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: μ = 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 α = 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

$$\boldsymbol{\sigma}\ \textbf{?}$$

$\bar{X} \sim N(100\ g, \frac{(\mathbf{4\ g})^2}{50}) = N(100\ g, 0.32\ g^2) = N(100\ g, (0.5657\ g)^2)$ 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 $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4\ g}{\sqrt{50}}$ korvataan estimaatillaan $\frac{s}{\sqrt{n}} = \frac{3.8\ g}{\sqrt{50}}$

Siis

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

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

↑

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(50-1) ≤ -2.791 tai t(49) ≥ 2.791) ja lyhyemmin merkittynä

= P(t ≤ -2.791 tai t ≥ 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·(1 – $\mathbf{F_{t(49)}(2.791)}$)

↖ Excelistä **0.996266** (Vrt. ϕ(2.79) = 0.9974.)

≈ 2·0.004

= 0.008

Koska

- hylkäämisvirheen tekemisen riski **p** ≈ **0.008** < **0.01 = α**,

- 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

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

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

↑

↓

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

….

= $0.008 < 0.01 = \alpha$

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

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 $p = 0.008 < 0.01 = \alpha$.

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 $t = \frac{98.5-100}{\frac{3.8}{\sqrt{50}}}$ = -2.791.

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

t-jakauman taulukossa

- vapausasteluku f ≈ 50 määrää rivin ja

- otsikon ”2 – suuntaisen testin kriittisiä rajoja” alta $\alpha = 0.01$ 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 $p = 0.01 = \alpha$.

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 $\alpha = 0.01$, jolloin luottamustaso $c = 1 - 0.01 = 0.99$.

Vapausasteluku f = 50 – 1 = 49 ≈ 50.

t-jakauman taulukosta saadaan (samasta kohdasta kuin äsken)

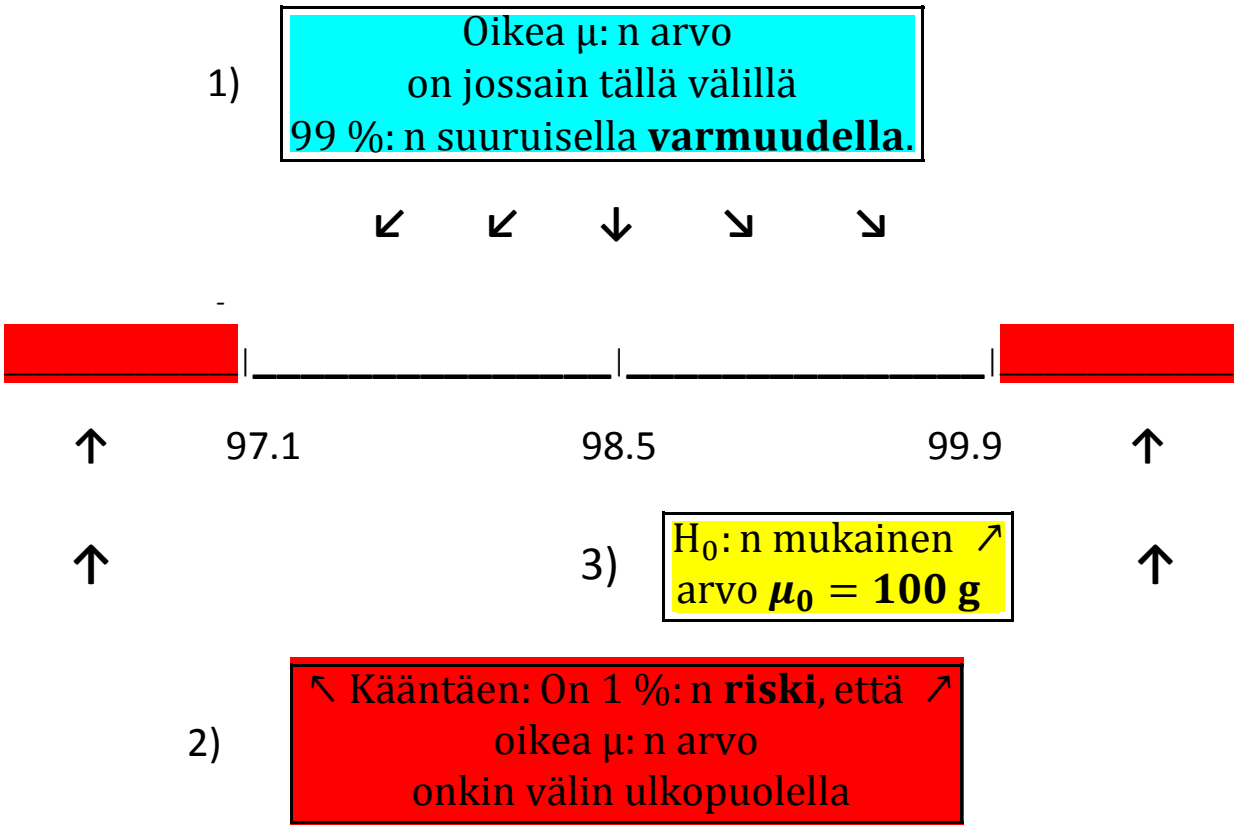
luottamuskerroin $t_{0.99}(50) = 2.678$.

