvin tilanne vastaa asetelmaa:

Perusjoukossa E_1	**Perusjoukossa 2**
muuttujan x keskimääräinen suuruus $EX = \mu_1$ hajonta $DX = \sigma_1$	muuttujan x keskimääräinen suuruus $EX = \mu_2$ hajonta $DX = \sigma_2$

↓ ↓

Perusjoukoista poimitaan **toisistaan riippumatta** otokset ja tiedetään,

että sattuman generoi otokset otantajakaumien

$$\bar{X}_1 \sim N\left(\mu_1, \frac{{\sigma_1}^2}{n_1}\right) \quad \text{ja} \quad \bar{X}_2 \sim N\left(\mu_2, \frac{{\sigma_2}^2}{n_2}\right)$$

mukaan, jos jakaumat ovat perusjoukoissa normaalisia tai edes otoskoot ovat ”suuria” (yli 30) .

↓ ↓

Otos 1	**Otos 2**
otoskoko n_1 muuttujan x otoskeskiarvo $\bar{x}_1$ otoshajonta s_1	otoskoko n_2 muuttujan x otoskeskiarvo $\bar{x}_2$ otoshajonta s_2

Todellisten keskiarvojen μ_1 ja μ_2 eron testaaminen perustuu

luonnollisesti otoskeskiarvojen $\bar{x}_1$ ja $\bar{x}_2$ eron tutkimiseen.

Koska otokset poimitaan toisistaan riippumatta, ovat **otoksesta laskettavat keskiarvot $\bar{X}_1$** ja **$\bar{X}_2$ myös riippumattomia**.

Normaalijakauman ominaisuuksien mukaan

"mekanismi" eli otantajakauma, joka tuottaa otokseen otoskeskiarvojen eron suuruuden $\bar{x}_1$ - $\bar{x}_2$ on

$$\bar{X}_1 - \bar{X}_2 = \bar{X}_1 + (-1)\cdot\bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1{}^2}{n_1} + (-1)^2 \cdot \frac{\sigma_2{}^2}{n_2}).$$

Jos H_0: $\mu_1 = \mu_2$ olisi tosi, kuten testauksen lähtökohtana oletetaan,

olisi $\mu_1 - \mu_2 = 0$ ja keskiarvojen eron otantajakauma on

$$\bar{X}_1 - \bar{X}_2 \sim N(0, \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}).$$

Silloin hylkäämisvirheen tekemisen riski voitaisiin laskea normaalijakauman avulla.

Tutkijat sopivat sen ylärajaksi erittäin ankaran kynnysarvon $\alpha = 0.001$.

Tässä "otoksista" saatiin $\bar{x}_1 - \bar{x}_2 = 5.2 - 5.7 = -0.5$ mMol/l.

On selvitettävä, kuinka todennäköistä on, että vähintään tämän suuruinen erotus olisi vain "sattuman leikkiä":

$$\mathbf{P(\bar{X}_1 - \bar{X}_2 \leq -0.5 \mid \mu_1 = \mu_2)}$$

$$= \mathbf{P}\left(\frac{\boldsymbol{(\bar{X}_1 - \bar{X}_2) - 0}}{\sqrt{\frac{\boldsymbol{\sigma_1}^2}{\mathbf{n_1}} + \frac{\boldsymbol{\sigma_2}^2}{\mathbf{n_2}}}} \leq \frac{\mathbf{-0.5 - 0}}{\sqrt{\frac{\boldsymbol{\sigma_1}^2}{\mathbf{n_1}} + \frac{\boldsymbol{\sigma_2}^2}{\mathbf{n_2}}}} \right) = \mathrm{P}\left(\frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{\boldsymbol{\sigma_1}^2}{200} + \frac{\boldsymbol{\sigma_2}^2}{150}}} \leq \frac{-0.5 - 0}{\sqrt{\frac{\boldsymbol{\sigma_1}^2}{200} + \frac{\boldsymbol{\sigma_2}^2}{150}}} \right)$$

$$= \mathrm{P}\left(\mathbf{Z} \leq \frac{-0.5}{\sqrt{\frac{\boldsymbol{\sigma_1}^2}{200} + \frac{\boldsymbol{\sigma_2}^2}{150}}}\right) \ (?)$$

- Todellisia hajontoja σ_1 ja σ_2 ei tietenkään tunneta, samoin kuin edellä oli, ja ne korvataan otoksesta lasketuilla estimaateillaan $s_1 = 1.2$ ja $s_2 = 1.6$.

- Tämä viittaa t-jakauman käyttöön testaamisessa, kuten paras menettely onkin.

- Tällöin eri tapauksia on niin monta, että tässä joudutaan rajoittumaan tilanteeseen, jossa molemmat **otoskoot n_1 ja n_2 ovat ”suuria”** eli yli 30.

Silloin

otoskeskiarvojen eron otoksista tiivistävä otantajakauma on

testisuure $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$ riittävän tarkasti.

Siis edellistä laskua voidaan jatkaa **normaalijakauman** avulla, kun

σ_1 ja σ_2 korvataan otoksesta lasketuilla estimaateillaan $s_1 = 1.2$ ja $s_2 = 1.6$.

$$\mathbf{P(\bar{X}_1 - \bar{X}_2 \leq 5.2 - 5.7 \mid \mu_1 = \mu_2)}$$

$$= \mathbf{P\left(Z \leq \frac{-0.5}{\sqrt{\frac{1.2^2}{200} + \frac{1.6^2}{150}}}\right)}$$

$$= P(Z \leq -3.21) = \phi(-3.21) = 1 - \phi(3.21) = 1 - 0.9993$$

$$= 0.0007.$$

Testaus on **2-suuntainen**, jolloin hylkäämisvirheen riski

p = **2** · $P(\bar{X}_1 - \bar{X}_2 \leq 5.2 - 5.7 \mid \mu_1 = \mu_2)$ **= 2·** 0.0007 = **0.0014** $>$ **0.001** = α.

Havaintoaineisto kyllä puhuu H_0:aa vastaan merkitsevästi, mutta ei kuitenkaan erittäin merkitsevästi, ja testaajat **eivät uskalla hylätä H_0:aa** (vielä tämän) aineiston perusteella.

Huom. Tässä näkyy selvästi vastahypoteesin H_1 asettelun tärkeys.

