6. Todennäköisyys ja satunnaisuus

Johdatus yhteiskuntatilastotieteeseen, osa 2

Kurssi on jaettu kahteen erilliseen osaan. Molemmissa on viisi teemaa.

Jos et ole suorittanut osaa 1, vaan tullut suoraan mukaan tässä vaiheessa, on hyvä vilkaista osaa 1, koska siihen viitataan toisinaan osan 2 kuluessa.

Kuva: Aino Huovio | www.helsinkidesignweek.com/series/pechakucha-night-29-march-meet-the-speakers/

Kimmo Vehkalahti
Vanhempi yliopistonlehtori, Soveltavan tilastotieteen dosentti
Yhteiskuntadatatieteen keskus (Centre for Social Data Science)
Valtiotieteellinen tiedekunta, Helsingin yliopisto

Teemat 6–10 (sekä Teemojen 4 ja 5 "sytykeosuudet") perustuvat (osin) oppikirjani **Kyselytutkimuksen mittarit ja menetelmät** (Helsingin yliopisto 2019 / Finn Lectura 2014 / Tammi 2008) lukuihin 4–7.

Kurssin puoliväli* ja Osan 2 teemat

Osa 1 keskittyy **mittaukseen**, **tiedonkeruuseen** ja **kuvailevaan tilastotieteeseen** yleisotsikkonaan **tilastollinen lukutaito**.

Osassa 2 painottuu **tilastollinen päättely**, joka puolestaan rakentuu voimakkaasti **todennäköisyys**-käsitteen varaan.

Osa 2 ei ole erillinen kokonaisuus vaan se kietoutuu monilta kohdin osan 1 teemoihin. Tämä vastaa käytäntöä, sillä tilastollisessa päättelyssä ei ole mitään mieltä ilman aineiston perusteellista kuvailua.

* **Osan 2 voi suorittaa erikseen myöhemminkin, mutta yleensä on suositeltavinta tehdä se suoraan osan 1 jatkoksi. Useiden muiden kurssien esitietovaatimuksena ovat molemmat osat.**

Teema 6: Todennäköisyys ja satunnaisuus

Käsite **todennäköisyys** kohdataan päivittäin mm. sääennusteissa, uutisraporteissa, peliarvonnoissa, tulevaisuuden suunnitelmissa jne. *Mitä todennäköisyys tarkoittaa ja miten sitä pitäisi tulkita?*

Todennäköisyys voidaan määritellä matemaattisena käsitteenä, mutta tässä keskitytään vain seuraavien käsitteiden tulkintoihin:

- **objektiivinen** todennäköisyys
- **subjektiivinen** todennäköisyys
- **klassinen** todennäköisyys

Tulkinnasta riippumatta todennäköisyydelle on ominaista, että eri vaihtoehtojen toteutuminen on luonteeltaan **satunnaista**, ts. **sattuma** määrää, mikä vaihtoehto kulloinkin toteutuu. Tällaista tapahtumaa kutsutaan **satunnaisilmiöksi**.

Lähteitä ja ''todennäköistä'' oheislukemistoa

- **Aalto-yliopisto:** Todennäköisyyslaskennan ja tilastotieteen peruskurssi
- **Vaasan yliopisto (Tommi Sottinen)**: Päätöksenteko epävarmuuden vallitessa
- **New York University (Nassim Nicholas Taleb)**:
- The Black Swan: The Impact of the Highly Improbable
- Antifragile: Things That Gain from Disorder
- (suom.) Musta joutsen: Erittäin epätodennäköisen vaikutus
- (suom.) Antihauras: Asioita, jotka hyötyvät epäjärjestyksestä

1. Objektiivinen todennäköisyys

Tapahtuman suhteellinen esiintymisfrekvenssi pitkässä, riippumattomassa koesarjassa

Tarkastellaan tilannetta, jossa satunnaisilmiö esiintyy useita kertoja. Vaihtoehdon todennäköisyydeksi määritellään sen **esiintymiskertojen suhteellinen frekvenssi**.

Esimerkki: rahanheitto – vaihtoehdot *kruuna* tai *klaava*

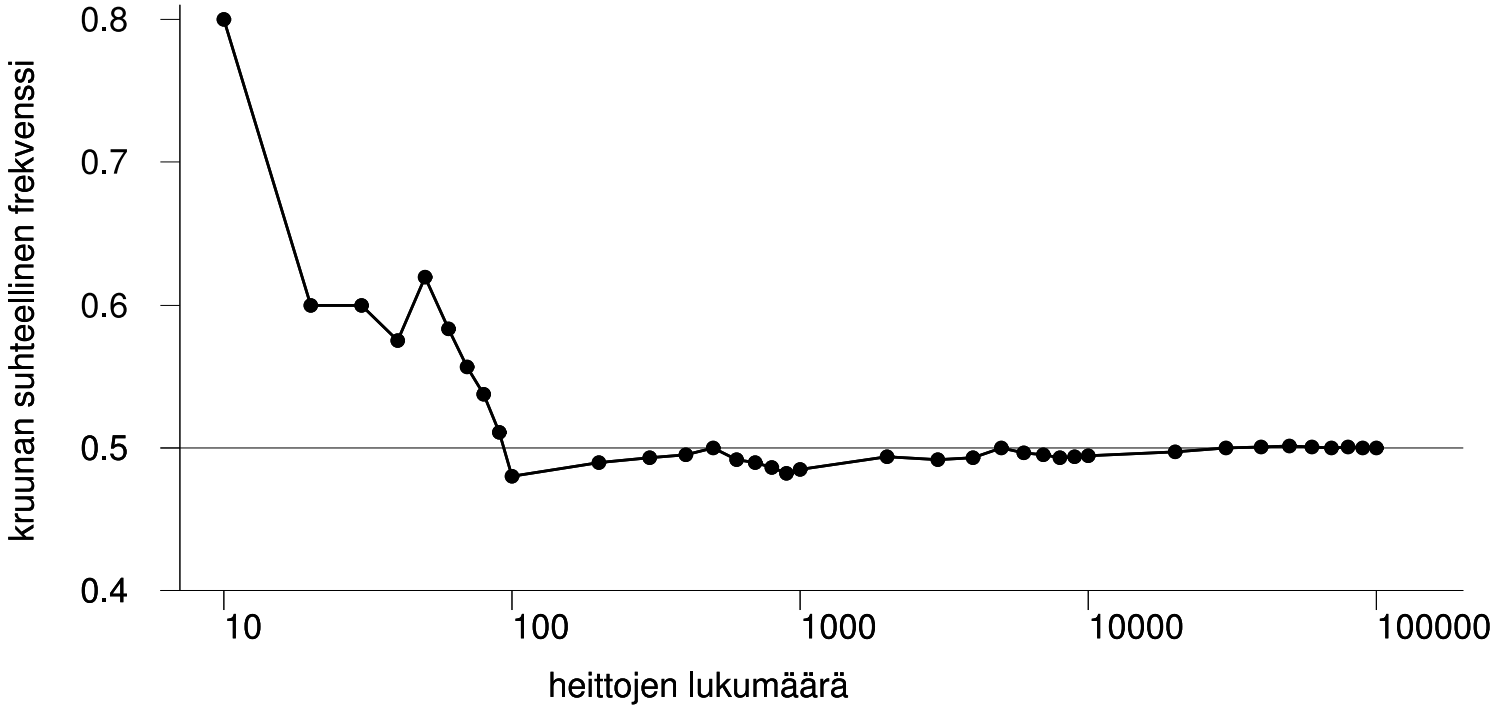
Historiaa (koesarjoja oikeilla kolikoilla):

- Comte de Buffon (1700-luvulla)
- 4 040 heittoa, 2 048 kruunaa ($\frac{2048}{4040} = 0.5069$)
- Karl Pearson (1800-luvulla)
- 24 000 heittoa, 12 012 kruunaa ($\frac{12012}{24000} = 0.5005$)
- John Kerrich (1900-luvulla)
- 10 000 heittoa, 5 067 kruunaa ($\frac{5067}{10000} = 0.5067$)

Odotettu arvo on $\frac{1}{2} = 0.5$, kunhan raha on *harhaton*. (Huomaa: yhdessäkään kokeessa suhteellinen frekvenssi ei ole tasan 0.5).

Rahanheittokoe Survolla: 100 000 heittoa, 49 996 kruunaa

''Heitot'' perustuvat (pseudo)satunnaislukugeneraattorin käyttöön, jossa aitoa satunnaisuutta jäljitellään täysin deterministisesti:



- **Simulointikokeissa** on luotava mahdollisimman **satunnaisia** lukusarjoja.
- Toisaalta tieteelliset kokeet on voitava **toistaa** täsmälleen **samanlaisina**.

Vaikuttaako ristiriitaiselta? Sattuman jäljittely ei ole helppoa!

Objektiivinen todennäköisyys: johtopäätöksiä

Edellä olevissa kokeissa poikkeamat ennakko-odotusten mukaisesta arvosta $\frac{1}{2}$ ovat tulkittavissa **satunnaisvaihteluksi**. Poikkeamien merkittävyyttä voidaan testata tilastollisesti (ks. Teema 9).

Kurssin alussa perehdyttiin mittaukseen ja tiedonkeruuseen, jotka molemmat tuovat tilastolliseen tutkimukseen **epävarmuuksia**.
Osa epävarmuuksista on luonteeltaan satunnaisvaihtelua.

Tilastollinen päättely edellyttää, että satunnaisilmiö noudattaa todennäköisyyden lakeja. Eri vaihtoehtoihin on voitava liittää todennäköisyydet, jotka kuvastavat ilmiön säännönmukaisuutta, kun sitä toistetaan (ns. *frekventistinen päättely*).

Objektiivinen todennäköisyys edellyttää useampia satunnaisilmiön **empiirisiä havaintoja**. On tapana puhua myös todennäköisyyden empiirisestä tulkinnasta tai **frekvenssitulkinnasta**.

Todennäköisyyksien merkintätapoja

Oletetaan, että satunnaisilmiö on toistunut n kertaa,
ja että tapahtuma A on esiintynyt f kertaa.

Tällöin A:n **frekvenssi** on f ja A:n **suhteellinen frekvenssi** $\frac{f}{n}$.

Frekvenssitulkinnan mukaisesti A:n **todennäköisyys** on

$$P(A) = \frac{f}{n}$$

P tulee englannin kielen sanasta *probability* (todennäköisyys).

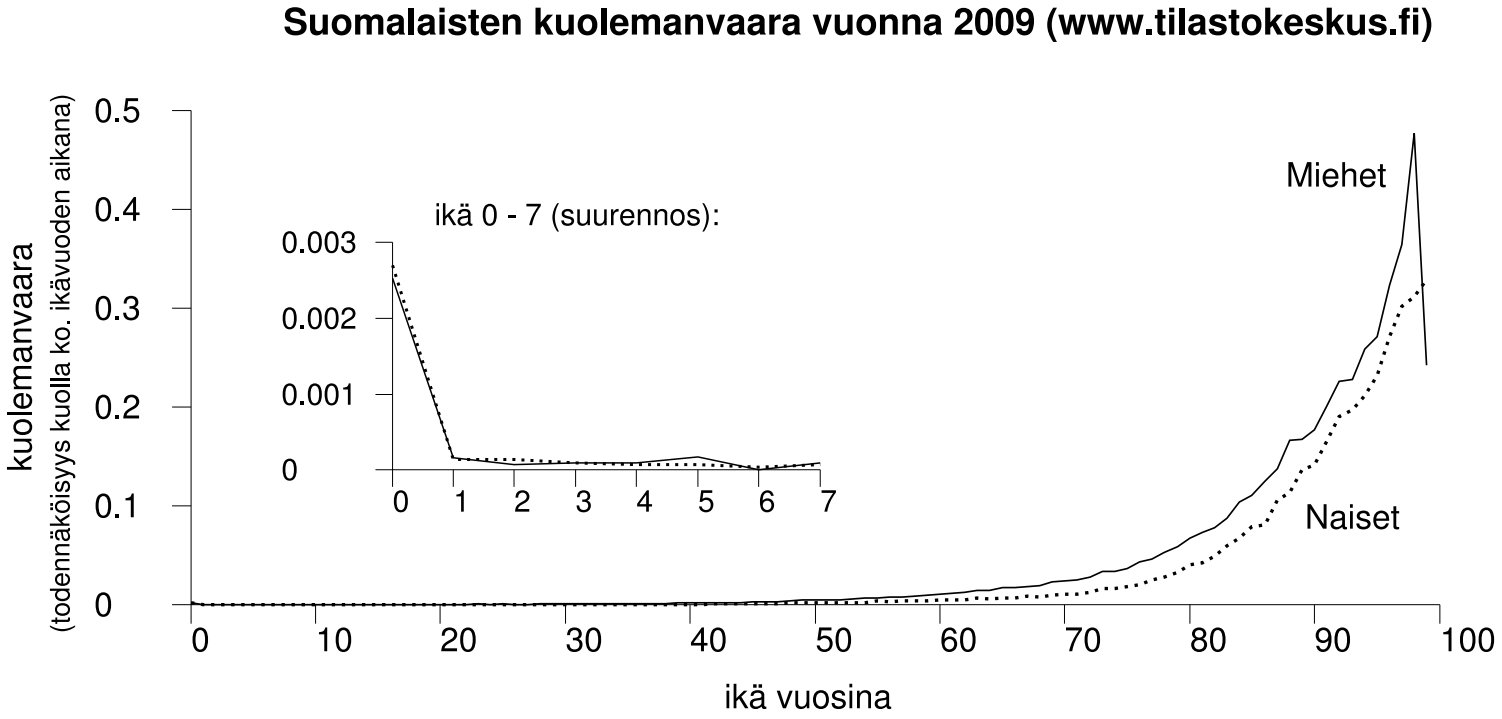
Koska $0 \leq f \leq n$, niin $0 \leq P(A) \leq 1$. Ääritapaukset:

- P(''A on mahdoton'') = 0
- P(''A on varma'') = 1

Todennäköisyydestä käytetään usein myös merkintää $P(A) = p$.

Todennäköisyyden frekvenssitulkinta: tilastot

Jotta suhteellisen frekvenssin tulkinta todennäköisyydeksi olisi luotettavaa, tarvitaan verrattain paljon havaintoja. Ainakaan Suomessa tämä ei ole mikään ongelma, sillä tietoja kerätään ahkerasti erilaisiin rekistereihin ja tilastoihin. Esimerkki:



Todennäköisyyden frekvenssitulkinta: perustelu

Edellä käsitelty todennäköisyyden frekvenssitulkinta perustuu **suurten lukujen lakiin**, jonka voi sanallisesti kuvata seuraavasti:

Tapahtuman suhteellinen frekvenssi $\frac{f}{n}$ *lähestyy tapahtuman todennäköisyyttä* p*, kun toistojen lukumäärä kasvaa.*

Kyseessä on niin sanottu *stokastinen* lähestyminen: todennäköisyys, että $\frac{f}{n}$ eroaa p:stä tulee yhä pienemmäksi, toisin sanoen näiden poikkeama tulee yhä epätodennäköisemmäksi.

Suurten lukujen laki takaa **tilastollisen stabiliteetin**, jonka varassa voidaan tehdä luotettavia johtopäätöksiä.

Tilastojen osalta frekvenssitulkinta nojaa siihen, että tietoja on kerätty riittävän pitkältä ajalta **ja** että ne ovat ajallisesti tarkasteltuina vertailukelpoisia.

Ajatus toistokokeesta ei kuitenkaan sovellu kaikkiin ilmiöihin.

2. Subjektiivinen todennäköisyys

Tapahtuman suhteellinen uskottavuus ilmiön (jopa yksittäisessä) realisaatiossa

- P(''Suomenlahdella sattuu öljykatastrofi'') = **?**
- P(''InterCity-juna törmää hotellin seinään'') = **?**
- P(''aurinkokunnan ulkopuolelta löytyy elämää'') = **?**
- P(''Helsingin Jokerit voittaa Gagarin Cupin 2022'') = **?**

Subjektiiviset todennäköisyydet ovat tyypillisesti eri alojen asiantuntemukseen perustuvien **riskiarvioiden** yhdistelmiä.

Objektiivinen todennäköisyyskäsite soveltuu näihin huonosti.
Sen sijaan tapahtumista voidaan lyödä vetoa. Tilastollisissa analyyseissa on yllättäen käyttöä **vedonlyöntisuhteelle** (*odds*).
Sitä tarvitaan niin lääketieteessä kuin yhteiskuntatieteissä.

Subjektiivisia todennäköisyyksiä hyödynnetään myös tutkittavan ilmiön ennakkotietämyksen täsmentämiseen ns. *Bayes-päättelyssä*, jota sovelletaan yhä useammilla tieteen ja tekniikan aloilla.
Tällä kurssilla pääpaino on frekventistisessä päättelyssä.

Toteamuksia epävarmuudesta ja todennäköisyydestä

''Uncertainty is a fundamental characteristic of any scientific data value.''

Drosg, Manfred (2009). *Dealing with Uncertainties: a Guide to Error Analysis.*
Second edition, Springer. (p. 17)

''Riski on mitattavissa objektiivisella todennäköisyydellä, epävarmuus ei.''

Albert, Michael (2004). Vakuutuksen taloudellinen ja yhteiskunnallinen tehtävä. Teoksessa Hellsten, Katri & Helne, Tuula (toim.) Vakuuttava sosiaalivakuutus? Helsinki: Kelan tutkimusosasto, 22–42. (s. 35)

''Epätäydelliseen informaatioon liittyy joko riskiä, jolloin eri lopputuloksia voidaan arvioida tietyllä laskettavissa olevalla todennäköisyydellä tai epävarmuutta, jolloin näitä todennäköisyyksiä ei voida objektiivisesti arvioida.''

Suoniemi, Ilpo; Tanninen, Hannu ja Tuomala, Matti (2003). Hyvinvointipalveluiden rahoitusperiaatteet. Helsinki: Sosiaali- ja terveysministeriö. (s. 33)

''All statements that express uncertainty are couched in terms of probability.''

Schmitt, Samuel A. (1969). *Measuring Uncertainty: an Elementary Introduction to Bayesian Statistics.* Addison–Wesley. (p. 1)

''Probability that Israel is planning to launch a major offensive against Syria in 30 days – 10% or .1''

Schweitzer, Nicholas. (1976). Bayesian Analysis for Intelligence:
Some Focus on the Middle East. *Studies in Intelligence* 20(2), p. 31–44.

3. Klassinen todennäköisyys

Tapahtuman suhteellinen mahdollisuus symmetrisessä perusjoukossa

- alkeistapahtuma: tapahtuma, jota ei voida jakaa osiin
- toisensa poissulkevat: eivät voi esiintyä yhtaikaa
- symmetria: alkeistapahtumilla sama todennäköisyys (reilu peli)
- historiallinen alkuperä: uhkapelit 1600-luvulla
- todennäköisyydet voidaan selvittää suorilla päättelyillä

Esimerkki: nopanheitto
– *silmäluvut* 1, 2, 3, 4, 5, 6.