99 %:n luottamusvälin 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:



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 $t^{(1)}_{0.001}(19)$ = 3.579.

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

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 tarkasti:

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 → $\mathbf{t_{\alpha}^{(2)}(f)}$ ja vastaavasti $\mathbf{t_{\alpha}^{(1)}(f)}$

merkitsevyystasolla ↑ ↖vapausastein”

Jos **X** $\sim$**N(μ, σ^2) perusjoukossa** ja **H_0: $\mu = \mu_0$** 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>:

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) \qquad \bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}\right)$$

↑ ↑

”Spekuloidaan” sillä, että μ_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 $\mathbf{t} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ tai $\mathbf{t} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$,

otoksesta estimoitu keskivirhe

t-jakautunut vapausastein f = n-1.

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

Sen arvo mittaa, kuinka paljon otoksesta laskettu keskiarvo $\bar{x}$ poikkeaa H_0:ssa oletetusta todellisesta keskiarvosta μ_0 otannassa tapahtuvan ”luonnollisen vaihtelun keskimääräisen suuruuden puitteissa”, jonka suuruus $\frac{s}{\sqrt{n}}$ estimoidaan otoksesta.

Todellisen keskimääräisen suuruuden μ testaamisen vaiheet

ovat samat kuin esimerkeissä:

- Testausmenetelmä toimii täsmälleen oikein, kun tutkittavan muuttujan arvojen jakauma perusjoukossa on **X** $\sim$**N(μ, σ^2).**

- **n:n suuruisesta otoksesta** lasketaan

(μ:n estimaatti) **otoskeskiarvo $\bar{x}$** ja (σ:n estimaatti) **otoskeskihajonta s**.

1) Asetetaan hypoteesit:

Nollahypoteesi **H_0: $\mu = \mu_0$,**

jossa μ_0 on jokin oletettu, väitetty, aikaisempi, tms. arvo.

Vastahypoteesi **H_1: $\mu \neq \mu_0$,**

jos **joudutaan käyttämään** 2-suuntaista testiä,

ja **H_1: $\mu < \mu_0$** tai **H_1: $\mu > \mu_0$,**

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, eli millä ”**merkitsevyystasolla**” testataan.

3) Lasketaan **testisuureen arvo**

$$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} \quad \text{tai} \quad t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$, kun korjaustekijä tarvitaan.