Jos todella olisi perusteita sulkea koejärjestelyn ulkopuolisen tiedon perusteella pois poikkeama toiseen suuntaan, p-arvo olisi puolta pienempi ja tässä aineiston ”todistusvoima H_0:aa vastaan” riittäisi sen hylkäämiseen.

Yleisestikin testaus etenee samalla tavalla kuin esimerkissä.

- Asetetaan hypoteesit.

- Sovitaan suurin siedettävissä oleva hylkäämisvirheen riski α.

- Lasketaan testisuureen arvon $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ avulla

($P(\bar{X}_1 - \bar{X}_2 \leq \bar{x}_1 - \bar{x}_2 \mid \mu_1 = \mu_2) =$)

$\mathbf{P(Z} \leq \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\mathbf{)}$ (tai $\mathbf{P(Z} \geq \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\mathbf{)}$, jos $\bar{x}_1 > \bar{x}_2$),

joka on p-arvo 1-suuntaisessa testissä.

Jos testi on 2-suuntainen, tämä todennäköisyys vielä kerrotaan 2:lla.

Menettely on approksimatiivinen ja sitä saa käyttää, jos molemmat otoskoot ovat yli 30.

Suhteellisen osuuden testaaminen, kun tutkittavana on kaksi perusjoukkoa

Esim.(jatkoa) Alueen kotitalouksista tehdyssä markkinatutkimuksessa

1500 suuruisessa otoksessa otokseen osuneista kotitalouksista 30 % käyttää hyödykettä Ö.

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?

Vastaavanlainen kampanja on aikaisemmin toiminut hyvin eikä käyttäjien osuus ole koskaan ainakaan pienentynyt. Tämän perusteella ollaan ”varmoja”, että perusteita on riittävästi 1-suuntaiseen testaamiseen.

Testauksessa päätetään käyttää merkitsevyystasona $\alpha = 0.05$.

Hypoteesit ovat

H_0: $\pi_1 = \pi_2$ eli käyttäjien suhteellinen osuus koko perusjoukossa

ennen (π_1) ja jälkeen (π_2) kampanjan on yhtä suuri.

H_1: $\pi_1 < \pi_2$ eli käyttäjien osuus oli pienempi ennen kampanjaa.

Tutkimusasetelmana on:

Perusjoukossa E_1 *kotitalouksissa ennen kampanjaa* Ö: *n käyttäjien* *suhteellinen osuus on* π_1	**Perusjoukossa E_2** *kampanjan jälkeen* Ö: *n käyttäjien* *suhteellinen osuus on* π_2

↓ ↓

Perusjoukoista poimitaan **toisistaan riippumatta** otokset ja tiedetään,

että sattuma tuottaa otokset otantajakaumien

$$\hat{P}_1 \sim N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \quad \text{ja} \quad \hat{P}_2 \sim N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

mukaan.

↓ ↓

Otos 1 otoskoko n_1 Ö: n käytön suhteellinen osuus on $\hat{p}_1$	**Otos 2** otoskoko n_2 Ö: n käytön suhteellinen osuus on $\hat{p}_2$

Todellisten keskiarvojen π_1 ja π_2 eron testaaminen perustuu vastaavasti kuin edellä

otoksista saatujen suhteellisten osuuksien $\hat{p}_1$ ja $\hat{p}_2$ eron tutkimiseen.

Koska otokset poimitaan toisistaan riippumatta, ovat **otoksesta laskettavat suhteelliset osuudet $\hat{P}_1$** ja **$\hat{P}_2$ myös riippumattomia**.

Normaalijakauman ominaisuuksien mukaan

otantajakauma, joka tuottaa otokseen otoskeskiarvojen eron suuruuden $\hat{p}_1$ - $\hat{p}_2$ on

$$\hat{P}_1 - \hat{P}_2 = \hat{P}_1 + (-1)\cdot\hat{P}_2 \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + (-1)^2 \cdot \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

Jos H_0: $\pi_1 = \pi_2$ olisi tosi, kuten testauksen lähtökohtana oletetaan,

olisi

- **$\pi_1 - \pi_2 = 0$** ja
- todellisilla suhteellisilla osuuksilla E_1:ssä ja E_2:ssa on sama tuntematon arvo

$\pi_1 = \pi_2 = \pi$ (?)

ja

keskiarvojen eron otantajakauma on

$$\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_2 \sim \mathbf{N}\left(0, \frac{\boldsymbol{\pi}\,(1-\boldsymbol{\pi}\,)}{\mathbf{n_1}} + \frac{\boldsymbol{\pi}\,(1-\boldsymbol{\pi}\,)}{\mathbf{n_2}}\right) = \mathbf{N}\left(0,\left(\boldsymbol{\pi}(1-\boldsymbol{\pi})\left(\frac{1}{\mathbf{n_1}} + \frac{1}{\mathbf{n_2}}\right)\right)\right)$$

ja esimerkkitilanteessa

$$\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_2 \overset{?}{\sim} \mathbf{N}\left(0,\left(\boldsymbol{\pi}(1-\boldsymbol{\pi})\left(\frac{1}{1500} + \frac{1}{1400}\right)\right)\right)$$

Hylkäämisvirheen riski on

$$\mathbf{p} = \mathbf{P}(\hat{\mathbf{P}}_1 - \hat{\mathbf{P}}_2 \leq \mathbf{0.30\text{-}0.34} \mid \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2)$$

$$= \mathrm{P}\left(\frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{\boldsymbol{\pi}(\mathbf{1} - \boldsymbol{\pi})\left(\frac{1}{1500} + \frac{1}{1400}\right)}} \leq \frac{(0.30 - 0.34) - 0}{\sqrt{\boldsymbol{\pi}(\mathbf{1} - \boldsymbol{\pi})\left(\frac{1}{1500} + \frac{1}{1400}\right)}}\right)$$

$\boldsymbol{\pi}$ on ↑ tuntematon ja se on korvattava estimaatillaan.

Nyt käy samalla tavalla kuin aikaisemmin. Kun $(\hat{P}_1 - \hat{P}_2)$:n otantajakaumassa keskivirhe korvataan estimaatillaan, "siirrytään" normaalijakaumasta t-jakaumaan.

Tässä kuitenkin kumpikin otoskoko on niin "suuri" ($>$ 30), että todennäköisyys voidaan laskea normaalijakauman avulla.