P(''nopanheitossa saadaan silmäluku i'') = $\frac{1}{6}, i = 1, 2, \ldots, 6.$

Yleisesti: Oletetaan, että satunnaisilmiön alkeistapahtumia on n, ja näistä tapahtumaan A johtavia alkeistapahtumia on k kpl.

A:n (klassinen) todennäköisyys on näiden ''A:lle suotuisten'' alkeistapahtumien suhteellinen osuus

$$P(A) = \frac{k}{n}$$

Klassinen todennäköisyys: esimerkkejä

Esimerkki: rahanheitto

Rahanheittoa voidaan käsitellä myös klassisena todennäköisyytenä.

Alkeistapahtumia vastaavat todennäköisyydet päätellään suoraan:

P(''tulee kruuna'') = P(''tulee klaava'') = $\frac{1}{2}$.

Esimerkki: kahden nopan heitto (tai saman nopan kahdesti)

Mahdollisia tulospareja (alkeistapahtumia) on $6 \cdot 6 = 36$ kpl:

$$\begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array}$$

Symmetrisyyden perusteella jokaisen todennäköisyys on $\frac{1}{36}$.

Klassinen todennäköisyys: esimerkkejä

Heitetään kahta noppaa. Olkoon tarkasteltava tapahtuma

A = ''Saadaan kummallakin nopalla sama silmäluku''.

Tapahtumalle A suotuisat alkeistapahtumat saadaan helposti poimittua edellä olevasta luettelosta. Suotuisia on kaikkiaan 6 kpl:

$$\begin{array}{cccccc} (1,1) & (2,2) & (3,3) & (4,4) & (5,5) & (6,6) \end{array}$$

Tällöin $P(A) = \frac{6}{36} = \frac{1}{6} \approx 0.17$. Jos taas
A = ''Silmälukujen summa on vähintään 9'', suotuisia ovat

$$\begin{array}{cccc} & & & (3,6) \\ & & (4,5) & (4,6) \\ & (5,4) & (5,5) & (5,6) \\ (6,3) & (6,4) & (6,5) & (6,6) \end{array}$$

ja $P(A) = \frac{10}{36} \approx 0.28$.

Kurssikokeisiin osallistuneiden sukupuoli ja tiedekunta

Lähde: WebOodi/KV (2008–2011)	**Sukupuoli**		(N=1981)
Tiedekunta (tieto puuttuu: N=37)	nainen	mies	**yhteensä**
Valtiotieteellinen[1]	790	324	1114
Matemaattis–luonnontieteellinen[2]	186	283	469
Käyttäytymistieteellinen[1]	123	31	154
Humanistinen[1]	56	22	78
Maatalous–metsätieteellinen[3]	35	20	55
Bio- ja ympäristötieteellinen[3]	31	14	45
Lääketieteellinen[4]	9	0	9
Teologinen[1]	7	2	9
Oikeustieteellinen[1]	2	3	5
Farmasian[3]	3	2	5
Eläinlääketieteellinen[3]	1	0	1
yhteensä	1243	701	1944

[1]Keskustakampus [2]Kumpulan kampus [3]Viikin kampus [4]Meilahden kampus

Esimerkkeinä satunnaisesti valittuja opiskelijoita

Seuraavissa tarkasteluissa äskeisen taulukon esittämää kurssikokeiden osallistujajoukkoa kutsutaan **perusjoukoksi**.

Tehdään tästä perusjoukosta **satunnaisia** ''opiskelijavalintoja'' tarkastellen erilaisia, toisensa poissulkevia alkeistapahtumia
''*Satunnaisesti valittu opiskelija* on . . .''.

Alkeistapahtumia on 1944 kpl, kun jätetään puuttuvat tiedot pois. Jokaisella opiskelijalla on sama mahdollisuus tulla valituksi, siis
valintatodennäköisyys on $\frac{1}{1944} \approx 0.0005$.

Esimerkkinä määritellään tarkemmin tapahtumat H ja K:

- H = ''*[. . .]* on humanistisesta tiedekunnasta''
- K = ''*[. . .]* on jostakin keskustakampuksen tiedekunnasta''

Tapahtumalle H suotuisia alkeistapahtumia on 78. Vastaavasti K:lle suotuisia on 1114+154+78+9+5=1360, joten saadaan

$P(H) = \frac{78}{1944} \approx 0.04$ ja $P(K) = \frac{1360}{1944} \approx 0.70$.

Todennäköisyyslaskennan päättelysäännöt

Monimutkaisempien tapahtumien todennäköisyyksien määräämistä voidaan olennaisesti helpottaa päättely- tai laskusääntöjen avulla. Ideana on palauttaa tarkastelu yksinkertaisempiin tapahtumiin ja niiden yhdistelmiin.

Tärkeimmät käsitteet ja päättelysäännöt:

- toisensa poissulkevat tapahtumat ja yhteenlaskusääntö
- yhtaikaiset tapahtumat ja yhteenlaskusääntö
- komplementtitapahtuma ja sen todennäköisyys
- ehdollinen todennäköisyys ja tulosääntö
- riippumattomuus ja tulosääntö

Päättelysääntöihin perehdytään klassisen todennäköisyyden avulla, koska se on käsitteellisesti yksinkertaisinta. Samat säännöt pätevät kuitenkin riippumatta sovellettavasta todennäköisyyskäsitteestä.

Todennäköisyyslaskennan päättelysääntöjen avulla on tarkoitus oppia yleisemmin **ymmärtämään satunnaisilmiöiden luonnetta.**

Toisensa poissulkevat tapahtumat ja yhteenlaskusääntö

Tapahtuma A ja tapahtuma B ovat toisensa poissulkevia, mikäli ne eivät voi esiintyä yhtaikaa. Esiintyy siis joko A tai B (ei A **ja** B).

Yhteenlaskusääntö toisensa poissulkeville tapahtumille:

$$P(A \text{ tai } B) = P(A) + P(B)$$

Oletetaan, että opiskeluoikeuden saisi vain yhteen tiedekuntaan. Tällöin tapahtumat

A = ''*[. . .]* on maatalous–metsätieteellisestä tiedekunnasta''
B = ''*[. . .]* on bio- ja ympäristötieteellisestä tiedekunnasta''
ovat toisensa poissulkevia.

Taulukosta voidaan päätellä, että

$P(A \text{ tai } B) = \frac{55+45}{1944} = \frac{100}{1944} \approx 0.05.$

Yhtaikaiset tapahtumat ja yhteenlaskusääntö

Jos tapahtuma A ja tapahtuma B eivät ole toisensa poissulkevia, niillä on yhteinen osuus (A **ja** B), joka on vähennettävä pois todennäköisyyksiä laskettaessa (muuten se tulee kahteen kertaan).

Yhteenlaskusääntö yhtaikaisille tapahtumille:

$$P(A \text{ tai } B) = P(A) + P(B) - P(A \text{ ja } B)$$

Esimerkiksi tapahtumat

A = ''*Satunnaisesti valittu opiskelija* on nainen''
B = ''*[. . .]* on valtiotieteellisestä tiedekunnasta''
voivat tietenkin esiintyä yhtaikaa.

Taulukosta nähdään, että $P(A \text{ ja } B) = \frac{790}{1944}$, joten

$P(A \text{ tai } B) = \frac{1243+1114-790}{1944} = \frac{1567}{1944} \approx 0.81.$

Ilman yhteisen osuuden vähennystä tulos olisi järjetön (yli 1)!

Komplementtitapahtuma ja sen todennäköisyys

Jos tapahtuma A **ei** esiinny tarkastellussa satunnaisilmiössä,
niin tällöin esiintyy sen **komplementtitapahtuma** A^C.

Komplementtitapahtuman todennäköisyys:

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

Esimerkiksi jos tarkasteltava tapahtuma on
A = ''*Satunnaisesti valittu opiskelija* on nainen'',
niin sen komplementtitapahtuma on A^C = ''*Satunnaisesti valittu opiskelija* on mies''.

Edellä olevan taulukon perusteella
$P(A^C) = 1 - P(A) = 1 - \frac{1243}{1944} = 1 - 0.64 \approx 0.36.$

Samaan tulokseen päädytään suoralla päättelyllä
$P(A^C) = \frac{701}{1944} \approx 0.36.$

Ehdollinen todennäköisyys

Tarkastellaan tapahtumia A ja B *olettaen että* B *esiintyy*.

Käytännön tutkimuksessa tärkeän tyyppinen kysymys on:

Mikä on A*:n todennäköisyys, jos* B *otetaan huomioon?*

Tätä kutsutaan A:n **ehdolliseksi todennäköisyydeksi** ehdolla B,
merkitään $P(A|B)$, ja luetaan *''*A*:n todennäköisyys ehdolla* B*''*.

Tarkastellaan esimerkiksi tapahtumia
A = ''*Satunnaisesti valittu opiskelija* on nainen'' ja
B = ''*[. . .]* on valtiotieteellisestä tiedekunnasta''.
Kun B:n esiintyminen ''otetaan huomioon'', riittää määrätä naisten suhteellinen osuus valtiotieteellisen tiedekunnan opiskelijoista.

Taulukosta nähdään suoraan, että $P(A|B) = \frac{790}{1114} \approx 0.71$.

Huomaa: tilanne muuttuu, jos ehtotapahtumaksi vaihdetaan A:
$P(B|A) = \frac{790}{1243} \approx 0.64$.

Ehdollinen todennäköisyys ja tulosääntö

Tarkastellaan nyt **yhdistettyä tapahtumaa** A ja B. Olkoot jälleen
A = ''*Satunnaisesti valittu opiskelija* on nainen''
B = ''*[. . .]* on valtiotieteellisestä tiedekunnasta'', jolloin A ja B = ''*[. . .]* on nainen valtiotieteellisestä tiedekunnasta''.

Tulosääntö yhdistetylle tapahtumalle:

$$P(A \text{ ja } B) = P(A|B)P(B) = P(B|A)P(A)$$

Taulukon avulla päätellään jälleen, että
$P(A \text{ ja } B) = P(A|B)P(B) = \frac{790}{1114} \cdot \frac{1114}{1944} \approx 0.41,$

$P(A \text{ ja } B) = P(B|A)P(A) = \frac{790}{1243} \cdot \frac{1243}{1944} \approx 0.41.$

Toisinaan voi olla helppo määrätä $P(A \text{ ja } B)$, jonka avulla saadaan

$$P(A|B) = \frac{P(A \text{ ja } B)}{P(B)} \quad \text{tai} \quad P(B|A) = \frac{P(A \text{ ja } B)}{P(A)}.$$

Riippumattomuus ja tulosääntö

Jos ehdolliseen todennäköisyyteen $P(A|B)$ liittyvässä tilanteessa ehtotapahtumalla B ei ole vaikutusta A:n todennäköisyyteen, niin tapahtumat A ja B ovat **riippumattomia**, toisin sanoen A:n todennäköisyys ei muutu, otettiin B huomioon tai ei.

Siis on vain kaksi vaihtoehtoa, jotka voidaan ilmaista lyhyesti:

Jos $P(A|B) = P(A)$, niin A ja B **ovat** riippumattomia.
Jos $P(A|B) \neq P(A)$, niin A ja B **eivät** ole riippumattomia.

Tulosääntö riippumattomille tapahtumille: *(vrt. ed. sivu)*

$$P(A \text{ ja } B) = P(A)P(B)$$

Tätäkin voi hyödyntää riippumattomuuden selvittämisessä:

Jos $P(A \text{ ja } B) = P(A)P(B)$, niin A ja B **ovat** riippumattomia.
Jos $P(A \text{ ja } B) \neq P(A)P(B)$, niin A ja B **eivät** ole riippumattomia.

Riippumattomuudesta ja riippuvuudesta

Riippumattomuus on keskeisimpiä tilastollisia käsitteitä (vrt. Teema 4, jossa asiaa lähestyttiin **riippuvuuden** näkökulmasta).

Teoriatasolla täydellinen riippumattomuus on osoitettavissa vain melko yksinkertaisissa tilanteissa.

Käytännössä joudutaan tekemään riippumattomuutta koskevia **oletuksia** – joko intuitiivisesti tai sen perusteella, mitä tutkittavasta ilmiöstä etukäteen tiedetään. Saatetaan esimerkiksi **olettaa**, että *kyselytutkimukseen osallistuneiden henkilöiden vastaukset ovat toisistaan riippumattomia* tai että *ihmisen onnellisuus ei riipu käytettävissä olevista tuloista* tai että *gradujen arvosanajakaumat ovat samanlaisia oppiaineesta riippumatta* jne.

Riippumattomuutta koskevien oletusten pätevyyttä on syytä **testata** tilastollisesti. Aiheeseen perehdytään Teemassa 9, kunhan ensin otetaan haltuun riittävästi sitä tukevaa käsitteistöä.

Riippumattomuus: esimerkki

Oheinen esimerkki on pelkistetty tilanne, jossa riippumattomuus ja riippuvuus voidaan nähdä täsmällisesti. **Käytännön tilanteet ovat harvoin** *(jos koskaan!)* **näin yksiselitteisiä.**

Oletetaan, että uurnassa on 3 numeroitua arpaa (1, 2 ja 3), joita nostetaan satunnaisesti. *Oletetaan, että on nostettu arpa nro 3.* Nostamisen jälkeen on kaksi vaihtoehtoista toimintatapaa:

- **Palautetaan** arpa nro 3 uurnaan.Tällöin todennäköisyys saada arpa nro 1 on $\frac{1}{3}$. Täsmällisemmin: $P(\text{"nostetaan nro 1"} \mid \text{"nostettu nro 3"}) = \frac{1}{3} = P(\text{"nostetaan nro 1"})$ Ehto ei siis vaikuta, joten tapahtumat ovat toisistaan **riippumattomia**.
- **Ei palauteta** arpaa nro 3 uurnaan. Tällöin todennäköisyys saada arpa nro 1 on $\frac{1}{2}$, koska jäljellä on enää kaksi arpaa. Täsmällisemmin: $P(\text{"nostetaan nro 1"} \mid \text{"nostettu nro 3"}) = \frac{1}{2} \neq P(\text{"nostetaan nro 1"})$ Ehto siis vaikuttaa, joten tapahtumat **riippuvat** toisistaan.

Tavat 1 ja 2 vastaavat erilaisia otoksen poimintatapoja (Teema 8).

7. Todennäköisyyksien laskentaa

Teemassa 6 tutustuttiin todennäköisyyden ja satunnaisuuden käsitteisiin sekä todennäköisyyslaskennan perusteisiin.

Seuraavaksi tätä aihepiiriä syvennetään hieman perehtymällä **satunnaismuuttujiin** ja **todennäköisyysjakaumiin**.

Otsikkoon innoittivat prof. *Seppo Mustosen* erikoiskurssin (1998) alkusanat:

> *Matemaatikon ja tilastotieteilijän ammattikuvaan kuuluvilla aloilla kohdataan tehtäviä, joissa ei riitä se, että tietää miten lasketaan vaan tulee todella osata laskea.* ***On myös tärkeätä, että pystyy kohtuullisen nopeasti ja hyvin arvioimaan erilaisten tapahtumien todennäköisyyksiä.*** *Ihmisen luontainen taito on tässä suhteessa kehnonlainen.*
>
> — Seppo Mustonen

Laskemisella Mustonen viittaa **tietokoneisiin**, joista hänellä on **50++ vuoden** kokemus. **Survo** on Sepon "tilastollinen laboratorio" ja elämäntyö: www.survo.fi/julkaisut.

Mainitun erikoisen kurssin aiheita olivat mm. *Temppelin pylväsongelma*, *Oikean kolikon tunnistaminen*, *Symmetrinen satunnaiskulku m-ulotteisessa avaruudessa* ja *Maailman ympäri* (**Simo Frangénin** tuolloin juontama television viihdeohjelma). Näihin kiehtoviin teemoihin voi edelleen uppoutua sivulla www.survo.fi/demos.

Korostamani virke koskee useiden muidenkin alojen ammattikuvia. Mm. erilaisten riskien arviointi on tärkeällä sijalla päätöksenteossa. On siis hyvä perehtyä todennäköisyyslaskentaan vähän tarkemmin.

Diskreetit satunnaismuuttujat

Satunnaismuuttuja

- kuvaa satunnaisilmiön tapahtumavaihtoehtoja numeerisesti
- saa arvonsa satunnaisesti, toteutuvan vaihtoehdon mukaan
- voi olla tyypiltään **diskreetti** tai **jatkuva** (vrt. Teema 3)

Keskitytään aluksi diskreetteihin satunnaismuuttujiin.

Esimerkki: lapsen sukupuolen määräytyminen satunnaisilmiönä.
Määritellään satunnaismuuttuja X seuraavasti:

$$X = 1, \text{ jos syntyy tyttö, ja}$$
$$X = 0, \text{ jos syntyy poika.}$$

Periaatteessa voidaan olettaa, että $P(X = 1) = P(X = 0) = \frac{1}{2}$, joskin todellisuudessa $P(X = 0) > P(X = 1)$. Numeerinen koodaus voidaan valita miten tahansa, kunhan eri tapahtumat koodataan eri arvoilla.

Pistetodennäköisyysfunktio

Diskreetin satunnaismuuttujan **pistetodennäköisyysfunktio** on

$$P(X = x)$$

kaikille X:n mahdollisille (diskreeteille) arvoille x.
Usein voidaan lyhentää $P(x)$. Kuten aiemmin, $0 \leq P(x) \leq 1$.

Satunnaismuuttujia merkitään isoilla kirjaimilla (X, Y, Z jne.) ja niiden saamia arvoja vastaavasti pienillä kirjaimilla (x, y, z jne.)

Pistetodennäköisyysfunktiota on tapana visualisoida pylväskuvaa muistuttavalla esityksellä, jossa jokaista arvoa x kuvaa vastaavan todennäköisyyden $P(X = x)$ mittainen jana.

Janojen pituuksien, ts. todennäköisyyksien $P(X = x)$ summa on 1.

Todennäköisyysjakauma

Kaikkien satunnaisilmiöön liittyvien tapahtumien todennäköisyydet kuvaa **todennäköisyysjakauma**. Se määritellään joko

- satunnaismuuttujan arvoilla ja niiden todennäköisyyksillä tai
- pistetodennäköisyysfunktion avulla.