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

2-suuntaisessa testissä t-jakauman taulukosta etsitään

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

1-suuntaisessa testissä t-jakauman taulukosta etsitään

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

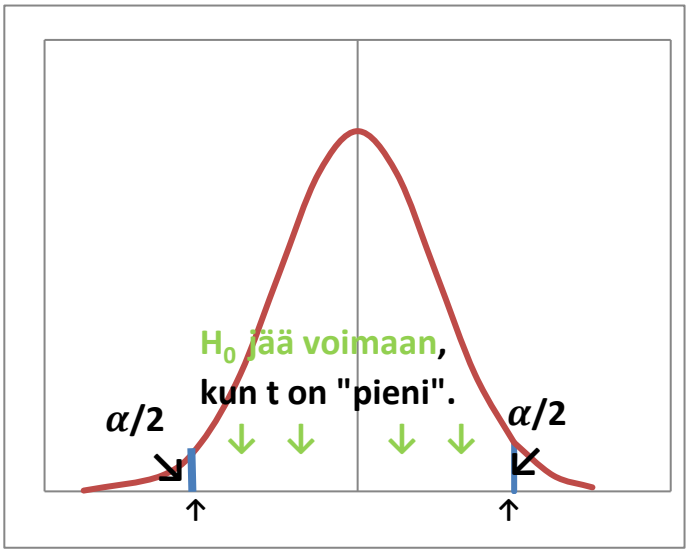
5.a) **2-suuntaisessa testissä**:

Jos laskettu testisuureen arvo

$\mathbf{t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} < -t_\alpha^{(2)}(n-1)}$ **tai** $\mathbf{t > t_\alpha^{(2)}(n-1)}$ (t ”joutuu jakauman hännälle”),

niin hylkäämisvirheen riski $p < \alpha$,

ja H_0 hylätään. Muuten H_0 jää voimaan.



Kun $t < -t_{\alpha}^{(2)}$ tai $t_{\alpha}^{(2)} < t$ H_0 hylätään.

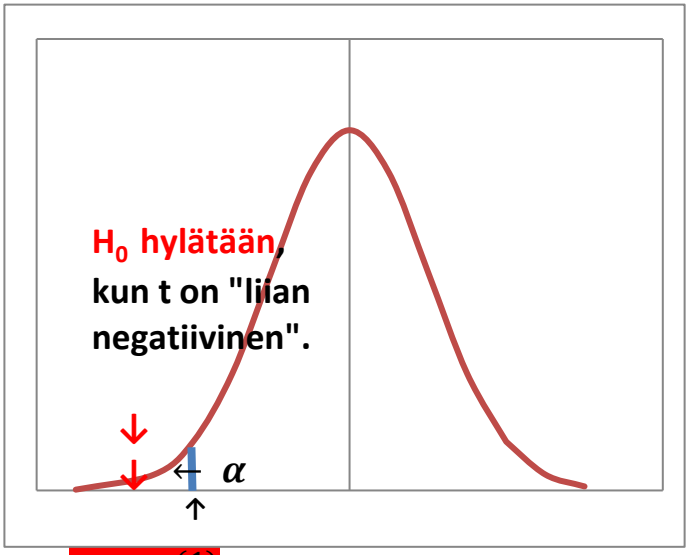
1-suuntaisessa testissä

- Kun vastahypoteesina on $H_1: \mu < \mu_0$ (testataan jakauman ”vasemmalla hännällä”.)

ja laskettu testisuureen arvo

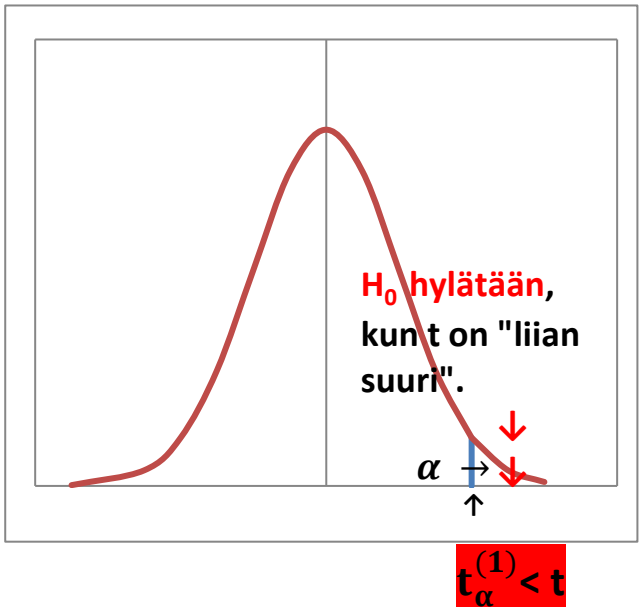
$$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} < -t_{\alpha}^{(1)}(n-1),$$

niin H_0 hylätään.