Kun $\boldsymbol{\pi_1} = \boldsymbol{\pi_2} = \boldsymbol{\pi}$, perusjoukot ovat Ö:n käytön osalta samanlaiset ja paras estimaatti $\boldsymbol{\pi}$:lle saadaan **yhdistämällä otosten informaatio**:

Esimerkin tilanteessa

$n_1 = 1500$ ja $\hat{p}_1 = 0.30$, joten käyttäjiä on $1500 \cdot 0.30 = 450$ kotitaloutta ja

$n_2 = 1400$ ja $\hat{p}_1 = 0.34$, joten käyttäjiä on $1400 \cdot 0.34 = 476$ kotitaloutta.

Kotitalouksia on yhteensä 1500+1400 = 2900, ja käyttäjiä 450+476 = 926.

Estimaatti $\boldsymbol{\pi}$:lle on

$\hat{p} = \frac{926}{2900} \approx 0.3193$ $\left(= \frac{1500 \cdot 0.30 + 1400 \cdot 0.34}{1500 + 1400} = \frac{n_1 \cdot \hat{p}_1 + n_2 \cdot \hat{p}_2}{n_1 + n_2}\right)$.

Siis edellisessä

$$\mathbf{p = P(\hat{P}_1 - \hat{P}_2 \leq 0.30\text{-}0.34 \mid \boldsymbol{\pi_1} = \boldsymbol{\pi_2})}$$

$$= P\left(\frac{(\widehat{P}_1 - \widehat{P}_2) - 0}{\sqrt{\mathbf{0.3193(1 - 0.3193)}\left(\frac{1}{1500} + \frac{1}{1400}\right)}} \leq \frac{-0.04 - 0}{\sqrt{\mathbf{0.3193(1 - 0.3193)}\left(\frac{1}{1500} + \frac{1}{1400}\right)}}\right)$$

$\approx$ P(Z $\leq$ - 2.31) = ϕ(-2.31) = 1 - ϕ(2.31) = 1- 0.9896

= 0.0104

p = 0.0104 $<$ 0.05 = α ja H_0 hylätään 5 %:n merkitsevyystasolla.

Suhteellisten osuuksien testaus etenee samalla tavalla kuin esimerkissä:

- Asetetaan hypoteesit.

- Sovitaan suurin siedettävissä oleva hylkäämisvirheen riski α.

- Lasketaan **testisuureen arvo**

$$\mathbf{Z} = \frac{\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2}{\sqrt{\hat{\mathbf{p}}(\mathbf{1} - \hat{\mathbf{p}})\left(\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}\right)}}$$, missä

$\hat{p}_1$ on tutkittavan ominaisuuden E_1:stä poimitussa n_1:n suuruisessa ja

$\hat{p}_2$ on tutkittavan ominaisuuden E_2:stä poimitussa n_2:n suuruisessa

otoksessa ja

$\hat{p}$ on otoksista saatujen suhteellisten osuuksien otoskoon suuruudella painotettu keskiarvo

$$\hat{\mathbf{p}} = \frac{\mathbf{n_1} \cdot \hat{\mathbf{p}}_1 + \mathbf{n_2} \cdot \hat{\mathbf{p}}_2}{\mathbf{n_1} + \mathbf{n_2}} \, .$$

- Testisuureen avulla lasketaan

($P(\hat{P}_1 - \hat{P}_2 \leq 0.30 - 0.34 \mid \pi_1 = \pi_2) =$)

$$\mathbf{P}\left(\mathbf{Z} \leq \frac{\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2}{\sqrt{\hat{\mathbf{p}}(\mathbf{1} - \hat{\mathbf{p}})\left(\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}\right)}}\right) \quad \left(\text{tai } \mathbf{P}\left(\mathbf{Z} \geq \frac{\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2}{\sqrt{\hat{\mathbf{p}}(\mathbf{1} - \hat{\mathbf{p}})\left(\frac{\mathbf{1}}{\mathbf{n_1}} + \frac{\mathbf{1}}{\mathbf{n_2}}\right)}}\right), \text{ jos } \hat{p}_1 \geq \hat{p}_2 \right),$$

joka on p-arvo 1-suuntaisessa testissä.

Jos testi on 2-suuntainen, tämä todennäköisyys vielä kerrotaan 2:lla.

Menettely on approksimatiivinen ja sitä saa käyttää, jos molemmat otoskoot ovat yli 30.

χ^2-riippumattomuustesti

Kun tutkittavien muuttujien mittaustaso on vain luokitteluasteikon taso, riippuvuuden tutkimisen lähtökohtana **2-ulotteinen frekvenssijakauma**.

Esim. (jatkoa) 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?

Nollahypoteesina on tässä (kuten **aina riippuvuutta tutkittaessa**)

H_0: **Muuttujat** sukupuoli ja käyttö ovat **riippumattomia**.

Vastahypoteesi on

H_1: Lisäravinteiden käyttö **riippuu** sukupuolesta.

Kvalitatiivisten muuttujien riippuvuutta voidaan tutkia monen testin avulla. Eräs eniten käytetyistä on χ^2- riippumattomuustesti:

Aluksi oletetaan taas, että **H_0 olisi tosi**.

Sitten tutkitaan millaiselta jakauman **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** (r_1)
Mies (M)	?	?	300 (r_2)
Yhteensä	**150** (s_1)	350 (s_2)	**500** (n)

Todennäköisyyslaskennan tapahtumien riippumattomuuden määritelmän avulla pääteltiin jo aikaisemmin:

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

reunafrekvenssit

↙ ↘

$$\mathbf{e_{11}} = 500 \cdot 0.12 = 500 \cdot P(N\cap K) = 500 \cdot P(N)\cdot P(K) = \cancel{500} \cdot \frac{\mathbf{200}}{\cancel{\mathbf{500}}} \cdot \frac{\mathbf{150}}{\mathbf{500}}$$

otoskoko ↗

$= 60$ tapauksessa.

Samalla tavalla riippumattomuuden voimassa ollessa on

<u>odotettu frekvenssi</u>

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

↙ ↘

$$\mathbf{e_{12}} = \frac{\mathbf{200 \cdot 350}}{\mathbf{500}} = 140$$

↖ otoskoko

- miehille, jotka käyttävät (M∩K) ja miehille, jotka eivät käytä (M∩T)

$$\mathbf{e_{21}} = \frac{\mathbf{200 \cdot 350}}{\mathbf{500}} = 90 \qquad \mathbf{e_{22}} = \frac{\mathbf{300 \cdot 350}}{\mathbf{500}} = 210.$$

(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)

Seuraavaksi 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?