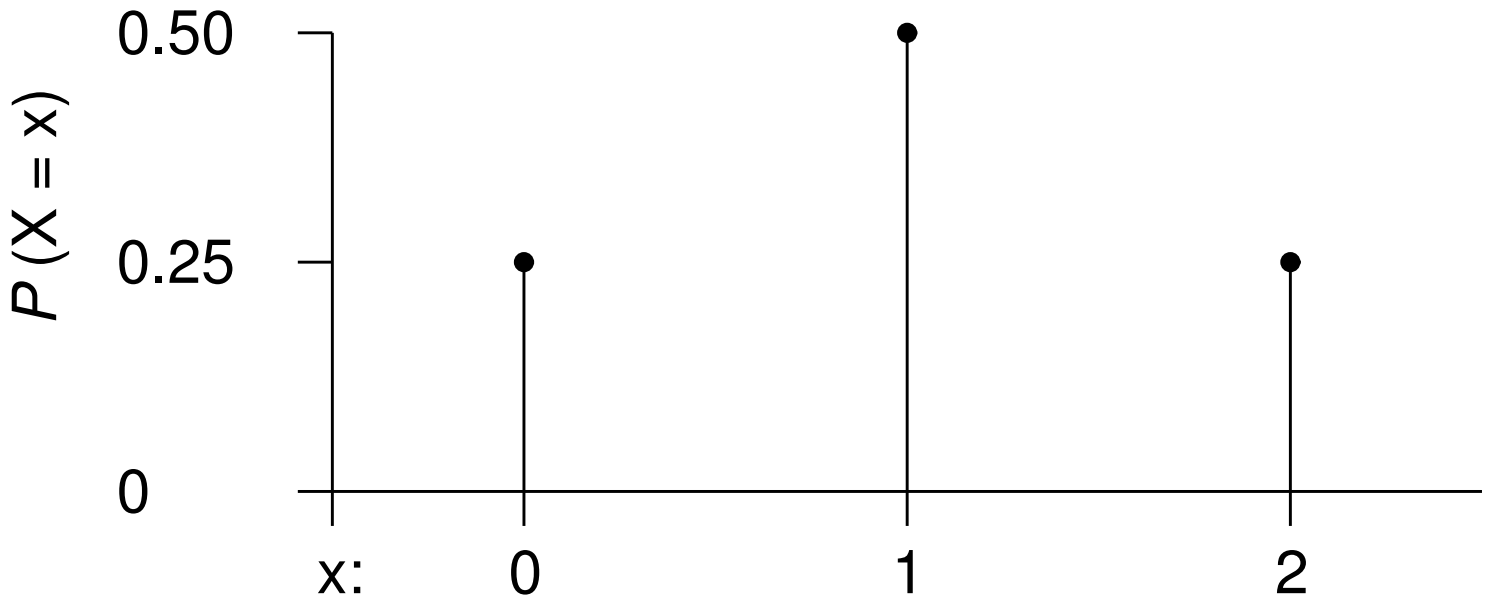
Esimerkki: Kahden rahan heitto; X = ''kruunujen lukumäärä''.

Satunnaismuuttujan X mahdolliset arvot ovat 0, 1 tai 2.
Niiden todennäköisyydet $P(X = x)$ ovat vastaavasti

$$P(X = 0) = \frac{1}{4} = 0.25,$$

$$P(X = 1) = \frac{1}{2} = 0.5 \text{ ja}$$

$$P(X = 2) = \frac{1}{4} = 0.25$$



Todennäköisyysjakauma suhteellisista frekvensseistä

Asuntotyyppi (KPT 2001, $N = 3927$)	Koodi	%
Omistusasunto	1	57.9
Vuokra-asunto	2	38.3
Asumisoikeus- tai osaomistusasunto	3	2.4
Jokin muu	4	1.4

Taulukon perusteella voidaan määritellä satunnaismuuttuja X (asuntotyyppi) ja sen diskreetti todennäköisyysjakauma

$$
\begin{aligned}
P(X = 1) &= 0.579 \\
P(X = 2) &= 0.383 \\
P(X = 3) &= 0.024 \\
P(X = 4) &= 0.014.
\end{aligned}
$$

Näillä voidaan tehdä laskelmia tavalliseen tapaan, esimerkiksi $P(\text{''ei omistusasunto''}) = P(X = 1)^C = 1 - P(X = 1) = 0.421.$

Diskreetti tasainen jakauma

Symmetrisiin alkeistapahtumiin liittyvää todennäköisyysjakaumaa kutsutaan **diskreetiksi tasaiseksi jakaumaksi**.

Oletetaan, että satunnaisilmiö voidaan jakaa symmetrisiin alkeistapahtumiin $A_1, A_2, \ldots, A_n$. Tällöin

$$P(A_i) = \frac{1}{n}, \text{ kun } i = 1, 2, \ldots, n.$$

Kun määritellään satunnaismuuttuja X siten, että

$$X = i, \text{ jos tapahtuma } A_i \text{ esiintyy},$$

niin X:llä on diskreetti tasainen jakauma

$$P(X = i) = \frac{1}{n}, \text{ kun } i = 1, 2, \ldots, n$$

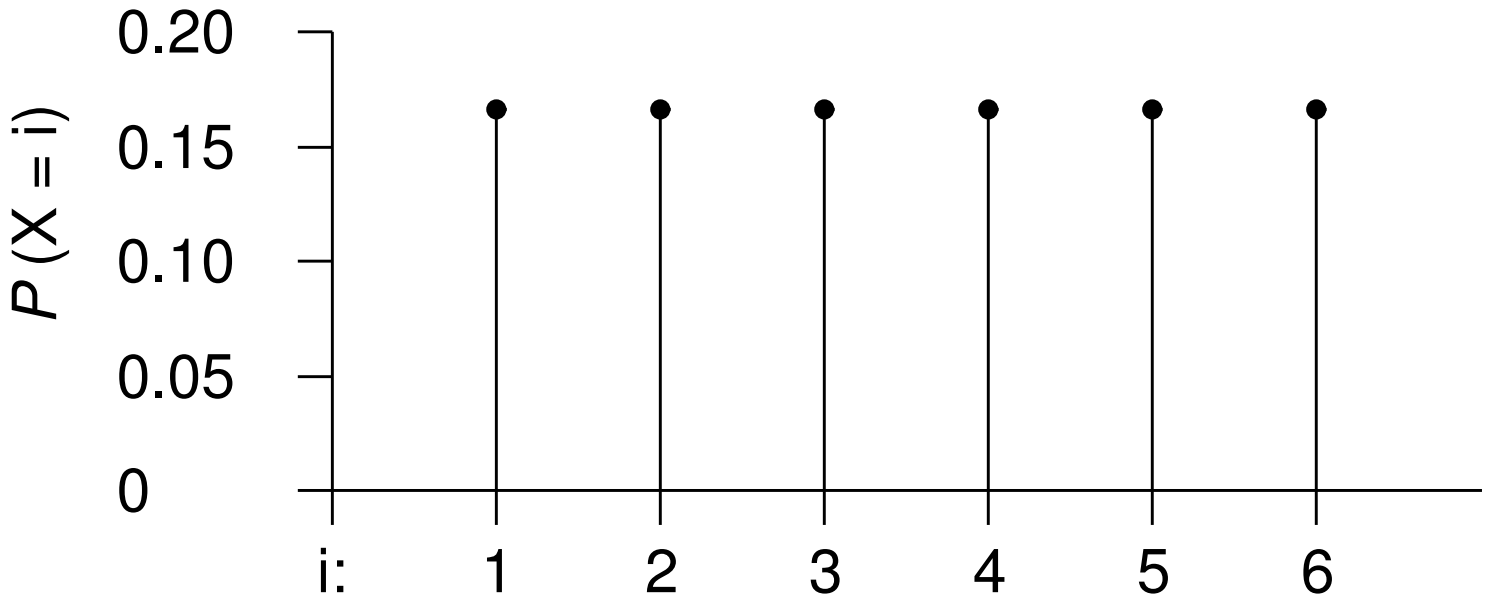
Vrt. esim. opiskelijoiden satunnainen valinta perusjoukosta (Teema 6).

Diskreetti tasainen jakauma: esimerkki

Esimerkki: Yhden nopan heittoa kuvaa diskreetti tasainen jakauma

$$P(\text{”saadaan silmäluku } i\text{”}) = \frac{1}{6}, \\ \text{kun } i = 1, 2, \ldots, 6.$$

Jakauman graafinen esitys näyttää seuraavalta:



Huomaa, että $\frac{1}{6} \approx 0.17$.

Binomijakauma: tärkein diskreetti jakauma

Diskreeteistä todennäköisyysjakaumista tärkein on binomijakauma. Se perustuu sarjaan **koetoistoja**, joissa oletetaan seuraavaa:

- joka kokeessa on vain **kaksi** mahdollista tulosvaihtoehtoa (onnistuu/ei onnistu, kruuna/klaava, kyllä/ei jne.)
- ''onnistumisen'' todennäköisyys p on joka toistossa **sama**
- koetoistot (n kpl) ovat toisistaan **riippumattomia**

Asetelma vastaa **harhaisen rahan** heittoa, sillä todennäköisyys p ei välttämättä ole $\frac{1}{2}$. Olennaista on, että p pysyy samana.

Oletusten voimassaoloa on käytännön sovellustilanteissa hyvä arvioida. Joka tapauksessa binomijakaumalla on paljon käyttöä niin yhteiskuntatieteissä kuin muillakin aloilla. Vaikka termistö viittaakin kokeelliseen tutkimukseen, se sopii yhtä hyvin myös otantatyyppisiin tutkimusasetelmiin, joissa koetoistoja vastaavat toisistaan riippumattomat havainnot.

Muun muassa yhteiskuntatieteissä yleisesti käytetty *logistinen regressiomalli* perustuu binomijakauman soveltamiseen.

Binomijakauma ja todennäköisyysfunktiot

Edellä olevia lukuja n ja p kutsutaan jakauman **parametreiksi**. Oletetaan nyt, että satunnaismuuttujan X jakauma on binomijakauma. Tätä merkitään lyhyesti $X \sim Bin(n, p)$, joka luetaan ''X noudattaa binomijakaumaa parametrein n ja p''.

Esimerkki: $n = 5$ ja $p = 0.2$. Rahanheitossa olisi kyse siitä, montako kruunaa saadaan 5 heitolla, kun raha on harhainen.

Eri lukumäärien todennäköisyydet saadaan binomijakauman pistetodennäköisyysfunktion avulla. Tällaiset funktiot sisältyvät useimpiin tilasto-ohjelmiin, jopa moniin laskimiin.

Esimerkiksi **R-ohjelmistolla** todennäköisyyksiä voi laskea aktivoimalla seuraavanlaisia funktioita; tässä siis binomijakauman pistetodennäköisyysfunktio:

> dbinom(0, 5, 0.2)
[1] 0.32768

> dbinom(1, 5, 0.2)
[1] 0.4096

> dbinom(2, 5, 0.2)
[1] 0.2048

> dbinom(3, 5, 0.2)
[1] 0.0512

> dbinom(4, 5, 0.2)
[1] 0.0064

> dbinom(5, 5, 0.2)
[1] 0.00032

Vastaavat **kertymäfunktiot** näyttävät seuraavilta (tulokseksi saadaan äskeiset arvot yhteen kerrytettyinä eli kumulatiivisesti yhteen laskettuina):

> pbinom(0, 5, 0.2)
[1] 0.32768

> pbinom(1, 5, 0.2)
[1] 0.73728

> pbinom(2, 5, 0.2)
[1] 0.94208

> pbinom(3, 5, 0.2)
[1] 0.99328

> pbinom(4, 5, 0.2)
[1] 0.99968

> pbinom(5, 5, 0.2)
[1] 1

Binomijakauman pistetodennäköisyysfunktio

Jos $X \sim Bin(n, p)$, niin sen pistetodennäköisyysfunktio on

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$
$$\text{kun } x = 0, 1, 2, ..., n,$$

jossa $\binom{n}{x}$ (luetaan ''n yli x'') tarkoittaa **binomikerrointa**. Kerroin ilmaisee, kuinka monella tavalla n alkion joukosta voidaan valita osajoukko, jossa on x alkiota. Se on lyhennysmerkintä kaavalle

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

jossa $n!$ on luvun n **kertoma**, esim. $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$.

Huomaa: 0! = 1. Edellä olevassa esimerkissä saataisiin 3 kruunalle

$$\begin{aligned} P(X=3) &= \binom{5}{3}(0.2)^3(1-0.2)^{(5-3)} \\ &= \frac{5!}{3!2!}(0.2)^3(0.8)^2 = 0.0512. \end{aligned}$$

Binomikerroin ja lottorivit

Binomikertoimen avulla saadaan määrättyä erilaisten lottorivien lukumäärä (ts. sellaisten yhdistelmien lukumäärä, jossa valitaan 7 numeroa 40:stä kiinnittämättä huomiota järjestykseen):

$$\begin{aligned} \binom{40}{7} &= \frac{40!}{7!(40-7)!} \\ &= \frac{40 \cdot 39 \cdot 38 \cdots 3 \cdot 2 \cdot 1}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 33 \cdot 32 \cdot 31 \cdots 3 \cdot 2 \cdot 1} \\ &= \frac{40 \cdot 39 \cdot 38 \cdot 37 \cdot 36 \cdot 35 \cdot 34}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 18643560. \end{aligned}$$

Erilaisia lottorivejä on siis n. 18.6 miljoonaa, ja siten todennäköisyys saada lotossa 7 oikein on $\frac{1}{18643560} = 0.00000005363782$.

Sivuhuomautuksena todettakoon, että yleisesti todennäköisyys saada lotossa k numeroa oikein on

$$\frac{\binom{7}{k}\binom{40-7}{7-k}}{\binom{40}{7}}$$

Kaavan avulla voi harjoitella binomikertoimen käyttöä ja selvittää samalla eri voittoluokkien todennäköisyydet. Kannattaa myös todentaa kaavan toimivuus, kun $k = 7$. Tarkempia tietoja löytyy mm. Veikkauksen lottosivuilta tai Wikipediasta.

Binomijakauma ja vakioveikkaus

Pelkistetty esimerkki binomijakauman pistetodennäköisyysfunktion soveltamisesta on urheiluvedonlyönti nimeltä *vakioveikkaus*, jossa veikataan, päättyvätkö jonkin urheilulajin pelikierroksen 13 ottelua kotivoittoon (''1''), tasapeliin (''X'') vai vierasvoittoon (''2'').

Kotivoiton (''1'') todennäköisyys on oikeasti usein suurempi, mutta oletetaan tässä ''1'', ''X'' ja ''2'' yhtä todennäköisiksi ($p = \frac{1}{3}$). Koska *veikkaajan kannalta* vain oikea vaihtoehto kiinnostaa, kaksi väärää yhdistetään ($1 - p = \frac{2}{3}$). Näin tulosvaihtoehtoja per ottelu on kaksi (''oikein''/''väärin''), ja voidaan soveltaa binomijakaumaa:

$$P(\text{”10 oikein”}) = \binom{13}{10} \left(\frac{1}{3}\right)^{10} \left(1 - \frac{1}{3}\right)^{3}$$

$$= 0.0014350918854$$

tai

$$P(\text{”13 oikein”}) = \binom{13}{13} \left(\frac{1}{3}\right)^{13} \left(1 - \frac{1}{3}\right)^{0}$$

$$= 0.00000062722547$$

(vrt. Loton 7-oikein -tulokseen - mitä havaitset?)

Valittuja paloja tieteellisistä artikkeleista (1/4)

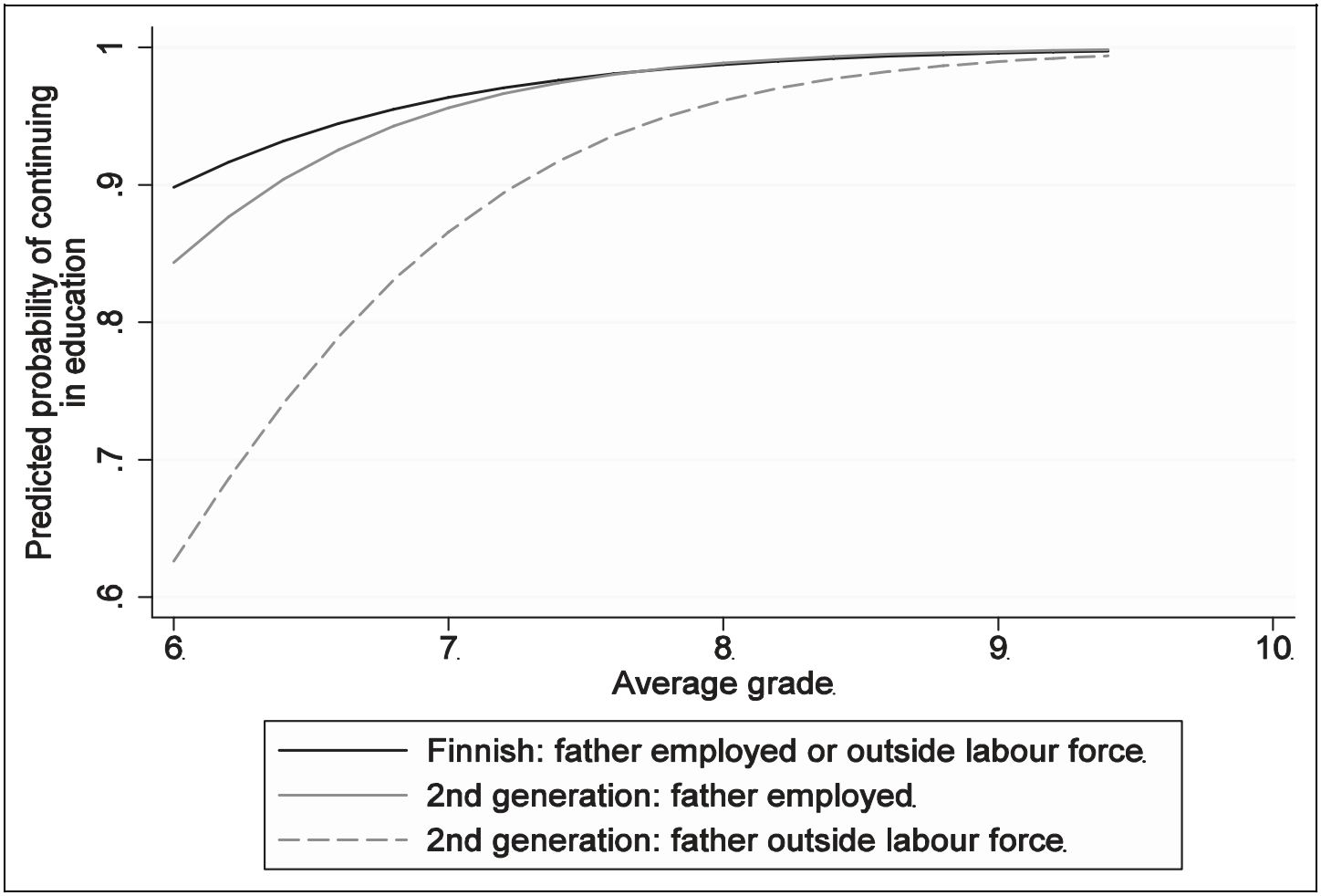


Figure 3. Effect of grades and father's labour force participation on continuation in education for different generations, predicted probabilities (range of grades 6–9.5, male student who is living in a household with two adults and a maximum of two children, and who has an employed mother, a father who is employed or outside the labour force, and at least one parent with tertiary education)

Kilpi–Jakonen, Elina (2011). Continuation to upper secondary education in Finland: Children of immigrants and the majority compared. *Acta Sociologica* 54(1)77–106.