$t < -t_{\alpha}^{(1)}$

- Kun vastahypoteesina on $\mathbf{H_1: \mu > \mu_0}$ (testataan jakauman ”oikealla hännällä”.)



ja laskettu testisuureen arvo

$$\mathbf{t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_\alpha^{(1)}(n\text{-}1)},$$

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:

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

tai

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

Hylkäämisvirheen tekemisen riski p on

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

- 2-suuntaisessa testissä

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

Jos

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

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

Menetelmää saa käyttää,

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

- menetelmä toimii riittävän hyvin, 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.

Esim. Terveystuotteen tuoteselosteen mukaan tuotteissa on keskimäärin 400 mg Woima-ainetta W.

Viimeisen käyttöpäivänsä ylittäneistä tuotteista poimittiin 22 suuruinen otos, josta mitattiin woima-aineen W pitoisuudet. Keskiarvoksi saatiin 354 mg ja hajonnaksi 96 mg ja jakauma näytti normaaliselta.

Testaa 5 %:n merkitsevyystasolla, onko todellinen keskimääräinen W:n määrä μ on pienentynyt vanhentuneissa tuotteissa?

H_0: μ = 400 mg

H_1: μ < 400 mg

$$t = \frac{354 - 400}{\frac{96}{\sqrt{22}}} = -2.247$$

f = 22-1 = 21 ja $\alpha = 0.05$

ja taulukosta saadaan kriittinen raja $t_{0.05}^{(1)}(21) = 1.721$.

Koska testisuureen arvo **t = -2.247 < -1.721,**

niin on **hylkäämisvirheen riski p < 0.05** ja H_0 voidaan hylätä 5 %:n merkitsevyystasolla.

(Excelistä: $p = P(\mathbf{t} \leq -2.247) = F_{t(21)}(-2.247) = 0.017759 < 0.05 = \alpha$)

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	muuttujan x
keskimääräinen suuruus $EX = \mu_1$	keskimääräinen suuruus $EX = \mu_2$
hajonta $DX = \sigma_1$	hajonta $DX = \sigma_2$

↓ ↓

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

että sattuman generoi otokset otantajakaumien

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

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

↓ ↓

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

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

Jos H_0: $\mu_1 = \mu_2$ olisi tosi, kuten testauksen lähtökohtana oletetaan, olisi $\mu_1 - \mu_2 = 0$ ja keskiarvojen eron otantajakauma on

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

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”:

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

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

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

- 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 useita, että tässä rajoitutaan tilanteeseen, jossa molemmat **otoskoot n_1 ja n_2 ovat ”suuria”** eli yli 30.

Silloin

otoskeskiarvojen eron otoksista tiivistävä otantajakauma on

testisuure Z = $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ $\sim$ 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 $\boldsymbol{Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}}}}$ avulla

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

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

joka on p-arvo 1-suuntaisessa testissä.

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

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

Esim. Palvelualalla P vertailtiin osa-aikaisten ja täyspäivätyötä tekevien tuntiansioita (€/h).

Poimituissa otoksissa olivat tulokset:

	otoskoko	keskiarvo	hajonta
Täyspäiväinen	47	12.8	2.4
Osapäiväinen	52	11.4	3.8

Alaan P hyvin perehtyneet tutkimuksen tekijät ovat ”varmoja”, ettei ero palkoissa ole ainakaan osapäiväisten eduksi, jos sitä on.

Tässä testataan 1 %:n merkitsevyystasolla, ovatko alalla P täyspäivätyötä tekevien todelliset keskimääräiset tuntiansiot suuremmat kuin osapäiväisillä.