Voidaan osoittaa, että(?)

otoksesta havaitun (o_{ij}) ja riippumattomuuden vallitessa keskimäärin odotettavissa olevan tilanteen (e_{ij}) välisen tilanteen välisen **eron suuruuden**

tiivistää tarkoituksenmukaisella tavalla(?)

χ^2- testisuure $\chi^2 = \Sigma\Sigma \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$,

joka noudattaa (normaalijakaumasta johdettua(?))

χ^2-jakaumaa vapausastein f = (k-1)·(m-1)

(k = luokkien määrä rivien ja m sarakkeiden suunnassa.)

Tässä on f = (2-1)·(2-1) = 1 ja

$$\chi^2 = \frac{(48-60)^2}{60} + \frac{(152-140)^2}{140} + \frac{(102-90)^2}{90} + \frac{(198-210)^2}{210} \approx 5.714.$$

Mitä suurempi χ^2-testisuureen arvo on, sitä enemmän otoksesta havaittu tilanne todistaa H_0:aa vastaan.

Excelistä saadaan(?) **hylkäämisvirheen riski p** eli

todennäköisyys, että H_0:n ollessa voimassa kuitenkin ”vain sattumalta” otoksessa havaittava tilanne ja riippumattomuuden vallitessa keskimäärin odotettu tilanne poikkeavat toisistaan näin paljon eli

taulukoiden erosta tehtävä

”yhteenveto” tulee saamaan otoksessa sattumalta

$\downarrow$ $\downarrow$ näin poikkeavan arvon

$\downarrow$ $\downarrow$

$p = P(\boldsymbol{\chi^2(1)} \geq 5.714) = 1 - P(\boldsymbol{\chi^2(1)} < 5.714) = 1 - \mathbf{F_{\chi^2(1)}(5.714)} = 1 - 0.98317$

≈ 0.017 $\uparrow$

χ^2-jakauman **kertymäfunktio**, kun vapausasteita f = 1.

Siis $p \approx 0.017 < 0.05$ ja H_0 voidaan hylätä 5 %:n merkitsevyystasolla, mutta $p > 0.01$ ja H_0 jää voimaan, jos voidaan sietää vain alle 1 %:n hylkäämisvirheen riski.

Tulos viittaa sukupuolen ja lisäravinteiden käytön riippuvuuteen melkein merkitsevästi, muttei kuitenkaan merkitsevästi.

Kertymäfunktion arvot Excelistä:

Formulas → More Functions → Statistical → CHISQ.DIST

Jos Exceliä ei ole käytössä, voidaan tässäkin testata taulukosta saatavien **kriittisten rajojen** avulla:

- Merkitsevyystasoa $\alpha = 0.05$ ja vapausastelukua $f = (2-1)\cdot(2-1) = 1$

vastaava kriittinen raja $\boldsymbol{\chi^2_{0.05}(1)} = 3.841$ ilmoittaa,

kuinka suuri testisuureen on oltava, jotta hylkäämisvirheen riski $p = 0.05$.

Tässä on $\chi^2 \approx 5.714 > 3.841$, joten $p < 0.05$
ja H_0 hylätään 5 %:n merkitsevyystasolla.

- Jos riskin yläraja $\alpha = 0.01$, niin $\boldsymbol{\chi^2_{0.01}(1)} = 6.635$ ja

$\chi^2 \approx 5.714 < 6.635$. Nyt hylkäämisvirheen riski $p > 0.01$ ja H_0 jää voimaan 1 %:n merkitsevyystasolla.

χ^2-testin kriittisiä rajoja merkitsevyystasoilla $\boldsymbol{\alpha}$

α →	0.10	**0.05**	0.02	**0.01**	0.005	0.001
f ↓						
1	2,706	**3,841**	5,412	**6,635**	7,879	10,828
2	4,605	5,991	7,824	9,210	10,597	13,816
3	6,251	7,815	9,837	11,345	12,838	16,266
4	7,779	9,488	11,668	13,277	14,860	18,467
5	9,236	11,070	13,388	15,086	16,750	20,515
6	10,645	12,592	15,033	16,812	18,548	22,458
7	12,017	14,067	16,622	18,475	20,278	24,322
8	13,362	15,507	18,168	20,090	21,955	26,124
9	14,684	16,919	19,679	21,666	23,589	27,877
10	15,987	18,307	21,161	23,209	25,188	29,588
11	17,275	19,675	22,618	24,725	26,757	31,264
12	18,549	21,026	24,054	26,217	28,300	32,909
13	19,812	22,362	25,472	27,688	29,819	34,528
14	21,064	23,685	26,873	29,141	31,319	36,123
15	22,307	24,996	28,259	30,578	32,801	37,697
16	23,542	26,296	29,633	32,000	34,267	39,252
17	24,769	27,587	30,995	33,409	35,718	40,790
18	25,989	28,869	32,346	34,805	37,156	42,312
19	27,204	30,144	33,687	36,191	38,582	43,820
20	28,412	31,410	35,020	37,566	39,997	45,315

(Formulas → More Functions → Statistical → CHISQ.INV)

Testauksen vaiheet ovat myös yleisesti samat kuin esimerkissä:

- Asetetaan hypoteesit

H_0: Muuttujat x ja y ovat **riippumattomia**.
H_1: Muuttujat x ja y ovat **riippuvat** toisistaan.

- Päätetään, kuinka suuri on suurin siedettävä hylkäämisvirheen riski α.

- Otoksesta havaitusta muuttujien x ja y

yhteisjakaumasta eli arvopareista (x_k, y_k), k = 1, …, n, tehdään

ristiintaulukoimalla **2-ulotteinen frekvenssijakauma**, jossa x:n arvot on jaettu **k luokkaan** ja y:n arvot **m luokkaan**.

- Tästä **havaittujen frekvenssien** $\mathbf{o_{ij}}$ taulukon reunafrekvenssien r_i ja s_j avulla lasketaan vastaavat

odotetut frekvenssit $\mathbf{e_{ij}} = \dfrac{\mathbf{r_i \cdot s_j}}{\mathbf{n}}$.