Valittuja paloja tieteellisistä artikkeleista (2/4)

Table 2. Factors Associated With Higher Probability of Accessing the Closest Center by Arrival

	Ambulance Arrival			Other Arrival		
Factor	Odds Ratio	95%CI	*P*	Odds Ratio	95%CI	*P*
Academic hospital	6.90	6.69–7.11	<0.001	6.18	5.96–6.42	<0.001
High-volume hospital	1.93	1.88–1.98	<0.001	1.78	1.73–1.84	<0.001
Rural area	1.22	1.13–1.31	<0.001	1.14	1.05–1.23	<0.001
Lower SES*	1.19	1.16–1.23	<0.001	1.10	1.05–1.15	<0.001
Male	1.04	1.02–1.07	<0.001	1.09	1.06–1.12	<0.001
Older patient†	1.02	1.02–1.03	<0.001	1.01	1.01–1.02	<0.001
Higher Charlson Index†	1.02	1.01–1.04	<0.001	0.98	0.96–0.99	<0.005

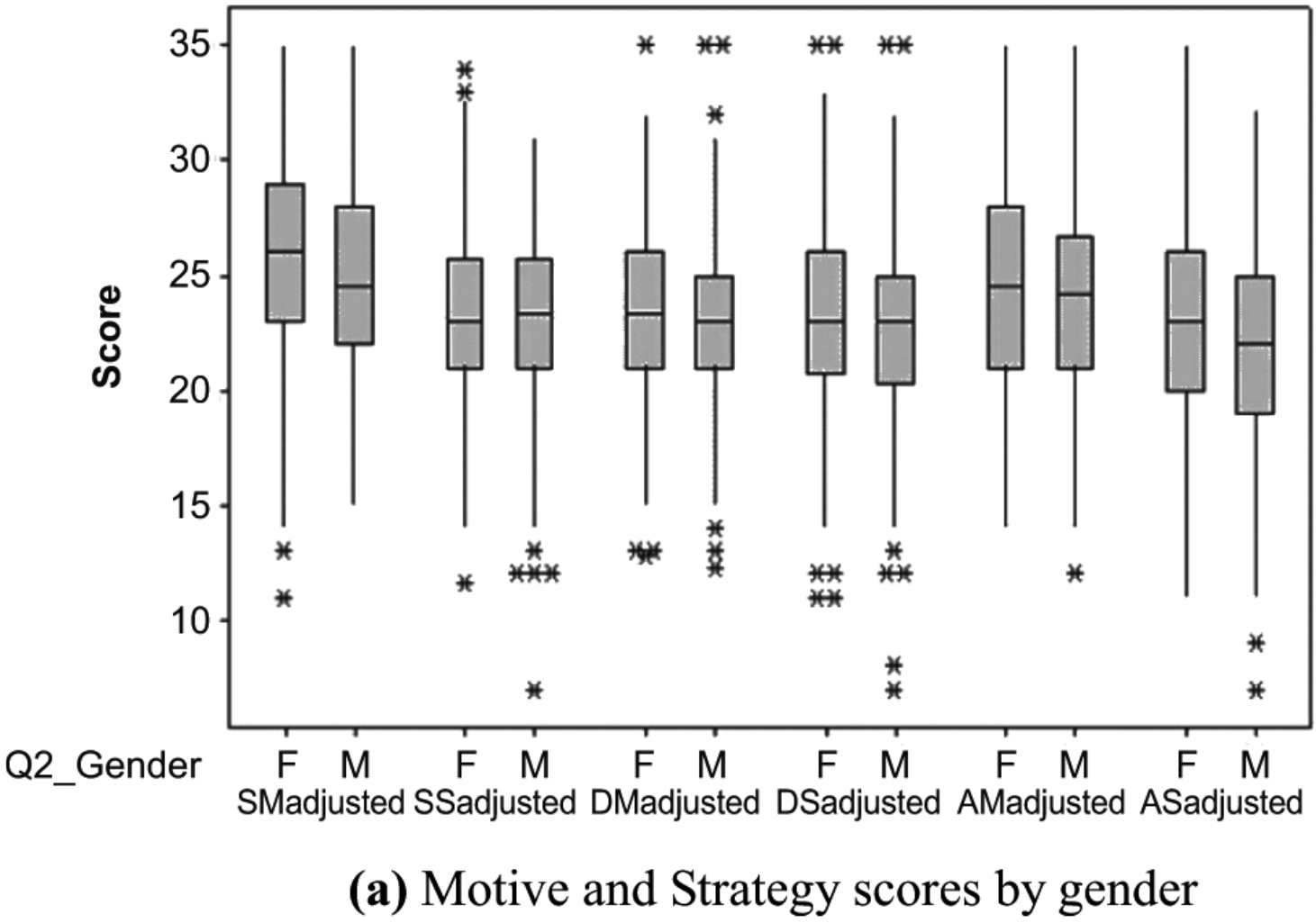
Odds ratios represent likelihood of travel beyond closest available center.
CI indicates confidence interval; Q, quintile; SES, socioeconomic status.
*SES Q1 (lowest) vs all others.
†Higher Charlson-Deyo Index corresponding to greater comorbidity.

Ahuja, Christopher; Mamdani, Muhammad; Saposnik, Gustavo (2012). Influence of socioeconomic status on distance traveled and care after stroke. *Stroke*, 43:233–235. Published online before print Oct 6, 2011, doi: 10.1161/STROKEAHA.111.635045

Valittuja paloja tieteellisistä artikkeleista (3/4)

3.3 Differences between the male and female students
The mean scores for Surface Motive ($p = 0.009$) and Achieving Strategy ($p = 0.018$) were significantly higher for female students, in other words, on average female students used more Surface Motive and Achieving Strategies compared to their male counterparts. Although the means of other subscales were not significantly different by gender ($p > 0.05$), female students had either roughly equal or slightly higher mean scores for the subcategories. The comparative box plot in Figure 1 shows the distribution of Motive and Strategy scores by gender for the three dimensions of the SPQ.

The distribution of the Approaches scores is shown in the comparative box plot in Figure 1. Although the medians were very similar for all three approaches, the means of Surface (male = 47.9, female = 49.3; $p = 0.045$) and Achieving Approaches (male = 45.8, female = 47.5; $p = 0.044$) scores were significantly higher for female students. This can be interpreted as females being more goal oriented, that is, to pass the unit and achieve the highest possible grade.



(a) Motive and Strategy scores by gender

Bilgin, Ayse A. B. (2011). Does learning in statistics get deeper or shallower? *International Journal of Educational Management*, 25(4)378–389.

Valittuja paloja tieteellisistä artikkeleista (4/4)

Is there joy in statistics?

We may not have convinced you about the joy, but hope this short article has convinced you of the need to be "statistics-literate". Both in producing research reports, and in the evaluation of other people's work, it is important to recognize the dependence of conclusions on the nature of the data analysed. One's paymasters, too, will have statistics-literate people who will soon spot flaws in the most impressively-presented case.

There is, though, some pleasure and satisfaction in coming to grips with abstract statistical concepts, learning to use the various tools and techniques to craft satisfactory results. Statistics can be a creative process. Statistical thinking leads to (or uses) rigorous logical thought, which is a good discipline for professional work, and engenders a critical outlook which, as a healthy critical outlook, is a valuable piece of mental equipment.

Further reading

Graham, A., Investigating Statistics: A Beginner's Guide, Hodder & Stoughton, Sevenoaks, 1990.

Paulos, J.A., Innumeracy: Mathematical Illiteracy and its Consequences, Penguin, Harmondsworth, 1988. (A short and excellent book for those who feel "no good at maths".)

Stephen, Peter; Hornby, Susan (1995). The joys of statistics. *Library Review* 44(8)56–62.

Paulos, J. A. (1991). Numerotaidottomuus: matemaattinen lukutaidottomuus ja sen seuraukset (suom. Klaus Vala), Art House. **Katso myös** *Solmu* (4/1999): Numerosokeus – kansainvälinen vitsaus (erit. kohta ''Todennäköisyyksien arviointi'')

Jatkuvat satunnaismuuttujat ja niiden jakaumat

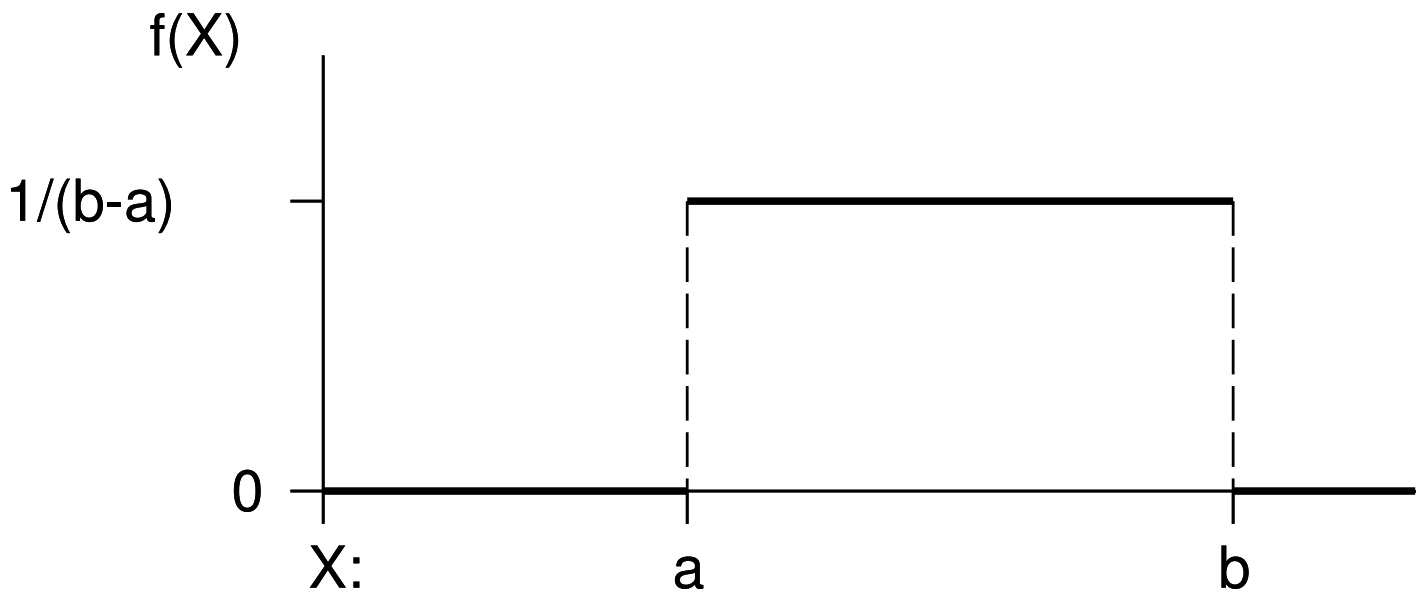
Jatkuva satunnaismuuttuja saa tietyltä väliltä mitä tahansa arvoja. Sen **todennäköisyysjakauman** määrittelee **tiheysfunktio**.

Diskreetin ja jatkuvan satunnaismuuttujan ero:

- **Diskreetti:** saa yksittäisiä (diskreettejä, erillisiä) arvoja.
- **Jatkuva:** jokaisen yksittäisen arvon todennäköisyys on 0, ja huomio kohdistuu erilaisiin väleihin liittyviin todennäköisyyksiin.

Esimerkki: **jatkuva tasainen jakauma**, jonka tiheysfunktio on

$$f(X) = \frac{1}{b-a}, \text{ jos } X \text{ kuuluu välille } (a, b),$$
$$\text{muulloin } 0.$$



- **Diskreeteissä** jakaumissa todennäköisyyksien summa on 1.
- **Jatkuvissa** jakaumissa tiheysfunktion kuvaajan ja vaaka-akselin väliin jäävä pinta-ala on vastaavasti 1.

Oheisessa kuvassa kyseessä on suorakulmion pinta-ala $(b-a) \cdot \frac{1}{b-a} = 1$.

Normaalijakauma: tärkein jatkuva jakauma

Jatkuvista todennäköisyysjakaumista tärkein on normaalijakauma, jolla on erittäin keskeinen merkitys tilastollisessa päättelyssä.

Normaalijakauman määrittelevä tiheysfunktio* on

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right\},$$

jossa symboleilla on seuraavat merkitykset:

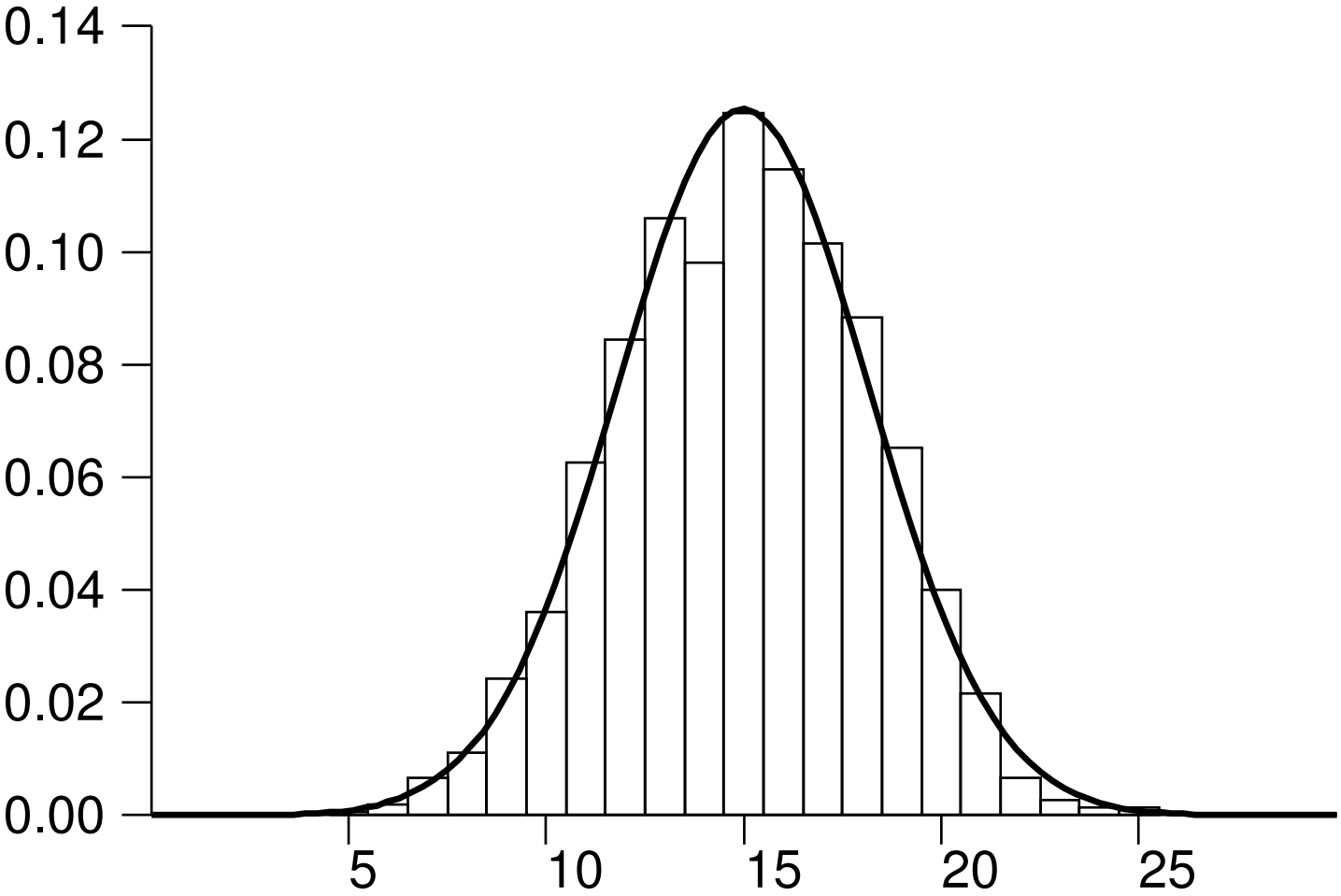
$$\begin{aligned}
\mu &= \text{normaalijakauman odotusarvo (”myy”)} \\
\sigma &= \text{normaalijakauman hajonta (”sigma”)} \\
\pi &= \text{ympyrän kehän ja halkaisijan suhde } (\pi = 3.14159\ldots) \\
\exp &= \text{eksponenttifunktio } (\exp(1) = e = 2.71828\ldots)
\end{aligned}$$

Normaalijakauma on määritelty kaikilla X:n arvoilla. Jakauman parametrit ovat μ ja σ^2, ja sitä merkitään $X \sim N(\mu, \sigma^2)$.

***** Hankalan oloista lauseketta ei käytännön laskuissa tarvita, vaan normaalijakaumaan liittyvät todennäköisyydet määrätään funktioiden tai taulukoiden avulla.

Normaalijakauman tiheysfunktion kuvaaja

Oheisessa kuvassa on KPT-aineiston ($N = 1518$) summamuuttujan (vrt. Teema 4) **Netti** empiiristä jakaumaa ilmentävä histogrammi sekä siihen sovitettu, teoreettista jakaumaa kuvaava normaalijakauman tiheysfunktio, jossa $\mu = 15$ ja $\sigma = 3.2$:



Pystyakseli kuvaa nyt todennäköisyyksiä, joihin histogrammin pylväiden pinta-alat ovat suoraan verrannollisia. Kokonaisuudessaan pinta-ala on 1, ja siis saman suuruinen kuin tiheysfunktion ja vaaka-akselin väliin jäävä alue.

Todennäköisyyksien laskentaa: kertymäfunktio

Tarkastellaan edellisen sivun Netti-muuttujaa (joka siis tässä kuvaa *tyytyväisyyttä kunnan internet-palveluihin*).

Mikä on todennäköisyys, että satunnaisesti valitun vastaajan tyytyväisyys (X) on (tämän summamuuttujan asteikolla 5–25)
a) yli 15?
b) välillä (20,25)?
c) välillä (5,10)?