:

Testattavat hypoteesit ovat:

H_0: $\mu_1 = \mu_2$ (todelliset keskipalkat yhtä suuria.)

H_1: $\mu_1 > \mu_2$ (täyspäiväisten keskipalkka suurempi.)

Testisuureen arvo on

$$Z = \frac{12.8 - 11.4}{\sqrt{\frac{2.4^2}{47} + \frac{3.8^2}{52}}} = 2.21$$

Hylkäämisvirheen tekemisen riski on

($p = P(\bar{X}_1 - \bar{X}_2 \geq 12.8 - 11.4 \mid \mu_1 = \mu_2)$, 1-suuntainen testi)

tässä

$p = P(Z \geq 2.21) = 1 - P(Z < 2.21) = 1 - \phi(2.21 = 1 - 0.9826$

$= 0.0174.$

Koska $p = 0.0174 > 0.01 = \alpha$, ei otoksen informaation perusteella ole riittäviä perusteita hylätä H_0 :aa 1 %:n merkitsevyystasolla.

(Jos näin tehdään, erehtymisen riskin suuruus on 1.74 %.)

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 <u>pienempi</u> ennen kampanjaa.

Tutkimusasetelmana on:

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

↓ ↓

Perusjoukoista poimitaan **toisistaan riippumatta** otokset ja tiedetään, että sattuma tuottaa otokset otantajakaumien

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

mukaan.

↓ ↓

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

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

ja esimerkkitilanteessa

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

(? ?)

Hylkäämisvirheen riski on

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

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

π 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 $\boldsymbol{\pi_1 = \pi_2 = \pi}$, 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

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

Siis edellisessä

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

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

$\approx P(Z \leq -2.31) = \phi(-2.31) = 1 - \phi(2.31) = 1 - 0.9896 = 0.0104$

$p = 0.0104 < 0.05 = \alpha$ 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**

$$\boldsymbol{Z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}+\frac{\hat{p}(1-\hat{p})}{n_2}}}} \quad \left(= \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right.$$, kuten yleensä oppikirjoissa)

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

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

- Testisuureen avulla lasketaan

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

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

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.

Huom. Tässä, toisin kuin kahden perusjoukon keskiarvotestissä edellisessä kappaleessa, t-jakauma on suoraan käytettävissä p-arvon laskemiseen.

Vapausasteluku on silloin $f = n_1 + n_2 - 2$ ja normaalijakauman kertymäfunktion sijasta käytetään t-jakauman kertymäfunktiota.

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**, siis

- 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

↙ ↘

$$\mathbf{e_{11}} = 500 \cdot 0.12 = 500 \cdot P(N \cap K) = 500 \cdot P(N) \cdot P(K) = \cancel{500} \cdot \frac{\mathbf{200}}{\cancel{500}} \cdot \frac{\mathbf{150}}{\mathbf{500}} = \frac{\mathbf{r_1 \cdot s_1}}{\mathbf{n}}$$

otoskoko ↗

= 60 tapauksessa.

Samalla tavalla riippumattomuuden voimassa ollessa on

odotettu frekvenssi

- naisille, jotka eivät käytä (N∩T) $\mathbf{e_{12}} = \frac{\mathbf{200 \cdot 350}}{\mathbf{500}} = 140$

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

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

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

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” $\chi^2(1)$ tulee saamaan otoksessa ihan vain sattumalta näin paljon nollasta poikkeavan arvon

↓ ↓

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

≈ 0.017 ↑

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)·(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 o_{ij}** taulukon reunafrekvenssien r_i ja s_j avulla lasketaan vastaavat

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

- Otoksesta havaittujen ja riippumattomuuden vallitessa odotettujen frekvenssien välinen ero tiivistetään

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

- Vapausasteluku on f = (k-1)·(m-1)

a) Excelin avulla lasketaan hylkäämisvirheen riski

p = P(χ^2**(f)** $\geq \chi^2$) = 1- P(χ^2**(f)** $< \chi^2$) = 1 – $F_{\chi^2(f)}(\chi^2)$

laskettu ↑ arvo kertymäfunktio ↑ , kun vapausasteita f

Jos **p** $<$ **α**, 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 χ^2_α**(f)**.