- Otoksesta havaittujen ja riippumattomuuden vallitessa odotettujen frekvenssien välinen ero tiivistetään

χ^2- testisuureeseen $\boldsymbol{\chi^2} = \sum\sum \frac{(\mathbf{o_{ij}} - \mathbf{e_{ij}})^2}{\mathbf{e_{ij}}}$.

- Vapausasteluku on f = (k-1)·(m-1)

a) Excelin avulla lasketaan hylkäämisvirheen riski

p = P($\boldsymbol{\chi^2}$**(f)** $\geq \chi^2$) = 1- P($\boldsymbol{\chi^2}$**(f)** $< \chi^2$) = 1 – $\mathbf{F}_{\chi^2(f)}(\chi^2)$

↑ ↑

laskettu ↑ arvo kertymäfunktio ↑ , kun vapausasteita f

Jos p $< \alpha$, niin H_0 hylätään. Muuten H_0 jää voimaan.

Tai

b) χ^2-jakauman taulukosta katsotaan

merkitsevyystasoa α ja vapausastelukua f = (k-1)·(m-1)

vastaava kriittinen raja $\boldsymbol{\chi^2_\alpha}$**(f)**.

Jos otoksesta laskettu testisuureen arvo $\boldsymbol{\chi^2} > \boldsymbol{\chi^2_\alpha}$**(f), niin H_0 hylätään**.

6 Regressioanalyysista

Pearsonin korrelaatiokerroin

Kvantitatiivisten muuttujien x ja y välisen riippuvuuden luonne hahmottuu hajontakuviosta:

- Pisteet (x_i,y_i) voivat keskittyä jonkin **suoran** ympärille, jolloin niiden välinen riippuvuus on **lineaarista**,

- tai ne keskittyvät jonkin käyrän ympärille, jolloin riippuvuus on **epälinaarista**.

- Riippuvuus on sitä voimakkaampaa mitä tiiviimpää keskittyminen on.

Tässä rajoitutaan muuttujien lineaarisen riippuvuuden asteen mittaamiseen ja määritellään tunnusluku, joka kuvaa

- kuinka tiiviisti (voimakkaasti) hajontakuvion pisteet keskittyvät pistejoukkoon ”mahdollisimman hyvin sopivan suoran ympärille”.

Esim. Taulukossa ovat viiden opiskelijan

x = laskettujen harjoitustehtävien määrät ja

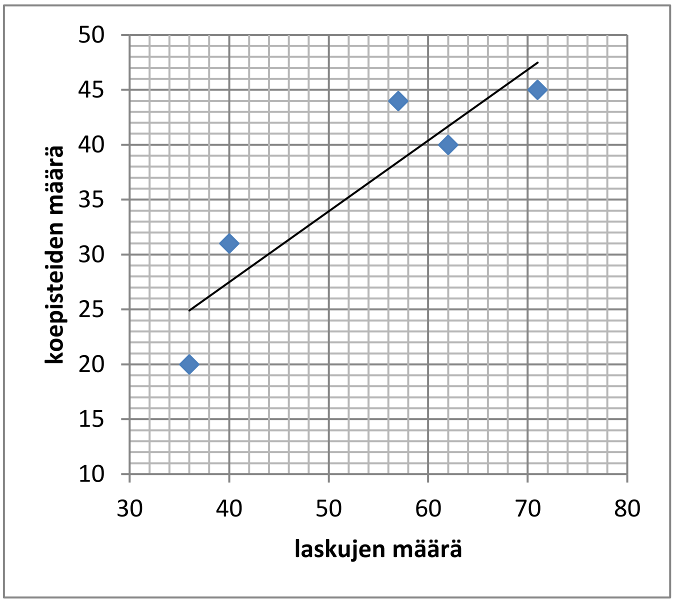
y = koepisteiden määrät

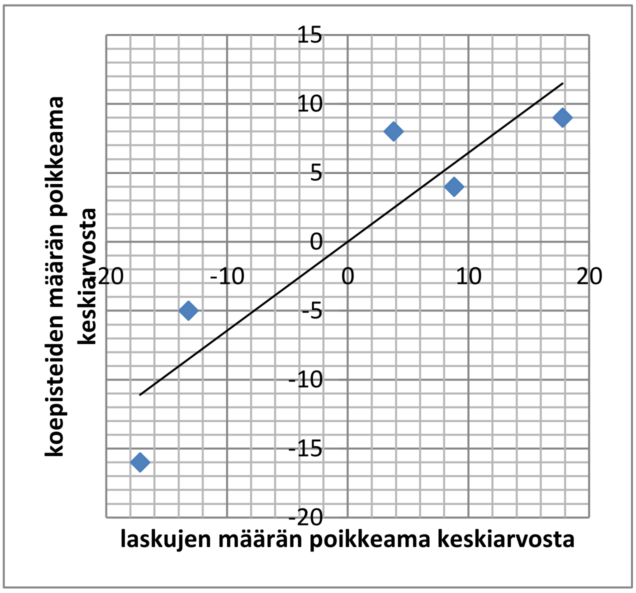
sekä näiden arvojen poikkeamat keskiarvosta eli keskistetyt arvot

$u_i = x_i - \bar{x}$ ja $v_i = y_i - \bar{y}$.

x_i	y_i	u_i	v_i
71	45	17,8	9
62	40	8,8	4
57	44	3,8	8
40	31	-13,2	-5
36	20	-17,2	-16

Alkuperäisistä ja keskistetyistä arvoista piirretyt hajontakuviot ovat:





Kuvioihin on myös piirretty ”niihin parhaiten sopivat” suorat(?).

- Pisteet keskittyvät suorien ympärille yhtä tiiviisti.

- Päämääränä on saada mitatuksi

<u>pisteiden suoran ympärille keskittymisen ”aste”</u>,

missä lähtökohtana voidaan näin käyttää keskistettyjä
arvoja.

- Taulukosta ja keskistettyjä arvoja vastaavasta hajontakuviosta näkyy,

että

arvot $u_i = x_i - \bar{x}$ ja $v_i = y_i - \bar{y}$ ovat ”samansuuntaisia”.

Siis

- kaikkien opiskelijoiden

tehtyjen harjoitustehtävien määrät ja koepisteet

poikkeavat keskiarvosta ” samaan suuntaan”.

Tätä saman- (vastakkais-, eri-) suuntaisuuden astetta mitataan(?)