On siis selvitettävä seuraavat todennäköisyydet:
a) $P(X > 15) = 1 - P(X \leq 15)$,
b) $P(20 < X \leq 25)$,
c) $P(5 < X \leq 10)$

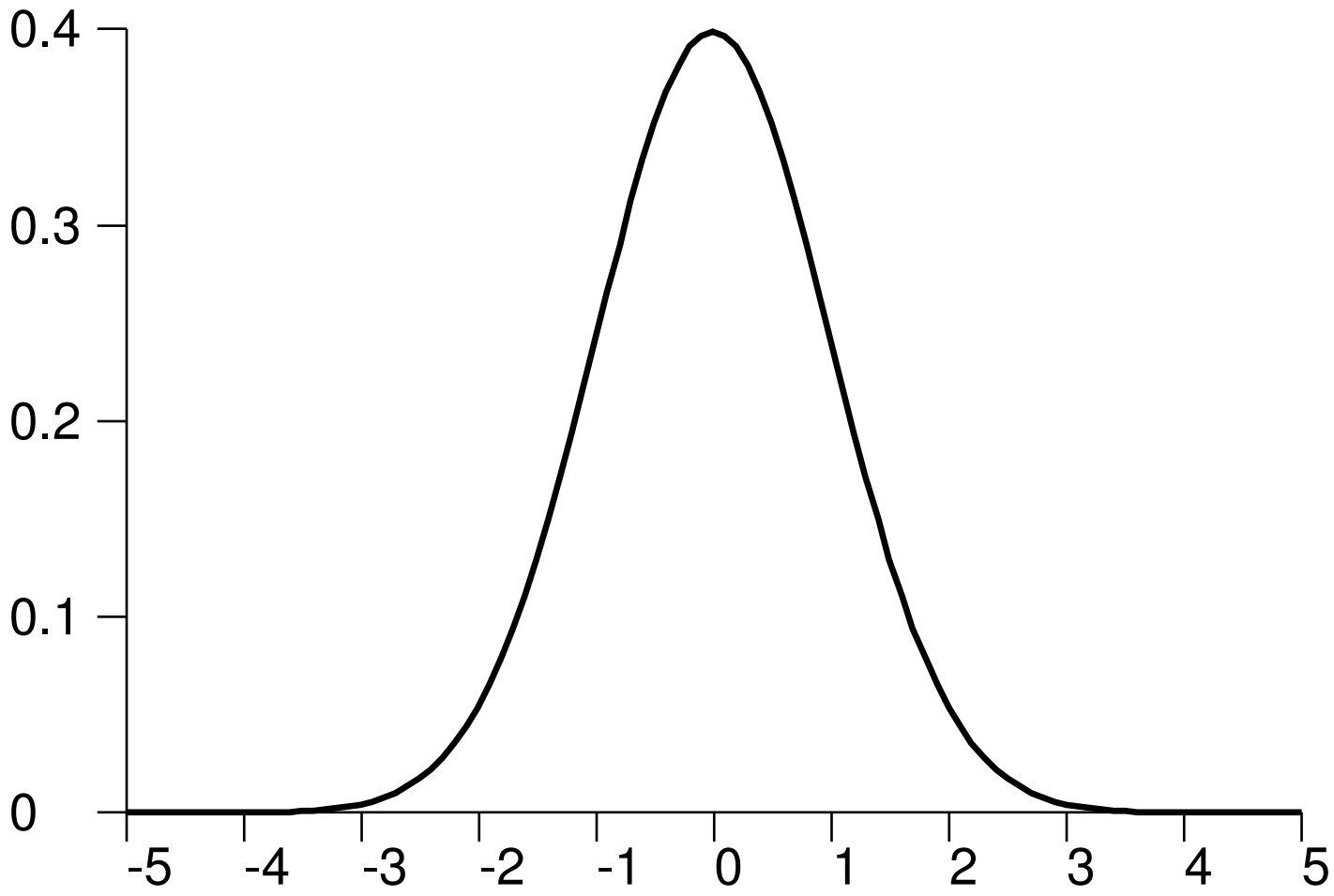
Tällaisiin kysymyksiin saadaan vastaukset normaalijakauman **kertymäfunktion** avulla. Kertymäfunktio ilmaisee, kuinka paljon todennäköisyyttä on kumulatiivisesti kertynyt kyseiseen pisteeseen mennessä tai kyseisellä välillä (vrt. binomijakauma edellä).

R:llä em. todennäköisyyksiä lasketaan aktivoimalla seuraavanlaisia funktioita (normaalijakauman kertymäfunktio):

```
> pnorm(15, 15, 3.2)
[1] 0.5

> pnorm(25, 15, 3.2) - pnorm(20, 15, 3.2)
[1] 0.0581961

> pnorm(10, 15, 3.2) - pnorm(5, 15, 3.2)
[1] 0.0581961
```

Standardoitu (eli normeerattu) normaalijakauma

Käytännössä on usein kätevintä siirtyä standardoituun jakaumaan $N(0, 1)$ eli helpommin tulkittaviin z-pisteisiin $z = (x - \mu)/\sigma$, joiden keskiarvo on 0 ja keskihajonta 1. Vertaa Teema 4, sivu 1: esim. ''keskiarvo $\pm$ 2 keskihajontaa'' vastaa suoraan väliä $[-2, 2]$.



Standardoidun normaalijakauman tiheysfunktio välillä $[-5, 5]$.

Väli $[-5, 5]$ riittää monissa sovellustilanteissa, sillä sen ulkopuolelle jää vain pieni osa todennäköisyydestä (< 0.0000006). Harvinaiset tapahtumat voivat kuitenkin olla **seurauksiltaan** järisyttäviä, eikä niitä saa unohtaa tai väheksyä. Onkin syytä opetella tilannekohtaisesti arvioimaan, onko normaalijakauman soveltaminen perusteltua. Kaikki ilmiöt eivät todellakaan ''noudata'' normaalijakaumaa, mutta joitain ilmiöitä voidaan riittävällä tarkkuudella kuvata normaalijakauman avulla.

Standardoitu normaalijakauma; z-pisteet

Standardoidun normaalijakauman $N(0, 1)$ ansiosta laskelmissa ei tarvitse käyttää muuttujien alkuperäisiä arvoja (kuten edellä Netti-muuttujan yhteydessä). Todennäköisyys $P(X \leq x)$ voidaan selvittää z-pisteiden avulla:

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P(Z \leq z),$$

jossa siis $z = (x-\mu)/\sigma$. Koska nyt $Z \sim N(0,1)$, niin kaikki (mihin tahansa) normaalijakaumaan liittyvät todennäköisyydet saadaan selville standardoidusta normaalijakaumasta.

Tällöin voidaan myös käyttää valmiita taulukoita, jotka löytyvät edelleen useimpien tilastotieteen oppikirjojen liitteistä.

Taulukoiden käyttäminen oli aikoinaan välttämätöntä. Nykyään se ei enää ole, koska ohjelmistot osaavat laskea tarvittavat todennäköisyydet (vrt. edellä). Ohjelmistoille standardoinnista ei ole erityistä hyötyä, mutta todennäköisyyksien **arvioinnin** kannalta se on edelleen hyvä tapa toimia.

8. Parametrien estimointi ja luottamusvälit

Todennäköisyyslaskennan perusteet (Teemat 6 ja 7) antavat hyvän pohjan siirtyä kurssin viimeiseen kokonaisuuteen, nimittäin **tilastolliseen päättelyyn**, jolla tarkoitetaan **perusjoukkoa** koskevien johtopäätösten tekemistä **satunnaisotoksen** perusteella.

Otantaan ja muihin tiedonkeruun menetelmiin ja käsitteisiin tutustuttiin jo Teemassa 3. Tästä eteenpäin käsitellään enimmäkseen tilanteita, joissa perusjoukko on määriteltävissä, ja siitä voidaan poimia edustava satunnaisotos.

Käytännössä tilanne ei aina ole näin selkeä, jolloin tilastollinen päättelykin on epävarmemmalla pohjalla. Pitää kuitenkin muistaa, että tilastollinen päättely perustuu joka tapauksessa todennäköisyyksiin, jotka useissa käytännön tutkimustilanteissa tarkoittavat erilaisia riskejä ja epävarmuuksia. Yleisenä tavoitteena on näiden hallinta.

Riskien hallinta ei suinkaan ole helppoa eikä aina edes mahdollista. Tapahtumien todennäköisyyksien ohella on huomioitava niiden yhteiskunnalliset, taloudelliset ym. **seuraukset.** Tämä koskee etenkin **ennustamista** *(selittäminen on helpompaa!).*

Parametrit ja estimaattorit

Käsite **parametri** esiintyi vaivihkaa jo Teeman 7 yhteydessä.

Yleisesti parametri on **perusjoukon ominaisuus**, josta tehdään arvioita otoksen avulla. Tätä arviointia kutsutaan tilastotieteessä **estimoinniksi**, estimointikeinoa tai -kaavaa **estimaattoriksi** ja otoksen perusteella saatua lukuarvoa **estimaatiksi**.

Tämän kurssin puitteissa keskitytään lähinnä seuraaviin:

Parametri	Estimaattori
Odotusarvo μ	(Otos)**keskiarvo** $\bar{x}$
Hajonta σ	(Otos)**hajonta** s
Todennäköisyys p	**Suhteellinen frekvenssi** $\hat{p}$

Myös muut otoksesta laskettavat tunnusluvut kuten **mediaani**
tai **korrelaatiokerroin** (ks. Teemat 3 ja 4) ovat vastaavien, perusjoukkoa kuvaavien parametrien, estimaattoreita.

Odotusarvo: esimerkkinä odotettu voittosumma

Odotusarvo on satunnaismuuttujan **odotettavissa** oleva arvo.

Esimerkki: peli, johon osallistuminen maksaa 1.5 € kierrosta kohti.

Odotusarvo on satunnaismuuttujan **odotettavissa** oleva arvo.

Esimerkki: peli, johon osallistuminen maksaa 1.5 € kierrosta kohti.

Voittomäärä:	1 €	2 €	3 €
Todennäköisyys:	0.6	0.3	0.1

Mitä on odotettavissa, jos peliä pelataan 100 kierrosta?

Voitettuja pelejä:	$0.6 \cdot 100 = 60$	$0.3 \cdot 100 = 30$	$0.1 \cdot 100 = 10$
Voitettua rahaa:	$60 \cdot 1 = 60$ €	$30 \cdot 2 = 60$ €	$10 \cdot 3 = 30$ €
Voittosumma:	$60 + 60 + 30 = 150$ €		
Kierrosta kohti:	$150/100 = 1.5$ €		

Peli vaikuttaa reilulta, sillä *odotettu voitto kierrosta kohti* on sama kuin osallistumismaksu.

Tulos on tässä saatu intuitiivisesti laskemalla kierroksen *mahdollisten voittomäärien summa painotettuna vastaavilla todennäköisyyksillä*.

Tulos on oikein, mutta tarkastellaan sitä vielä simulointikokeella, jolloin satunnaisvaihtelu tuo päättelyyn pientä epävarmuutta.

Odotusarvo: äskeisen pelin simulointi Survolla

Simuloidaan edellä määriteltyä diskreettiä jakaumaa seuraavasti:

```
MATRIX P ///
1 0.6
2 0.3
3 0.1

MAT SAVE P                  / talletetaan jakauman arvot ja todennäköisyydet
FILE MAKE KOE,1,100,x,1     / luodaan tyhjä aineisto 100 koetoistoa varten
TRANSFORM KOE BY #DISTR(P) / käytetään satunnaislukugeneraattoria RND=rand(2008)

STAT KOE CUR+1              / katsotaan mitä saatiin:
Basic statistics: KOE N=100
Variable: x
mean=1.42     stddev=0.622475
x              f      %      *=2 obs.
       1      65   65.0 *********************************
       2      28   28.0 **************
       3       7    7.0 ***
```

Keskimääräinen voittosumma 100 kierroksen jälkeen olisi alle 1.5 €:

$$\bar{x} = \frac{1}{100}(1 \cdot 65 + 2 \cdot 28 + 3 \cdot 7) = 1.42 \text{ (euroa)}.$$

Kun peliä jatketaan, suhteelliset frekvenssit tarkentuvat (vrt. Teema 6), ja niinpä 10000 kierroksen jälkeen ollaan jo lähellä:

$$\bar{x} = \frac{1}{10000}(1 \cdot 6045 + 2 \cdot 2961 + 3 \cdot 994) = 1.4949 \text{ (euroa)}.$$

Diskreetin jakauman odotusarvo

Vastaava parametrin arvo, teoreettinen odotusarvo, saadaan diskreetissä jakaumassa laskemalla jakauman **mahdollisten arvojen** vastaavilla **todennäköisyyksillä painotettu summa**.

Äskeisessä esimerkissä

$$\mu = 1 \cdot 0.6 + 2 \cdot 0.3 + 3 \cdot 0.1 = 1.5 \text{ (euroa)}.$$

Binomijakaumassa päästään helpolla, sillä siinä odotusarvo saadaan suoraan jakauman parametrien n ja p tulona: $\mu = np$.

Esimerkki: heitetään harhatonta rahaa 10 kertaa.
Mikä on kruunujen lukumäärän odotusarvo?

$$X \sim Bin(n, p), \text{ jossa } n = 10 \text{ ja } p = \frac{1}{2},$$
$$\text{joten odotusarvo on}$$
$$\mu = np = 10 \cdot \frac{1}{2} = 5.$$

Diskreetin jakauman hajonta

Aivan kuten empiirissäkin jakaumissa (ks. Teemat 3 ja 4), on teoreettisissakin jakaumissa kiinnitettävä huomiota odotusarvon ohella **hajontaan**, siis siihen miten muuttujan arvot jakautuvat odotusarvon ympärille.

Diskreeteistä jakaumista tarkastellaan vain binomijakaumaa, jossa hajontakin saadaan suoraan jakauman parametrien n ja p avulla:

$$\text{Jos } X \sim Bin(n, p), \text{ niin } \mu = np \text{ ja } \sigma = \sqrt{np(1-p)}.$$

Äskeisessä esimerkissä, jossa $n = 10$ ja $p = \frac{1}{2}$, on siis $\mu = 5$ ja $\sigma = \sqrt{10 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{\frac{5}{2}} \approx 1.6$.

Hajonnan sijasta voitaisiin tarkastella *varianssia*, jonka lausekekin olisi vähän yksinkertaisempi (ei neliöjuurta), mutta hajonta on helpompi tulkita, koska se ilmaistaan samoissa yksiköissä kuin odotusarvo (vrt. Teema 4).

Jatkuvan jakauman odotusarvo ja hajonta

Jatkuvissa jakaumissa vastaavat tarkastelut johtavat integraaleihin, jotka tällä kurssilla sivuutetaan. Normaalijakauman parametrit, odotusarvo μ ja varianssi σ^2, tulivat esille jo Teemassa 7.

Vaikka normaalijakauman parametrina onkin varianssi (hajonnan neliö), on käytännön laskelmissa ja tulosten esittämisessä usein parempi käyttää sen neliöjuurta eli hajontaa (aivan samoin perustein kuin binomijakaumassa edellä).

Normaalijakauman merkinnässä $N(\mu, \sigma^2)$ viitataan kuitenkin jakauman parametreihin: odotusarvoon ja varianssiin, kuten on tapana myös usei(mmi)ssa oppikirjoissa.

Yksinkertainen satunnaisotanta ja otoksen poimintatapa

Perusjoukon parametrien estimointiin tarvitaan satunnaisotos. Otannan perusasetelma on **yksinkertainen satunnaisotanta**, jossa jokaisella perusjoukon alkiolla on sama todennäköisyys tulla valituksi otokseen. Tämä tarkoittaa, että otanta suoritetaan *palauttaen*, mikä voi vaikuttaa hieman yllättävältä.

Otos voidaan poimia **1) palauttaen** tai **2) palauttamatta**.

Poimintatapa vaikuttaa estimoinnissa ja tilastollisessa testauksessa käytettäviin kaavoihin. Palauttamatta perusjoukon koko pienenee joka poiminnalla (vrt. esim. lottoarvonta).

Käytännössä poimintatapojen ero on useimmiten merkityksetön. Muutaman satunnaisesti valitun henkilön poistaminen (siis poiminta palauttamatta) ei vaikuta seuraavien henkilöiden poimintatodennäköisyyksiin juuri millään lailla, ellei otos sitten muodosta huomattavan suurta osaa perusjoukosta.

Jatkossa (tällä kurssilla) oletetaan otos poimituksi **palauttaen**.

Keskiarvon otantajakauma (normaalinen perusjoukko)

Otoksesta laskettu keskiarvo $\bar{x}$ estimoi odotusarvoa μ. Poimimalla erilaisia otoksia saadaan otantaan liittyvän satunnaisvaihtelun myötä myös eri suuruisia keskiarvoja. Jotta saataisiin käsitys estimaatin tarkkuudesta, pitäisi tietää millainen on $\bar{x}$:n jakauma.

Asiaa voi havainnollistaa simulointikokeiden avulla.

Tulos on se, että mikäli havaintoarvot $x_1, x_2, \ldots, x_n$ muodostavat **riippumattoman otoksen** normaalijakaumasta $N(\mu, \sigma^2)$, niin

$$\bar{x} \sim N(\mu, \sigma^2/n),$$

siis keskiarvo noudattaa normaalijakaumaa, jossa odotusarvo on sama kuin perusjoukossa. Keskiarvon hajonta eli **keskivirhe** on sen sijaan $\sigma/\sqrt{n}$, siis selvästi pienempi kuin perusjoukon hajonta σ. Kun otoskoko n kasvaa, keskiarvon keskivirhe pienenee. Niinpä *päättely keskiarvon perusteella on tarkempaa suuremmilla otoksilla.*

Keskiarvon otantajakauma (yleinen tilanne)

Äskeinen tulos on sinänsä hieno, mutta käytännössä ei tiedetä,

- onko perusjoukko normaalisti jakautunut,
- mikä on perusjoukon hajonta σ, josta riippuu keskivirhe $\sigma/\sqrt{n}$.

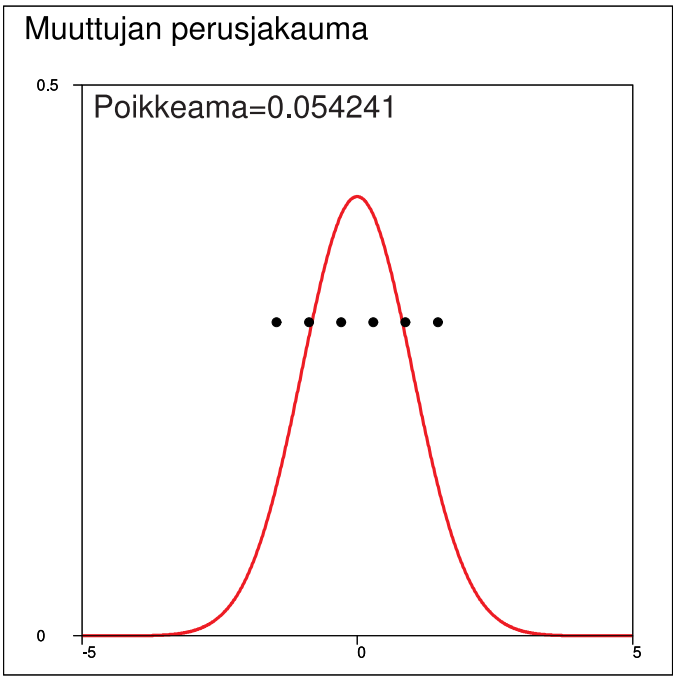
Varsinaisesti käytännön kannalta **merkittävä tulos** onkin, että **perusjoukon jakauman ei tarvitse olla normaalijakauma**, ja silti keskiarvon jakauma **lähestyy** normaalijakaumaa $N(\mu, \sigma^2/n)$, kunhan otoskoko n on ''riittävän'' suuri.