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

Käyttöedellytyksinä testille ovat:

Kaikkien odotettujen frekvenssien on oltava suurempia kuin 1 ja

korkeintaan 20% odotetuista frekvensseistä saa olla pienempiä kuin 5.

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?

Hypoteesit ovat

H_0: Suhtautuminen tulospalkkaukseen on **riippumatonta** iästä.
H_1: Suhtautuminen tulospalkkaukseen **riippuu** iästä.

Hylkäämisvirheen riskin maksimiarvoksi asetetaan tässä $\alpha = 5\%$.

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

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

(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

$$\chi^2 = \frac{(96-111.03)^2}{111.03} + \frac{(174-171.50)^2}{171.50} + \ldots + \frac{(122-134.52)^2}{134.52} \approx 6.561$$

Vapausasteluku f = (2-1)·(3-1) = 2

Koska $\boldsymbol{\chi^2 \approx 6.561 > \chi^2_{0.05}(2) = 5.991}$

niin hylkäämisvirheen tekemisen riski **p < 0.05,**

kun H_0 hylätään.

Pearsonin korrelaatiokertoimen testaus on seuraavassa luvussa.

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

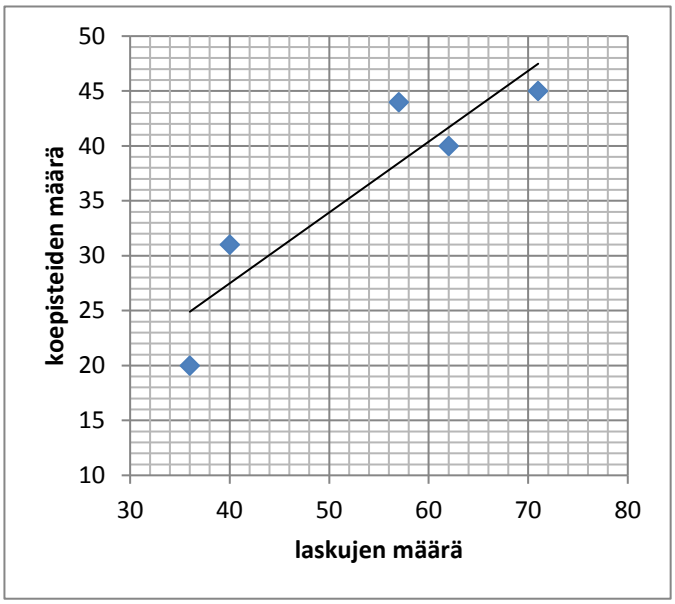
y = koepisteiden määrät

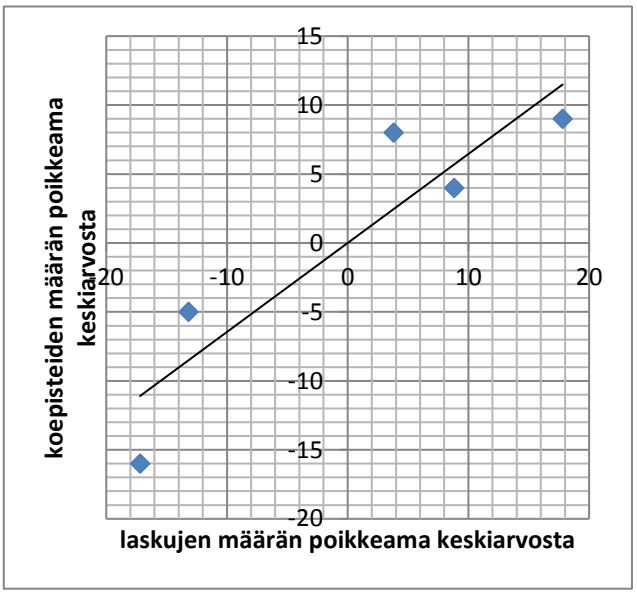
sekä näiden arvojen poikkeamat keskiarvosta eli **keskistetyt arvot**

$u_i = x_i - \bar{x}$ 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

pisteiden suoran ympärille keskittymisen ”aste”,

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

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

$\mathbf{C_{xy}} = \frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})$ 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: $\rho = 0$ eli muuttujat x ja y ovat korreloimattomia perusjoukossa

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

ja

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

Voidaan osoittaa, että testisuure

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

Edellisessä äärimmäisen pienessä esimerkkiaineistossa:

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

H_1: $\rho > 0$

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.