Pearsonin korrelaatiokertoimen

$$\mathbf{r} = \frac{\sum(\mathbf{x_i}-\bar{\mathbf{x}})(\mathbf{y_i}-\bar{\mathbf{y}})}{\sqrt{\sum(\mathbf{x_i}-\bar{\mathbf{x}})^2\sum(\mathbf{y_i}-\bar{\mathbf{y}})^2}} = \frac{\sum \mathbf{u_i}\mathbf{v_i})}{\sqrt{\sum \mathbf{u_i}^2 \sum \mathbf{v_i}^2}}$$ avulla.

Tämän määrittelyn täsmälliset perustelut saadaan n- ulotteisen reaaliavaruuden $\mathbf{R}^n$ ”geometriasta” ja ne täytyy tässä sivuuttaa.

Esimerkissä on

$$r = \frac{17.8 \cdot 9 + \cdots + (-17.2) \cdot (-16)}{\sqrt{17.8^2 + \cdots + (-17.2)^2}\sqrt{9^2 + \cdots + (-16)^2}} \sim 0.910$$

Pearsonin korrelaatiokerroin mittaa

- kvantitatiivisten muuttujien x ja y
- yhteisjakaumaan (x_1,y_1), (x_2,y_2), …, (x_n,y_n)
- sisältyvää **lineaarista riippuvuutta**.

Määritelmästä voidaan johtaa varsin helposti muita sääntöjä korrelaatiokertoimen laskemiseksi:

Eräs hyödyllinen muoto on

$$r = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$, missä

$\mathbf{C_{xy}} = \frac{\mathbf{1}}{\mathbf{n-1}}\sum(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{y_i} - \bar{\mathbf{y}})$ on muuttujien x ja y välinen **kovarianssi**,

s_x ja s_y ovat keskihajonnat.

Muuttujien x ja y arvojen välisellä

korrelaatiokertoimella ja hajontakuviolla on ominaisuudet:

1. Jos hajontakuvion pisteet (x_i, y_i)

keskittyvät jonkin **nousevan suoran ympärille**,

niin $\mathbf{r > 0}$.

Silloin sanotaan, että muuttujat ovat **positiivisesti korreloituneita**.

2. Jos hajontakuvion pisteet (x_i, y_i)

keskittyvät jonkin **laskevan suoran ympärille**,

niin $\mathbf{r < 0}$.

Silloin sanotaan, että muuttujat ovat **negatiivisesti korreloituneita**.

3. Jos pisteet (x_i,y_i) **eivät keskity minkään suoran ympärille**,

niin $\mathbf{r \approx 0}$.

Silloin sanotaan, että muuttujat ovat (linaarisesti) **korreloimattomia**.

4. Kaikissa aineistoissa on

$\mathbf{-1 \leq r \leq 1}$.

- Erityisesti r = 1 , kun pisteet ovat täsmälleen nousevalla suoralla.
- Kun r = -1 , pisteet ovat täsmälleen laskevalla suoralla.

5. Mitä tiiviimmin pisteet keskittyvät hajontakuviossa siihen ”parhaiten sopivan suoran ympärille ”

eli mitä voimakkaampaa muuttujien välinen korrelaatio on,

sitä suurempi $|r|$ on.

Hajontakuvio kannattaa aina piirtää korrelaatiota tutkittaessa.

- Muuttujien välillä voi olla voimakas epälineaarinen riippuvuus.
- Kuitenkin Pearsonin korrelaatiokerroin, joka mittaa vain lineaarista riippuvuutta voi olla $r \approx 0$.

Korrelaatiokertoimen käsite voidaan yleistää niin, että myös epälineaarista riippuvuutta voidaan mitata.

Otoksesta lasketun korrelaatiokertoimen arvon tulkinnassa on oltava varovainen:

- Pienestä otoksesta laskettu näennäisesti suuri $|r|$:n arvo voi johtua sattumasta.

Kertoimen arvon merkitsevyys on **testattava**.

Testaus perustuu t-jakaumaan:

Hypoteesit ovat:

H_0: $\boldsymbol{\rho}$ = 0 eli muuttujat x ja y ovat korreloimattomia perusjoukossa

H_1: $\boldsymbol{\rho \neq 0}$, jos joudutaan testaamaan 2-suuntaisesti,

ja

H_1: $\boldsymbol{\rho > 0}$ tai **H_1: $\boldsymbol{\rho < 0}$**, jos 1-suuntaiseen testiin ovat riittävät perusteet.

Voidaan osoittaa, että(?) testisuure

$\mathbf{t} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ noudattaa t-jakaumaa vapausastein f = n-2.

Edellisessä äärimmäisen pienessä esimerkkiaineistossa:

H_0: $\boldsymbol{\rho}$ = 0 eli tehtyjen tehtävien määrä ja koepisteet korreloimattomia

H_1: $\boldsymbol{\rho > 0}$

ja testataan 5 %:n merkitsevyystasolla.

Testisuureen arvo on

t = $\frac{0.910\sqrt{5-2}}{\sqrt{1-0.910^2}}$ = 3.801   ja vapausasteluku f = 5-2 = 3.

Excelistä saadaan hylkäämisvirheen riski

p = P($\mathbf{r} \geq 0.910 \mid \boldsymbol{\rho} = \mathbf{0}$)

= P($\mathbf{t(3)} \geq 3.801 \mid \boldsymbol{\rho} = \mathbf{0}$) = 1 - P($\mathbf{t(3)} < 3.801$)

= 1 – 0.984009

= 0.016 $<$ 0.05

eli näin pienikin aineisto tukee tässä jopa lähes merkitsevästi sitä, että muuttujat ovat positiivisesti korreloituneita.

Suuresta otoksesta saatu melko pienikin r:n arvo voi testissä

osoittaa, että muuttujien välillä ”on jotain tekemistä keskenään” perusjoukossa.

Tällöin tulkinta on kuitenkin hyvin epämääräinen:

Muuttujat selittävät toistensa käyttäytymisestä ”jotain”, mutta vain heikosti.

<u>Pienimmän neliösumman suora</u>

Pearsonin korrelaatiokerroin r mittaa,

- kuinka tiiviisti hajontakuvion pisteet (x_i,y_i) keskittyvät

- ” jonkin pistejoukkoon mahdollisimman hyvin sopivan suoran ” ympärille.