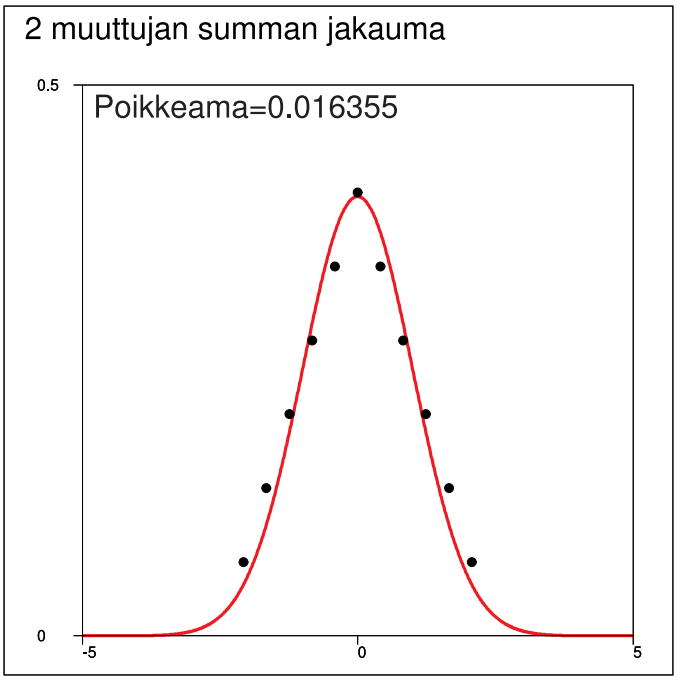
Tätä tärkeää tulosta kutsutaan nimellä todennäköisyyslaskennan **keskeinen raja-arvolause** *(Central Limit Theorem)*. Tulos selittää, minkä vuoksi normaalijakauma soveltuu niin hyvin monenlaisiin sovellustilanteisiin (vaikkei tietenkään kaikkiin mahdollisiin!).

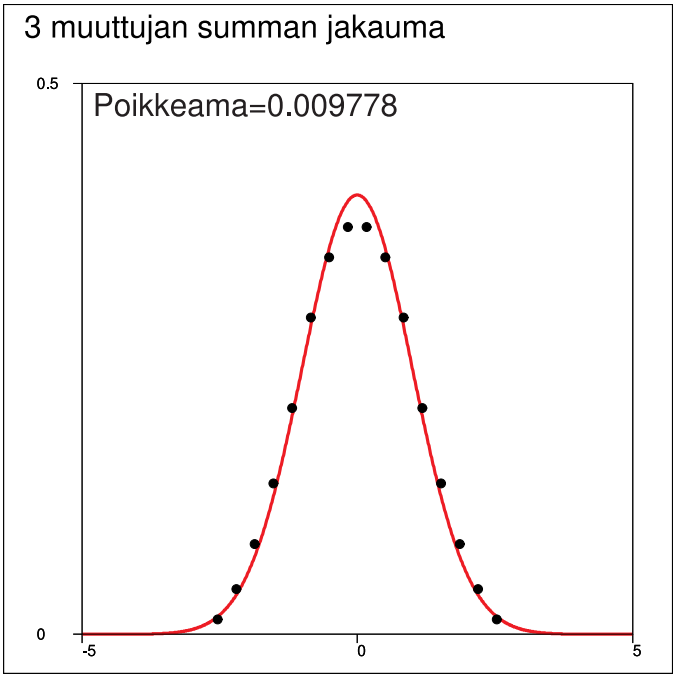
Tätäkin voi tarkastella simulointikokeilla, joista selviää muun muassa, miltä näyttää, kun alun perin diskreeteistä muuttujista muodostettu summamuuttuja lähestyy normaalijakaumaa. *(Käytännön tutkimuksen kannalta tyypillinen ja tärkeä asetelma!)*

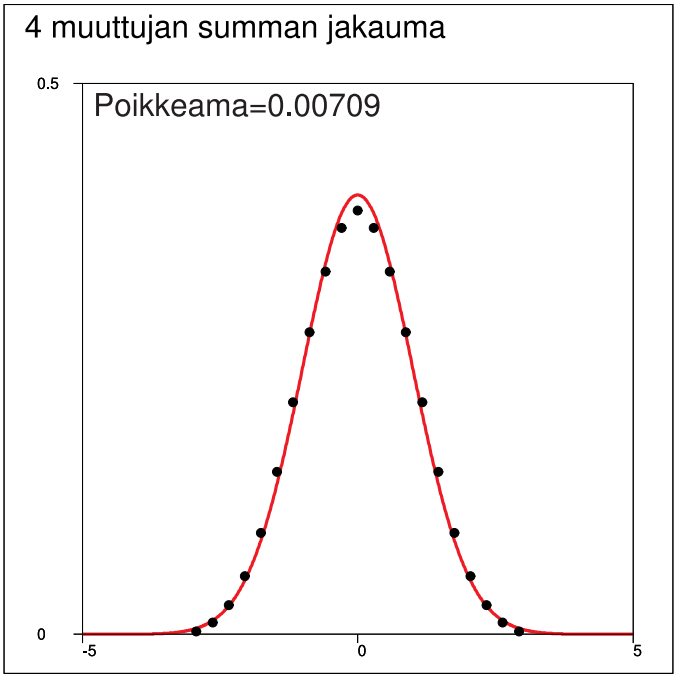
Keskeinen raja-arvolause toimii (näkymiä Survo-demosta)

Tässä muuttujan perusjakauma on nopanheitosta tuttu tasajakauma:





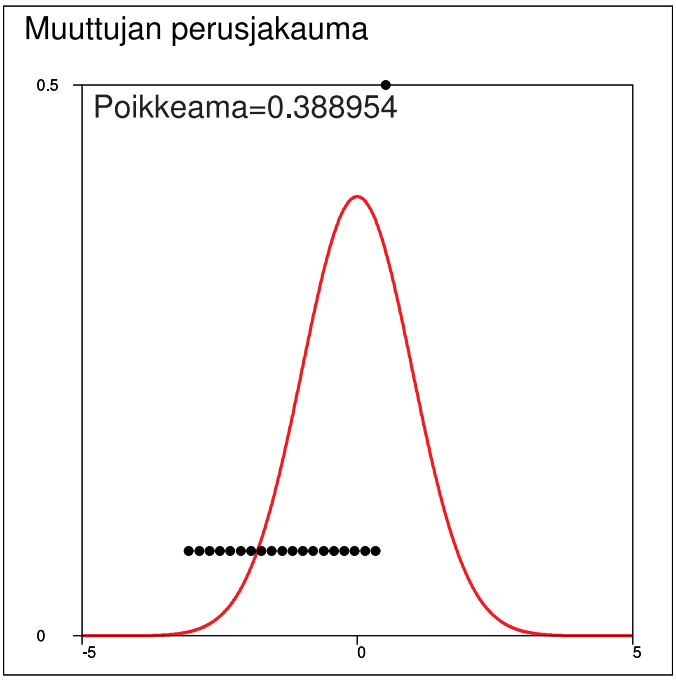


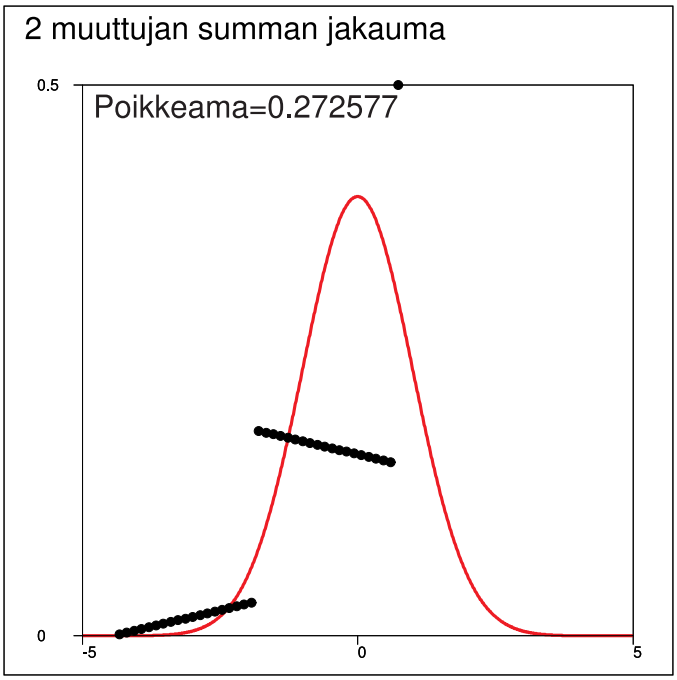


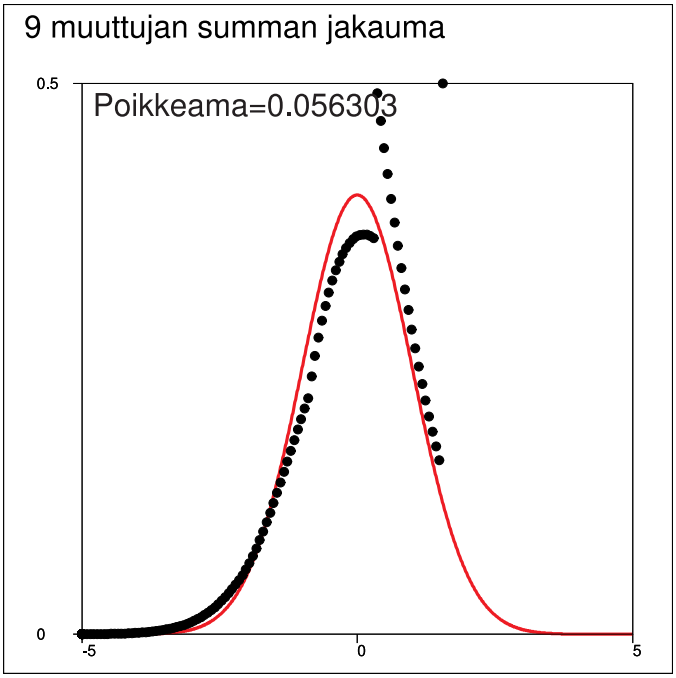
Tässä jo *kahden* osion summamuuttuja lähestyy nopeasti normaalijakaumaa.

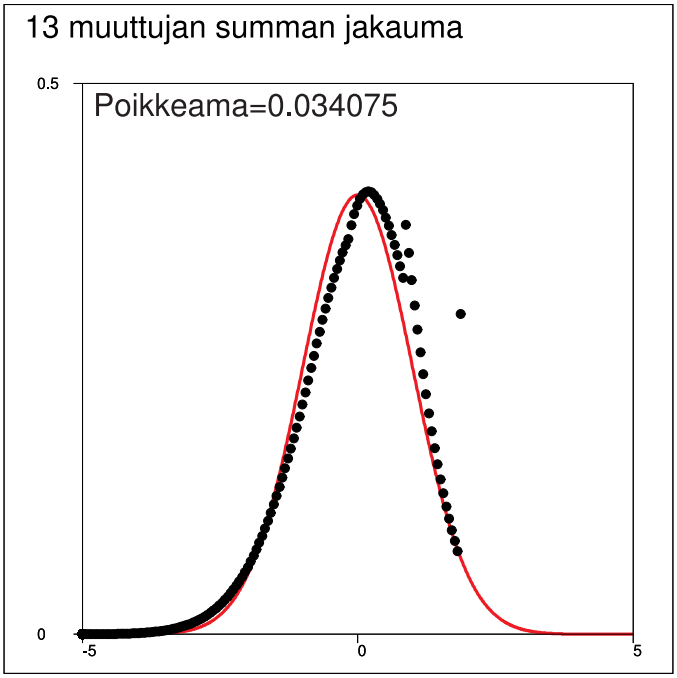
Keskeinen raja-arvolause toimii (näkymiä Survo-demosta)

Tässä perusjakauma on muuten diskreetti tasajakauma (20 pistettä), mutta viimeinen arvo on **50 kertaa** muita todennäköisempi (jakauma todella vino):









Lähestyminen on selvästi hitaampaa (etenkin jos otoskoko on kovin pieni).

Suhteellinen frekvenssi ja todennäköisyys

Sivulla 2 mainituista estimaattoreista on vielä käsittelemättä suhteellinen frekvenssi $\hat{p}$ (''p-hattu''), jota vastaava perusjoukon parametri on suhteellinen osuus tai todennäköisyys p.

Tyypillisiä suhteellisen frekvenssin sovelluksia ovat mm. puolueiden kannatusosuudet tai mitkä tahansa prosenttiosuuksien tarkastelut.

Yleisesti jonkin perusjoukon alkioilla joko **on** tai **ei ole** jotain ominaisuutta (esim. *''kannattaa''* tai *''ei kannata''*). Poimitaan nyt n riippumattoman havainnon satunnaisotos ja tarkastellaan havaintoyksiköitä, joilla on ko. ominaisuus. Niiden lukumäärä eli frekvenssi on f, ja vastaava suhteellinen frekvenssi $\hat{p} = f/n$.

Teeman 7 perusteella $f \sim Bin(n, p)$. Suhteellinen osuus p tulkitaan siis todennäköisyydeksi poimia perusjoukosta sellainen alkio, jolla on em. ominaisuus. Sivun 6 mukaan f:n odotusarvo on np ja hajonta $\sqrt{np(1-p)}$.

Suhteellisen frekvenssin otantajakauma

Suhteellisen frekvenssin $\hat{p}$ otantajakauma voitaisiin myös määrätä binomijakauman ominaisuuksien perusteella, mutta se johtaisi tässä yhteydessä tarpeettoman monimutkaisiin lausekkeisiin.

Tähänkin voidaan soveltaa keskeistä raja-arvolausetta, jonka mukaan suhteellisen frekvenssin $\hat{p}$ otantajakauma lähestyy normaalijakaumaa $N(p, \frac{p(1-p)}{n})$, kunhan n on ''riittävän'' suuri.

Odotusarvo on suhteellinen osuus p, mutta **keskivirhe** $\sqrt{\frac{p(1-p)}{n}}$ riippuu sekä suhteellisesta osuudesta p että otoskoosta n.

Tulos on siinä mielessä teoreettinen, että p on käytännössä tuntematon (aivan vastaavasti kuin σ edellä). Eteenpäin päästään korvaamalla nämä parametrit vastaavilla estimaattoreilla ($\hat{p}$ ja s).

Huomaa, että f:n alun perin *diskreettiä binomijakaumaa* **approksimoidaan** tässä yhteydessä *jatkuvalla normaalijakaumalla*.

Parametrien luottamusvälit

Yksittäinen estimaatti (kuten keskiarvo tai suhteellinen frekvenssi) ei sellaisenaan välttämättä ole mielekäs, koska se voi vaihdella huomattavasti otoksesta toiseen. **Luottamusväli** antaa paljon

konkreettisemman käsityksen parametrin todellisesta arvosta (ja samalla estimoinnin tarkkuudesta).

Luottamusvälillä tarkoitetaan sellaista otoksen havainnoista riippuvaa väliä, joka valitulla **luottamustasolla** (usein 95 %) peittää perusjoukon tuntemattoman parametrin arvon.

Luottamusvälejä käytetään monenlaisissa sovellustilanteissa tukemaan ja havainnollistamaan erilaisia estimointituloksia.

Seuraavassa tarkastellaan odotusarvon ja suhteellisen osuuden luottamusvälejä, jotka saadaan määrättyä kyseisten parametrien estimaattoreiden (siis keskiarvon ja suhteellisen frekvenssin) edellä esitettyjen otantajakaumien perusteella.

Odotusarvon luottamusväli

Odotusarvon μ luottamusväli luottamustasolla $1 - \alpha$ on

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}},$$

jossa $\bar{x}$ on keskiarvo, s on hajonta, n on otoskoko ja $z_{\alpha/2}$ on $N(0, 1)$-jakauman z-piste, joka vastaa todennäköisyyttä $\alpha/2$.

Lauseke $\frac{s}{\sqrt{n}}$ on keskiarvon hajonnan eli **keskivirheen** estimaattori, jossa perusjoukon hajonta σ on korvattu sen estimaattorilla s.

Luottamusväli peittää perusjoukon tuntemattoman odotusarvon todennäköisyydellä $1 - \alpha$, jossa α valitaan (usein $\alpha = 0.05$).