Pienimmän neliösumman suora

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:

Esim. (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

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

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**

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

↑ ↑

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 tässä jättämään tässä kesken ja

seuraavassa on (ääriarvotehtävä sivuuttaen) **ratkaisu:**

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

kulmakerroin b ja vakio a lasketaan säännöillä

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

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

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

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

x = harjoitustehtävien määrät ja

y = koepisteiden määrät:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	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

$$b = \frac{10143 - \frac{266 \cdot 180}{5}}{15030 - \frac{266^2}{5}}$$ = 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 (selitysosuuden),

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 **ei ratkaise** kausaalisuus-ongelmaa.

Regressiokertoimen (-ien) tilastollinen merkitsevyys voidaan myös

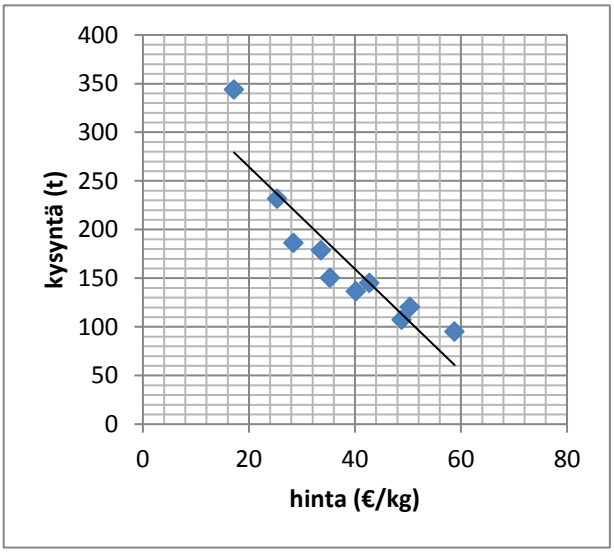
testata.

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

Esim. 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

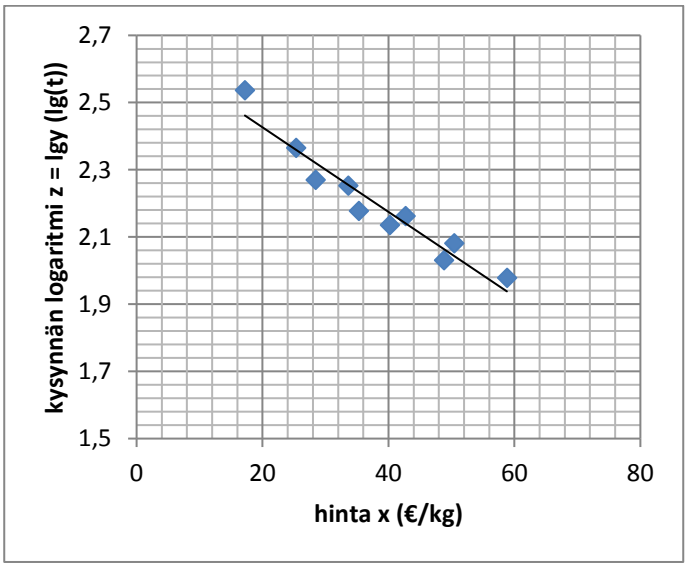


Lineaarinen sovite on melko hyvä, mutta

- 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:

<u>Herkkumatikan kysyntä logaritmisena</u>



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_iz_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

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

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

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

$= \mathbf{476.4 \cdot 0.971^{x}}$

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.