- Nyt täsmennetään, mitä suoraa (lineaarista funktiota) silloin tarkoitetaan

- ja miten sen lauseke saadaan selville:

<u>Esim.</u> (jatkoa) Viiden opiskelijan x = laskujen määrä ja y = koepisteet

x_i	y_i
71	45
62	40
57	44
40	31
36	20

- Pisteet keskittyvät hajontakuviossa tiiviisti nousevan suoran ympärille.

- Silloin koepisteiden määrän y ”keskimäärin odotettavissa oleva arvo ”, kun laskuja on x kpl, voidaan ennustaa jonkin(?) lineaarinen funktion f lausekkeen

y = f(x) = a + bx avulla.

On ratkaistava ongelmat:

- Minkä periaatteen mukaan määritellään kuvioon ”mahdollisimman sopiva ” suora?

- Miten x:n ja y:n yhteisjakauman informaatiosta saadaan tätä suoraa vastaavan lineaarisen funktion lauseke?

Suora sopii hyvin hajontakuvioon, kun

- suora on ”lähellä” hajontakuvion pisteitä.

- Silloin suoraa vastaava lineaarinen funktio ennustaa hyvin jokaista

aineistossa olevaa laskujen määrää x vastaavan koepistemäärän y.

Siis

ennusteen $\hat{y}_i = a + bx_i$ ja todella havaitun arvon y_i

väliset poikkeamat (”ennustusvirheet” eli)

residuaalit

$$|d_i| = |y_i - \hat{y}_i| = |y_i - (a + bx_i)|$$

ovat ” pieniä ”.

Tässä menetelmän matemaattinen toimivuus

vaatii(?), että näiden poikkeamien keskimääräistä suuruutta on tarkasteltava kvadraattisen keskiarvon(?) näkökulmasta.

Siis lineaaristen funktioiden $y = a + bx$ kuvaajista hajontakuvioon

sopii parhaiten se, joka tuottaa

pienimmän residuaalien kvadraattisen keskiarvon

Käytännössä a ja b estimoidaan **minimoimalla jäännösneliösumma**

$$\Sigma (y_i - (\mathbf{a} + \mathbf{b}x_i)^2$$

$$\uparrow \quad \uparrow$$

a:n ja b:n suhteen. (x_i ja y_i arvot aineistosta saatuja ”lukuja”)

Jäännösneliösumma on kahden muuttujan funktio a:n ja b:n suhteen ja sen minimoinnissa tarvitaan osittaisderivaattoja(?).

Tämän vuoksi lasku joudutaan (tällä kurssilla) jättämään tässä kesken ja

seuraavassa on (ääriarvotehtävä sivuuttaen) ratkaisu:

Pienimmän neliösumman (pns-) suoran $y = a + bx$

kulmakerroin b ja vakio a lasketaan säännöillä

$$b = \frac{r_{xy}\, s_y}{s_x} \quad \text{ja} \quad a = \bar{y} - b\bar{x},$$

missä $\bar{y}$ ja s_y sekä $\bar{x}$ ja s_x ovat selitettävän y ja selittävän x muuttujan arvojen keskiarvot ja hajonnat ja r_{xy} on Pearsonin korrelaatiokerroin.

Tästä voidaan johtaa monta eri muotoa, eräs helppokäyttöinen on

$$b = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \cdot \Sigma y_i}{n}}{\Sigma x_i{}^2 - \frac{(\Sigma x_i)^2}{n}} \qquad \text{ja } a = \bar{y} - b\bar{x}.$$

Esim. (jatkoa) Viiden opiskelijan

x = harjoitustehtävien määrät ja

y = koepisteiden määrät:

	x_i	y_i	x_i^2	y_i^2	x_iy_i
	71	45	5041	2025	3195
	62	40	3844	1600	2480
	57	44	3249	1936	2508
	40	31	1600	961	1240
	36	20	1296	400	720
Σ	266	180	15030	6922	10143

Summista saadaan

keskiarvot $\bar{x} = 53.2$ ja $\bar{y} = 36.0$ ja hajonnat $s_x = 14.822$ ja $s_y = 10.512$

ja edellä laskettiin $r_{xy} = 0.910$.

$b = \frac{0.910 \cdot 10.512}{14.822} \sim 0.645$ ja $a = 36.0 - 0.645 \cdot 53.2 \sim 1.686$.

Koepisteiden y ja laskettujen laskujen määrän x välisen lineaarisen

yhteyden kuvaa parhaiten (**pns-periaatteen** mukaan)

suora (malli)

$y = 0.645x + 1.686.$

Taulukossa laskettujen neliösummien avulla saadaan b = $\dfrac{10143 - \dfrac{266 \cdot 180}{5}}{15030 - \dfrac{266^2}{5}}$ = 0.645 ja kuten edellä a = 1.686.

Suoran (Ks. edellä) piirtämistä varten tarvitaan kaksi pistettä: (esim.)

x = 30 , y = 0.645·30 + 1.686 = 21.0

x = 70 , y = 0.645·70 + 1.686 = 46.8

Samalla ennustetaan, että

30 laskua laskenut saa keskimäärin 21 koepistettä ja

70 laskua laskenut saa keskimäärin 47 koepistettä.

Jos hajontakuvion pisteet keskittyvät voimakkaasti pns-suoran ympärille,

- malli y = a + bx sopii hyvin aineistoon ja

samalla myös

-Pearsonin korrelaatiokertoimen arvo on suuri.

Siten r:n suuruutta voidaan käyttää

tällaisen (yhden selittävän muuttujan) mallin selitys-kyvyn kuvaamiseen.

Korrelaatiokertoimen neliö $\mathbf{r_{xy}^2}$ ilmaisee mallin y = a + bx

selitysasteen,

joka kertoo, kuinka suuren osan muuttujan y ” vaihtelusta”

(tarkemmin varianssista) muuttuja x selittää mallin avulla.

Yleisessä tapauksessa, kun selittäviä muuttujia on useampia,

selitysaste lasketaan(?)

selitettävän muuttujan y ”kokonaisvaihtelua” kuvaavan neliösumman(?) ja jäännösneliösumman avulla.

Laskettujen laskujen määrän ja koepisteiden määrän yhteyttä kuvaavan

lineaarisen mallin y = 0.645x + 1.686

selitysaste on $r^2 = 0.910^2 \approx 0.83$

Tämä voidaan(?) tulkita niin, että muuttujan x vaihtelu selittää n. 83 % muuttujan y vaihtelusta.