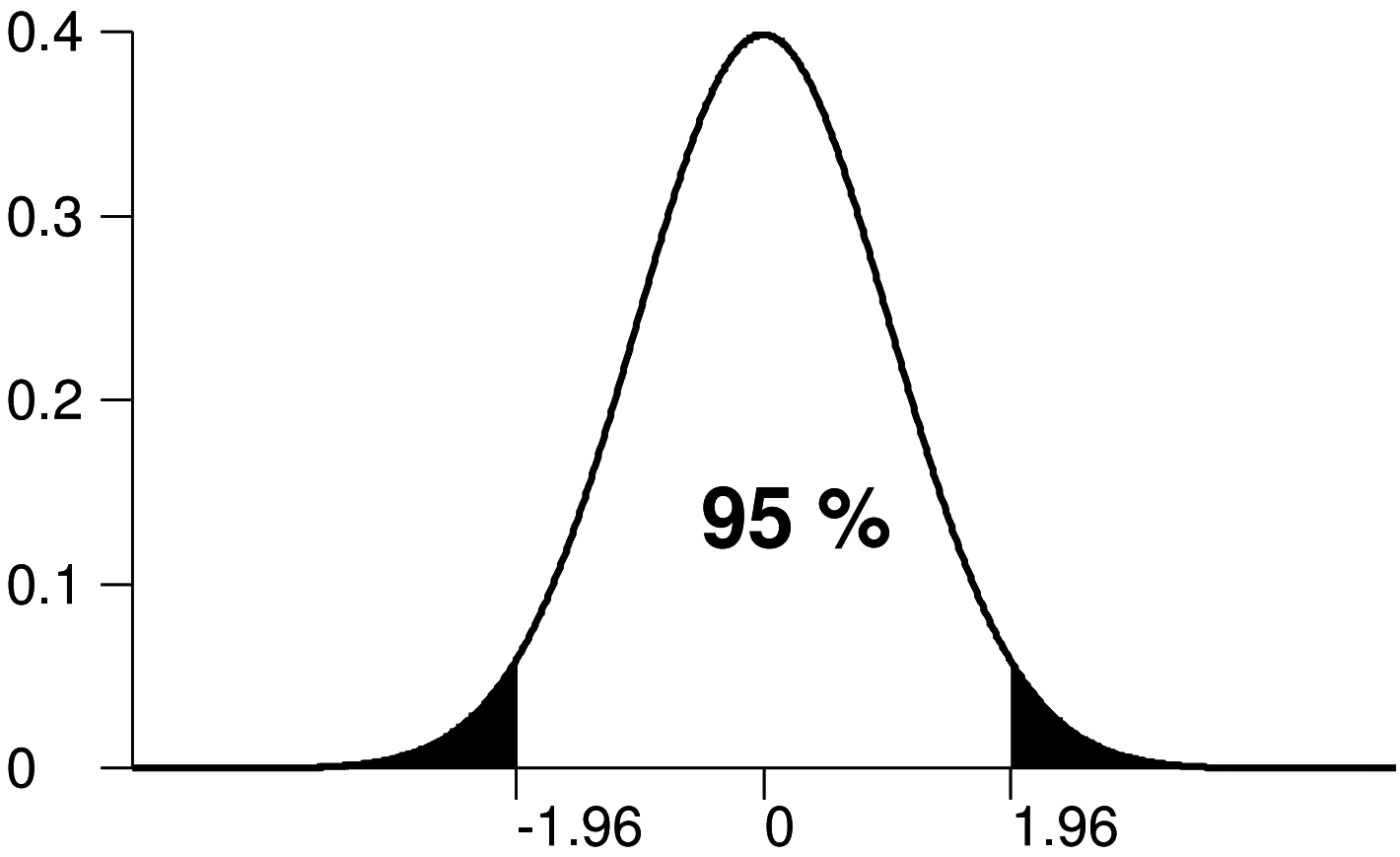
Todennäköisyyden frekvenssitulkinnan mukaan ajatellaan, että perusjoukosta poimitaan riippumattomia satunnaisotoksia, jolloin $100(1 - \alpha)$ % väleistä peittää odotusarvon μ. Täydellistä varmuutta ei saada, sillä 100α %:ssa otoksista näin ei käy.

Odotusarvon luottamusvälin tulkinta

Odotusarvon luottamusväli on siis

- symmetrinen väli keskipisteenään keskiarvo $\bar{x}$
- sitä kapeampi mitä suurempi on otoskoko n
- sitä leveämpi mitä suurempi on keskivirhe $\frac{s}{\sqrt{n}}$

Luottamusvälin leveyteen vaikuttaa myös haluttu luottamustaso. Usein $\alpha = 0.05$, ja luottamustaso siten $1 - 0.05 = 0.95$ (''**95 % luottamusväli**''). Kun otetaan jakauman molemmat ''hännät'' huomioon, tarkastellaan kohtaa $z_{\alpha/2}$ eli $z_{0.025}$. $N(0, 1)$-taulukon perusteella todennäköisyys 0.025 saavutetaan z:n arvolla -1.96. Jakauman keskellä (95 %) z on itseisarvoltaan tätä pienempi:



Vertaa Teema 4, s.1: ''keskiarvo $\pm$ 2 keskihajontaa kattaa noin 95 %''.

Suhteellisen osuuden luottamusväli

Ajatellaan jälleen, että jonkin perusjoukon alkioilla joko on tai ei ole jotain ominaisuutta, ja että p on niiden perusjoukon alkioiden suhteellinen osuus, joilla on tuo ominaisuus (ks. edellä).

Suhteellisen osuuden (tai todennäköisyyden) p luottamusväli luottamustasolla $1 - \alpha$ on

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

jossa $\hat{p}$ on suhteellinen frekvenssi, n on otoskoko ja $z_{\alpha/2}$ on $N(0,1)$-jakauman z-piste, joka vastaa todennäköisyyttä $\alpha/2$. Lauseke $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ on suhteellisen frekvenssin hajonnan eli **keskivirheen** estimaattori, jossa p on korvattu estimaattorilla $\hat{p}$.

Luottamusväli peittää perusjoukon tuntemattoman suhteellisen osuuden arvon todennäköisyydellä $1 - \alpha$, jossa α voidaan valita.

Suhteellisen osuuden luottamusvälin tulkinta

Suhteellisen osuuden luottamusväli on siis

- symmetrinen väli keskipisteenään suhteellinen frekvenssi $\hat{p}$
- sitä kapeampi mitä suurempi on otoskoko n
- sitä leveämpi mitä suurempi on keskivirhe $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Huomaa, että keskivirheeseen vaikuttaa $\hat{p}$: mitä lähempänä se on arvoa 0.5, sitä suurempi on hajonta. (Mitä enemmän $\hat{p}$ poikkeaa puolikkaasta, sitä enemmän tutkittavasta ilmiöstä tiedetään, jolloin hajonta on vastaavasti vähäisempää.)

Luottamusvälin leveyteen vaikuttaa myös valittu luottamustaso. Kuten edellä, käytännössä useimmiten valitaan $\alpha = 0.05$, jolloin luottamustasoksi tulee $1 - 0.05 = 0.95$ (''**95 % luottamusväli**'').

Kaikkiaan luottamusvälin leveyteen vaikuttavat varsin monet seikat, mikä on syytä muistaa luottamusvälien määräämisessä, käytössä ja erityisesti tulkinnassa.

Luottamusvälien perustelua ja pohdintaa

Esitetyt luottamusvälit ovat **approksimatiivisia:** niissä esiintyy normaalijakaumasta saatu kerroin $z_{\alpha/2}$. Perusteluina toimivat keskeinen raja-arvolause ja ''riittävän'' suuri otoskoko n.

Mikä sitten on käytännössä ''riittävän'' suuri otoskoko? Kysymys on vaikea, koska pelkkä otoskoko ei kuitenkaan ratkaise kaikkea tutkimukseen liittyvää epävarmuutta. Käytännössä joudutaan tasapainoilemaan mm. otoksen poimintakustannusten kanssa.

Erilaisia (enemmän tai vähemmän karkeita) arvioita voi tietenkin esittää. Tyypillisin ''vaatimus'' on, että $n > 30$, mutta ainakin suhteellisen osuuden yhteydessä se on liian yksinkertaistavaa. Joskus päättelyä tosin joudutaan tekemään pienistäkin otoksista.

Pienemmillä otoksilla voidaan normaalijakauman sijasta käyttää t-jakaumaa, johon palataan Teemassa 9. Toisaalta luottamusvälit ovat joka tapauksessa aika epäuskottavia, jos ne perustuvat liian pieniin havaintomääriin.

Johdattelua tilastolliseen hypoteesin testaukseen

Siirretään ajatuksia jo kohti **tilastollista hypoteesin testausta** tarkastelemalla eräitä taulukoita. Tässä kohdattaviin käsitteisiin perehdytään yksityiskohtaisemmin Teemassa 9.

Erään aiemman kurssin ikäjakauma tiedekunnittain: (lähde: WebOodi/KV)

	Ikä (vuotta)				
Tiedekunta	18–22	23–27	28–32	33–	yht.
Valtiotieteellinen	132	62	21	13	228
Matem.–luonnontiet.	47	30	9	6	92
Muut tiedekunnat	49	27	3	4	83
yhteensä	228	119	33	23	403

Aiemmin on tarkasteltu ehdollisia jakaumia, reunajakaumia ja erilaisia %-jakaumia. Esillä on ollut myös taulukko, jonka sarakeluokittelijana oli sukupuoli. Sen avulla tutkittiin ehdollista todennäköisyyttä ja riippumattomuutta.

Riippumattomuuden arviointi χ^2-testillä

Ovatko äskeisen taulukot rivi- ja sarakeluokittelijat toisistaan riippumattomia, ts. ovatko ikäjakaumat tiedekunnittain samanlaisia (kun keskitytään lähinnä kahteen kurssin kannalta suurimpaan tiedekuntaan ja samastetaan muut yhdeksi kategoriaksi)?

Muodollisemmin on kyse seuraavasta hypoteesin testauksesta
(H_0 on nollahypoteesi ja H_1 vaihtoehtoinen hypoteesi):

H_0 : Ikäjakaumat ovat samat tiedekunnasta riippumatta.
H_1 : Ikäjakaumissa on eroja tiedekunnittain.

Testaus tapahtuu yhteiskuntatieteissä hyvin yleisesti sovelletulla χ^2-testillä (''khi-toiseen'' tai ''khiin neliö'').

Tässä testi johtaa p-arvoon 0.604, joten H_0 jää voimaan. Ikäjakaumissa **ei ole** eroja tiedekunnittain.
Miten se ilmenee?
Tutkitaan tarkemmin eikä tyydytä pelkkään p-arvoon.
(Näihin käsitteisiin siis palataan tarkemmin Teemassa 9).

χ^2-testi: odotetut ja havaitut frekvenssit

Testin ideana on verrata toisiinsa **havaittuja** ja **odotettuja frekvenssejä**. Odotetut frekvenssit vastaavat H_0:n mukaista riippumatonta tilannetta, joten ne seuraavat aivan suoraan reunajakaumista. Esimerkiksi taulukon korostetun solun arvo on muodostettu laskemalla $\frac{33 \cdot 83}{403} = 6.7965$. Ohessa näin saadut **odotetut frekvenssit** on pyöristetty:

	Ikä (vuotta)				
Tiedekunta	18–22	23–27	28–32	33–	yht.
Valtiotieteellinen	129.0	67.3	18.7	13.0	228
Matem.–luonnontiet.	52.0	27.2	7.5	5.3	92
Muut tiedekunnat	47.0	24.5	**6.8**	4.7	**83**
yhteensä	228	119	**33**	23	**403**

χ^2-testi: kontribuutiot päättelyn ja elaboraation tukena

Odotettuja ja havaittuja frekvenssejä verrataan toisiinsa laskemalla niiden *neliöidyt erotukset* suhteessa odotettuihin frekvensseihin. Jokaista taulukon solua kohti saadaan kyseistä eroa kuvaava luku. Näitä lukuja kutsutaan toisinaan χ^2-kontribuutioiksi:

	Ikä (vuotta)				
Tiedekunta	18–22	23–27	28–32	33–	yht.
Valtiotieteellinen	0.07	0.42	0.29	0.00	0.78
Matem.–luonnontiet.	0.49	0.30	0.29	0.11	1.18
Muut tiedekunnat	0.09	0.25	**2.12**	0.11	**2.58**
yhteensä	0.65	0.97	**2.70**	0.22	4.54

Korostetut luvut paljastavat suurimmat erot (vrt. taulukot).

Mukana ovat myös χ^2-kontribuutioiden summat riveittäin ja sarakkeittain. Kokonaissumma 4.54 on χ^2-testisuure, jonka **vapausasteet** *(degrees of freedom)* ovat $df = (\text{rivien lkm} - 1) \times (\text{sarakkeiden lkm} - 1)$, kun summarivejä ei oteta mukaan, siis $(3-1)(4-1) = 6$. Testin p-arvo saadaan χ^2-jakauman kertymäfunktiosta (taulukosta tai ohjelmistojen avulla).

χ^2-testi, kun ''selviä'' eroja ilmenee

Vertailun vuoksi tarkastellaan myös tilannetta, jossa ikäjakaumat eroavat tiedekunnittain:

Toisen aiemman kurssin ikäjakauma tiedekunnittain (lähde: WebOodi/KV)

	Ikä (vuotta)				
Tiedekunta	18–22	23–27	28–32	33–	yht.
Valtiotieteellinen	86	43	6	12	147
Matem.–luonnontiet.	46	33	20	7	106
Muut tiedekunnat	35	28	5	4	72
yhteensä	167	104	31	23	325

Tässä tilanteessa testi johtaa p-arvoon 0.003, joten H_0 hylätään. Ikäjakaumissa **on eroja** tiedekunnittain. *Mitähän ne ovat??*

Päättely **ei saa jämähtää** vain (kvalitatiiviseen) henkäykseen: *''eroa on, p<0.005''* tms. Tutkijan on kyettävä sel(v)ittämään, mitä ja miten suuria erot ovat – ehkä jopa, mistä ne johtuvat!

χ^2-testi: odotetut ja havaitut frekvenssit

Tehdään vastaavat tarkastelut kuin edellä:

Odotetut frekvenssit vastaavat H_0:n mukaista riippumatonta tilannetta, joten ne seuraavat aivan suoraan reunajakaumista. Esimerkiksi taulukon ensimmäisen solun arvo saadaan laskemalla $\frac{167 \cdot 147}{325} = 75.5$.

	Ikä (vuotta)				
Tiedekunta	18–22	23–27	28–32	33–	yht.
Valtiotieteellinen	**75.5**	47.0	14.0	10.4	**147**
Matem.–luonnontiet.	54.5	33.9	10.1	7.5	106
Muut tiedekunnat	37.0	23.0	6.9	5.1	72
yhteensä	**167**	104	31	23	**325**

χ^2-testi: kontribuutiot päättelyn tukena

Lasketaan χ^2-kontribuutiot ja -testisuure vaiheittain:

$$\frac{(86-75.5)^2}{75.5} + \frac{(43-47)^2}{47.0} + \cdots + \frac{(4-5.1)^2}{5.1} = 1.45 + 0.35 + \cdots + 0.24 = 19.6$$

Muista, että pyöristetyillä luvuilla laskeminen on vaarallista: lopputulos voi olla yllättävän paljon pielessä (tässä laskelmat on oikeasti tehty tarkemmilla luvuilla).

Laskutoimitukset ovat tässä ainoastaan χ^2-kontribuutioiden havainnollistamisen vuoksi – käytännössä χ^2-testi (kuten kaikki muutkin analyysit) tehdään tietenkin tilastollisilla ohjelmistoilla.

Kerätään luvut taulukkoon (vastaavasti kuin edellä):

	Ikä (vuotta)				
Tiedekunta	18–22	23–27	28–32	33–	yht.
Valtiotieteellinen	1.45	0.35	**4.59**	0.25	**6.63**
Matem.–luonnontiet.	1.32	0.02	**9.67**	0.03	**11.05**
Muut tiedekunnat	0.11	1.07	0.51	0.24	1.92
yhteensä	2.87	1.44	**14.77**	0.51	19.60

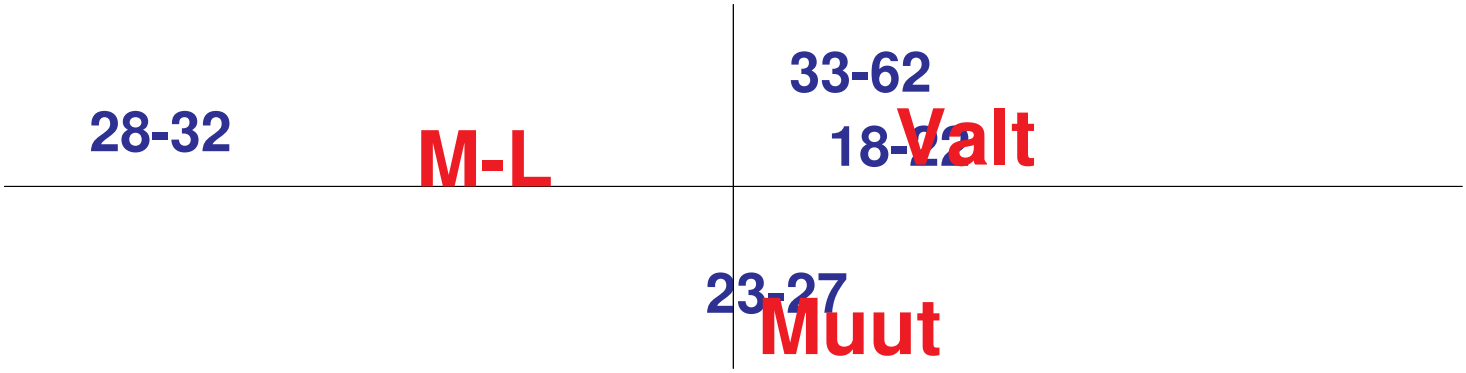
Korostetut luvut paljastavat, miksi H_0 hylätään (vrt. taulukot).

χ^2-testin visualisointi: korrespondenssianalyysi

Taulukoiden visualisointiin erikoistunut monimuuttujamenetelmä, **korrespondenssianalyysi**, on läheisessä yhteydessä χ^2-testiin. Tarkastellaan äskeisiä taulukoita vielä lyhyesti sen avulla (tällä kurssilla ei perehdytä menetelmän yksityiskohtiin):

Kevät 2008:

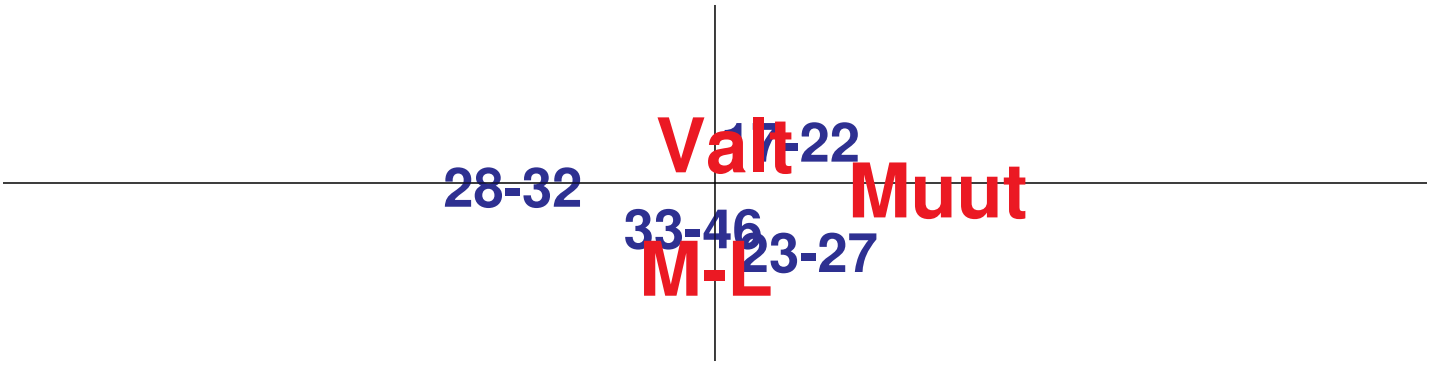
```
Correspondence analysis on data JK08K: Rows=3 Columns=4

    Canonical      Eigen-              Chi^2    Cumulative
    correlation    value                        percentage
1   0.2311         0.0534       17.3646554         88.61
2   0.0829         0.0069        2.23264005       100.00
                   0.0603      19.5973 (df=6 P=0.00326527)
```



Syksy 2008:

Correspondence analysis on data JK08S: Rows=3 Columns=4

	Canonical correlation	Eigen-value	Chi^2	Cumulative percentage
1	0.0899	0.0081	3.25890317	71.82
2	0.0563	0.0032	1.27855915	100.00
		0.0113	4.53746 (df=6 P=0.604347)	



Lisää aiheesta: *Kyselytutkimuksen mittarit ja menetelmät* (luku 7) ja kurssit monimuuttujamenetelmistä tai luokitteluaineistojen analysoinnista.