Huom. Tämä ei tarkoita, että ”nyt enää puuttuu 17 %”, ja asia on (kausaalisessa mielessä) täysin selvä.

On mahdollista, että käyttämällä jotakin (joitakin) toisia muuttujia selittäjinä selitysosuus voi olla suurempikin.

Tilanne voi olla:

Taustalla on (tuntematon)

todellinen ”syy”

↓

z

z ”määrää” x:n. z määrää myös v:n

malli 1. ↙ ↘ malli 2.

x v

x:llä selitetään y:tä. v:llä selitetään y:tä.

↘ ↙

y

selitettävä muuttuja

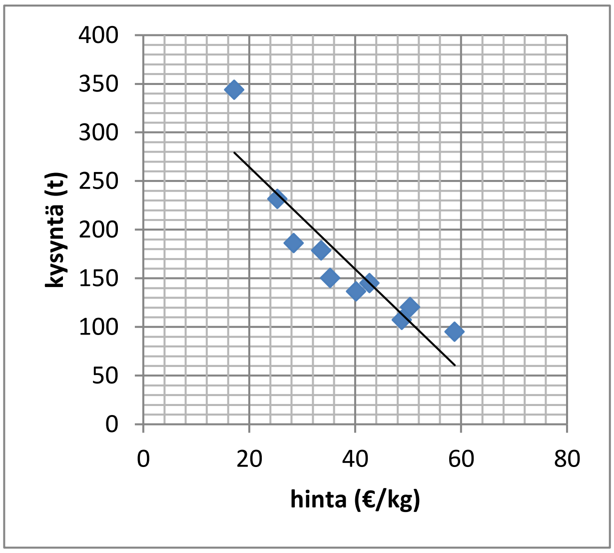
Korkea selitysaste <u>ei ratkaise</u> kausaalisuus-ongelmaa.

Regressiokertoimen (-ien) tilastollinen merkitsevyys voidaan myös **testata.**

Yhden selittäjän tapauksessa testi (kulma-)kertoimelle b on **sama kuin korrelaatiokertoimen r testi**. (Ks. edellä)

<u>Esim.</u> Ekonomisti E tutki herkkumatikan kysyntää markkinoilla ja sai(?) havainnot keskimääräisistä yksikköhinnoista x (€/kg) ja niitä vastaavista kysynnän määristä y (t):

x_i	y_i
40,2	136,5
50,4	120,6
17,2	343,8
33,6	178,8
58,8	95,2
42,7	145,3
48,8	107,4
28,4	186,3
35,3	150,4
25,3	231,7

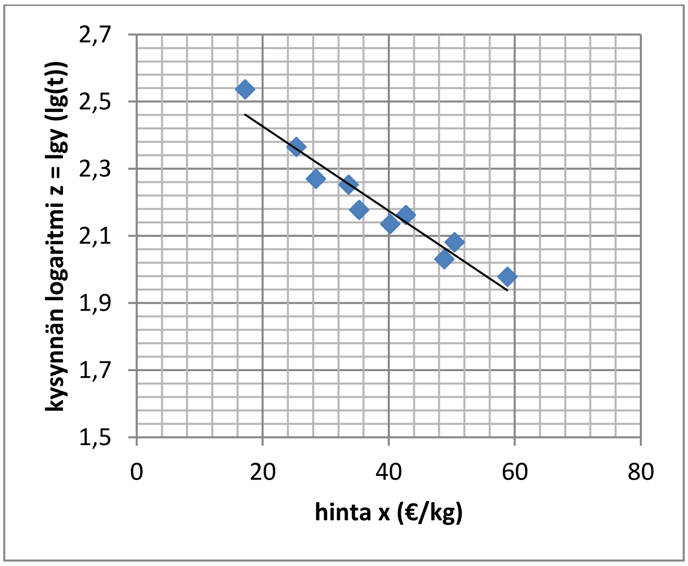


Lineaarinen sovite on melko hyvä, mutta

- alhaisilla yksikköhinnoilla kysyntä vähenee lineaarista voimakkaammin
- ja korkeilla taas hitaammin.
- ”Oikeaa” ratkaisua tällaisten muuttujien yhteyden (kuluttajien joukkokäyttäytymisen) mallintamiseen ei ole.

- Toimiva arvaus voisi tässä olla jokin vähenevä eksponenttifunktio,
jolloin logaritmi-muunnos linearisoisi hajontakuviosta näkyvän riippuvuuden:

Herkkumatikan kysyntä logaritmisena



Yhteys ei linearisoidu täysin, mutta yhteensopivuus

on parempi.

Arvot ovat seuraavassa taulukossa.

x_i	y_i	$z_i=\lg y_i$	x_i^2	$x_i z_i$
40,2	136,5	**2,135133**	1616,04	85,83233
50,4	120,6	**2,081347**	2540,16	104,8999
17,2	343,8	**2,536306**	295,84	43,62446
33,6	178,8	**2,252368**	1128,96	75,67955
58,8	95,2	**1,978637**	3457,44	116,3439
42,7	145,3	**2,162266**	1823,29	92,32874
48,8	107,4	**2,031004**	2381,44	99,11301
28,4	186,3	**2,270213**	806,56	64,47405
35,3	150,4	**2,177248**	1246,09	76,85685
25,3	231,7	**2,364926**	640,09	59,83263
380,7		21,98945	15935,91	818,9854

Arvopareihin (x_i,z_i) sovitettu malli z = a + bx :

$$b = \frac{818.9854 - \frac{380.7 \cdot 21.98945}{10}}{15935.91 - \frac{380.7^2}{10}} = -0.012583$$

a = 2.198945 – (-0.012583)·38.07 = 2.67798

Siis malli on logaritmisena z = -0.012583x + 2.67798.

”Takaisin” päästään vastaavalla eksponenttifunktiolla

y = $10^{-0.012583\mathrm{x} + 2.67798}$

= $10^{-0.012583\mathrm{x}} \cdot 10^{\,2.67798}$

= $(10^{-0.012583})^{\mathrm{x}} \cdot 10^{2.67798}$

= **476.4 ·** $\mathbf{0.971^{x}}$

Mallista nähdään, että

- kysyntä on 476.4 t, jos herkkumatikka on ilmaista (!!!),

- ja kysyntä pienenee kaikilla hintatasoilla x keskimäärin 2.9 %, kun hinta nousee 1 €/kg.