Esimerkki: graduarvosanat valtsikassa 2006–2010

Havaitut frekvenssit (lähde: tdk-neuvosto 4/2011, esityslistan liite, Alma):

Oppiaine	A	B	N	C	M	E	L	Yht
Viestintä	2	5	28	82	82	19	3	221
Sosiaalityö	0	15	47	70	58	18	3	211
Sosiologia	0	4	16	72	71	42	1	206
Valtio-oppi, maailma	2	16	47	55	48	15	1	184
Kansantaloustiede	0	4	11	33	66	44	4	162
Sosiaalipsykologia	0	16	18	39	50	25	4	152
Poliittinen historia	1	3	28	50	42	18	0	142
Sosiaalipolitiikka	0	7	16	40	39	14	2	118
Valtio-oppi, hallorg	2	6	24	33	29	7	0	101
Valtio-oppi, poltutk	2	6	28	14	23	11	1	85
Sos. ja kultt.antropologia	1	0	6	28	26	8	0	69
Kehitysmaatutkimus	1	4	4	22	19	7	0	57
Käytännöllinen filosofia	1	0	5	7	11	21	1	46
Talous- ja sosiaalihistoria	1	1	5	12	17	5	1	42
Tilastotiede	0	2	2	5	10	8	1	28
Yht	13	89	285	562	591	262	22	1824

Lähde: Valtiotieteellinen tiedekunta (2011).

Esimerkki: graduarvosanat valtsikassa 2006–2010

Odotetut frekvenssit (arvosanat oppiaineesta riippumattomia, summat kiinteät):

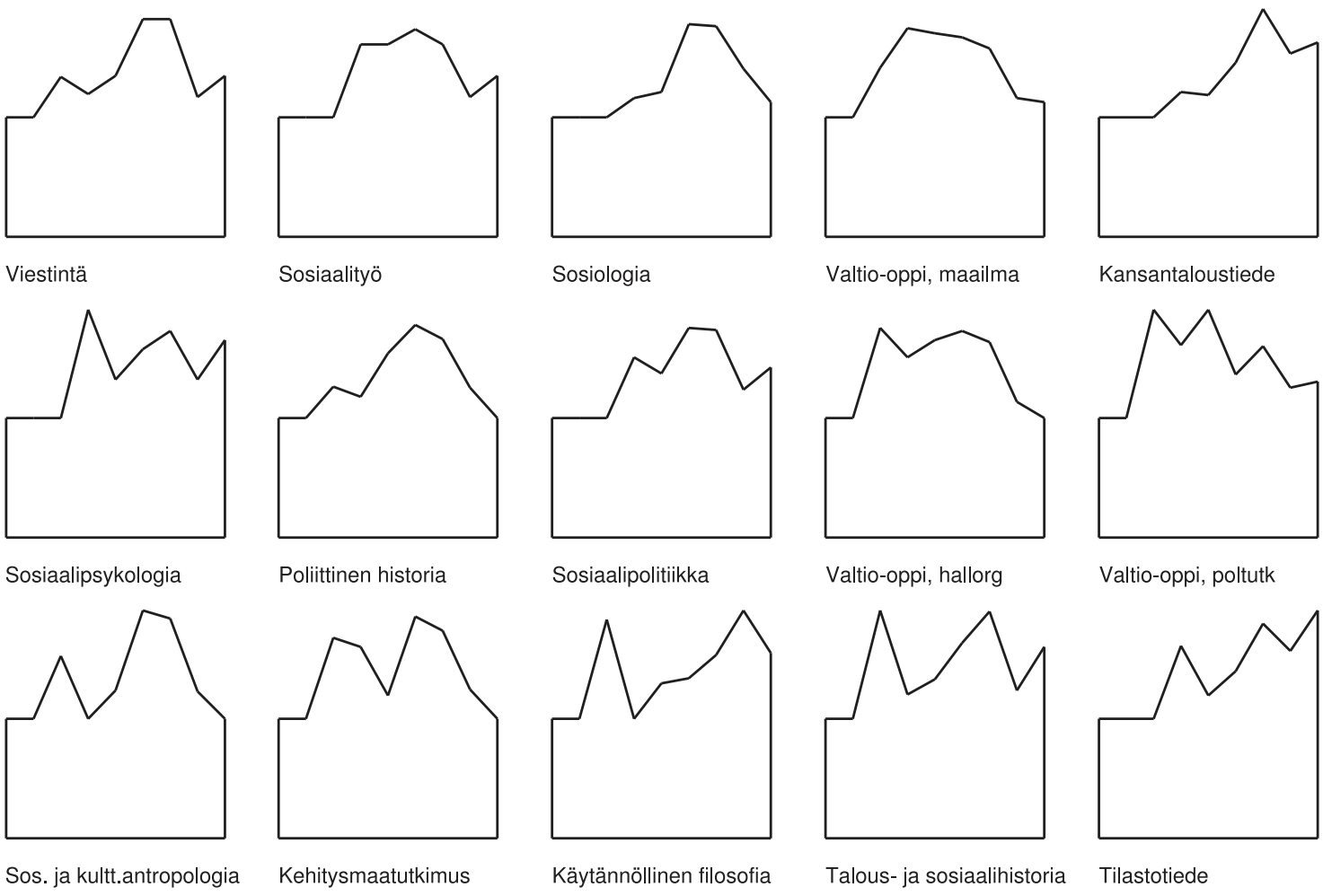
Oppiaine	A	B	N	C	M	E	L	Yht
Viestintä	2	11	35	68	72	**32**	3	221
Sosiaalityö	2	10	**33**	65	68	**30**	3	211
Sosiologia	1	10	**32**	63	67	**30**	2	206
Valtio-oppi, maailma	1	**9**	**29**	57	60	**26**	2	184
Kansantaloustiede	1	8	**25**	**50**	52	**23**	2	162
Sosiaalipsykologia	1	**7**	24	47	49	22	2	152
Poliittinen historia	1	7	22	44	46	20	2	142
Sosiaalipolitiikka	1	6	18	36	38	17	1	118
Valtio-oppi, hallorg	1	5	16	31	33	15	1	101
Valtio-oppi, poltutk	1	4	**13**	**26**	28	12	1	85
Sos. ja kultt.antropologia	0	3	11	21	22	10	1	69
Kehitysmaatutkimus	0	3	9	18	18	8	1	57
Käytännöllinen filosofia	0	2	7	14	15	**7**	1	46
Talous- ja sosiaalihistoria	0	2	7	13	14	6	1	42
Tilastotiede	0	1	4	9	9	4	0	28
Yht	13	89	285	562	591	262	22	1824

Esimerkki: graduarvosanat valtsikassa 2006-2010

Riviprosentit (arvosanojen %-jakauma oppiaineiden sisällä):

Oppiaine	A	B	N	C	M	E	L	Yht %
Viestintä	0.9	2.3	12.7	37.1	37.1	8.6	1.4	100
Sosiaalityö	0.0	7.1	22.3	33.2	27.5	8.5	1.4	100
Sosiologia	0.0	1.9	7.8	35.0	34.5	20.4	0.5	100
Valtio-oppi, maailma	1.1	8.7	25.5	29.9	26.1	8.2	0.5	100
Kansantaloustiede	0.0	2.5	6.8	20.4	40.7	27.2	2.5	100
Sosiaalipsykologia	0.0	10.5	11.8	25.7	32.9	16.4	2.6	100
Poliittinen historia	0.7	2.1	19.7	35.2	29.6	12.7	0.0	100
Sosiaalipolitiikka	0.0	5.9	13.6	33.9	33.1	11.9	1.7	100
Valtio-oppi, hallorg	2.0	5.9	23.8	32.7	28.7	6.9	0.0	100
Valtio-oppi, poltutk	2.4	7.1	32.9	16.5	27.1	12.9	1.2	100
Sos. ja kultt.antropologia	1.4	0.0	8.7	40.6	37.7	11.6	0.0	100
Kehitysmaatutkimus	1.8	7.0	7.0	38.6	33.3	12.3	0.0	100
Käytännöllinen filosofia	2.2	0.0	10.9	15.2	23.9	45.7	2.2	100
Talous- ja sosiaalihistoria	2.4	2.4	11.9	28.6	40.5	11.9	2.4	100
Tilastotiede	0.0	7.1	7.1	17.9	35.7	28.6	3.6	100
Yht %	0.7	4.9	15.6	30.8	32.4	14.4	1.2	100

Esimerkki: graduarvosanat valtsikassa 2006–2010



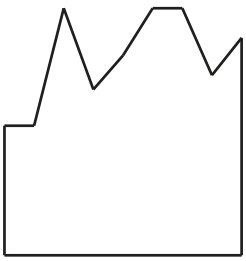
Esimerkki: graduarvosanat valtsikassa 2006–2010

Sarakeprosentit (arvosanojen %-jakauma oppiaineiden välillä):

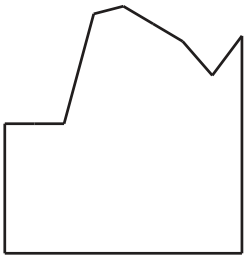
Oppiaine	A	B	N	C	M	E	L	Yht %
Viestintä	15.4	5.6	9.8	14.6	13.9	7.3	13.6	12.1
Sosiaalityö	0.0	16.9	16.5	12.5	9.8	6.9	13.6	11.6
Sosiologia	0.0	4.5	5.6	12.8	12.0	16.0	4.5	11.3
Valtio-oppi, maailma	15.4	18.0	16.5	9.8	8.1	5.7	4.5	10.1
Kansantaloustiede	0.0	4.5	3.9	5.9	11.2	16.8	18.2	8.9
Sosiaalipsykologia	0.0	18.0	6.3	6.9	8.5	9.5	18.2	8.3
Poliittinen historia	7.7	3.4	9.8	8.9	7.1	6.9	0.0	7.8
Sosiaalipolitiikka	0.0	7.9	5.6	7.1	6.6	5.3	9.1	6.5
Valtio-oppi, hallorg	15.4	6.7	8.4	5.9	4.9	2.7	0.0	5.5
Valtio-oppi, poltutk	15.4	6.7	9.8	2.5	3.9	4.2	4.5	4.7
Sos. ja kultt.antropologia	7.7	0.0	2.1	5.0	4.4	3.1	0.0	3.8
Kehitysmaatutkimus	7.7	4.5	1.4	3.9	3.2	2.7	0.0	3.1
Käytännöllinen filosofia	7.7	0.0	1.8	1.2	1.9	8.0	4.5	2.5
Talous- ja sosiaalihistoria	7.7	1.1	1.8	2.1	2.9	1.9	4.5	2.3
Tilastotiede	0.0	2.2	0.7	0.9	1.7	3.1	4.5	1.5
Yht %	100	100	100	100	100	100	100	100

Esimerkki: graduarvosanat valtsikassa 2006–2010

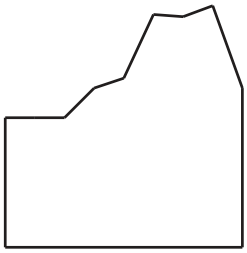
Graduarvosanaprofiilit oppiaineittain 2006 - 2010 (sarake-%):



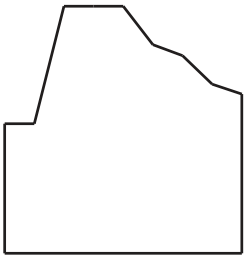
Viestintä



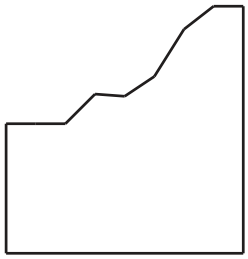
Sosiaalityö



Sosiologia

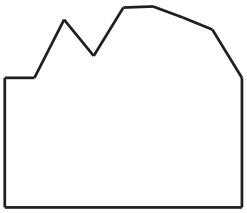


Valtio-oppi, maailma

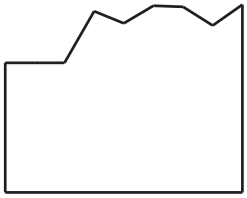


Kansantaloustiede

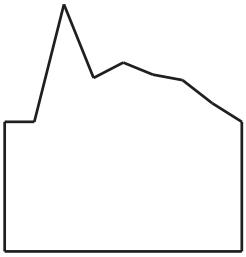
Sosiaalipsykologia



Poliittinen historia

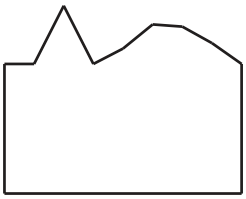


Sosiaalipolitiikka

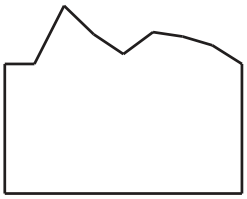


Valtio-oppi, hallorg

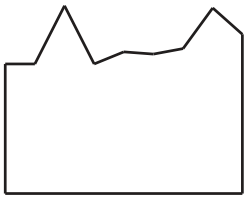
Valtio-oppi, poltutk



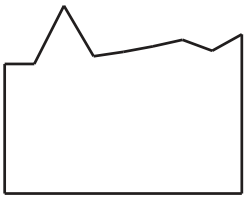
Sos. ja kultt.antropologia



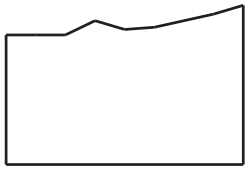
Kehitysmaatutkimus



Käytännöllinen filosofia



Talous- ja sosiaalihistoria



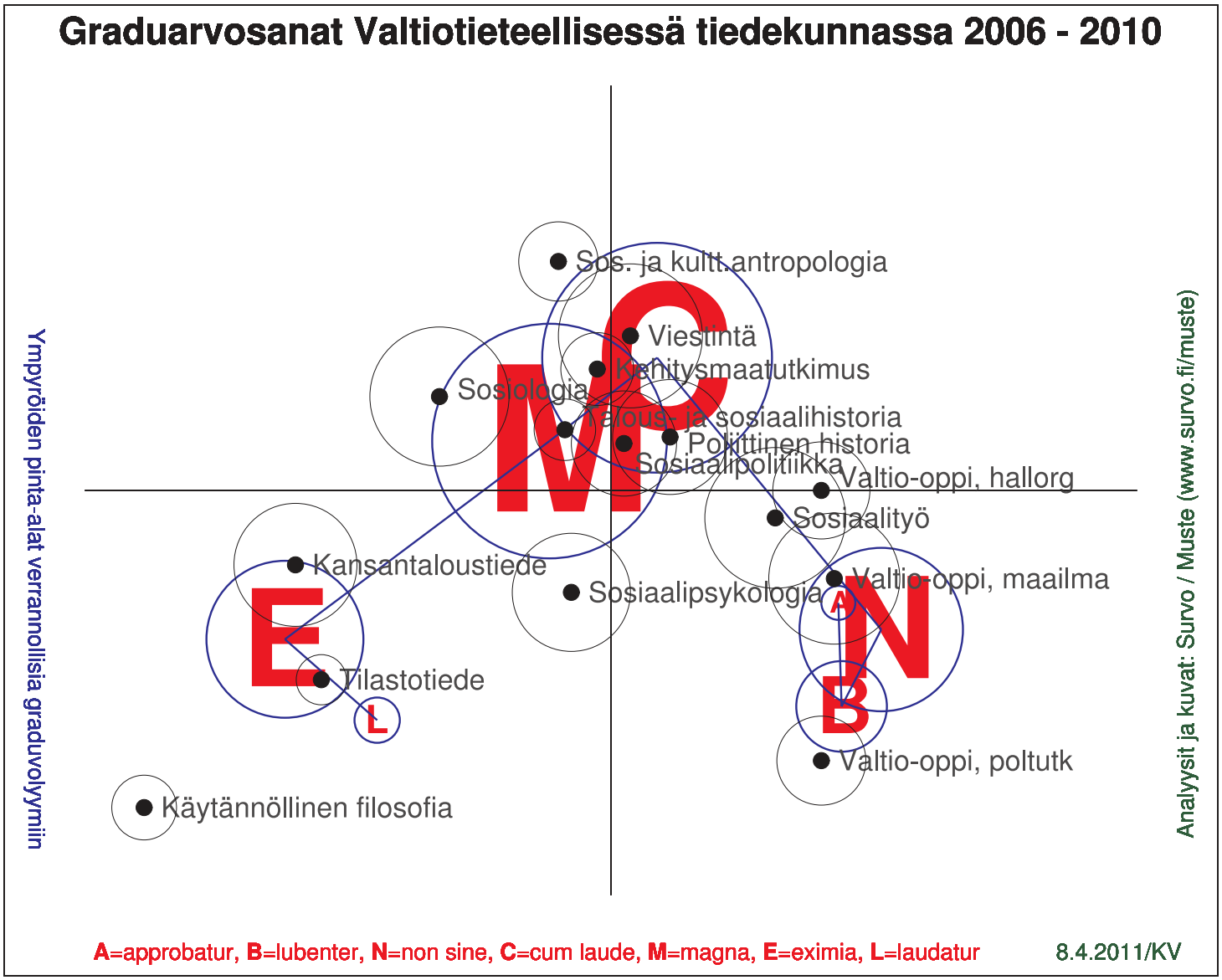
Tilastotiede

Esimerkki: graduarvosanat valtsikassa 2006–2010

Khi2-kontribuutiot (arvosanat riippuvat oppiaineesta; khi2=252.8(84), p=0.00):

Oppiaine	A	B	N	C	M	E	L	Yht
Viestintä	0.1	3.1	1.2	2.8	1.5	**5.1−**	0.0	14.0
Sosiaalityö	1.5	2.1	**6.0+**	0.4	1.6	**5.0−**	0.1	**16.7**
Sosiologia	1.5	3.6	**8.1−**	1.1	0.3	**5.2+**	0.9	**20.8**
Valtio-oppi, maailma	0.4	**5.5+**	**11.6+**	0.1	2.3	**4.9−**	0.7	**25.4**
Kansantaloustiede	1.2	1.9	**8.1−**	**5.7−**	3.5	**18.5+**	2.1	**41.0**
Sosiaalipsykologia	1.1	**9.9+**	1.4	1.3	0.0	0.5	2.6	**16.8**
Poliittinen historia	0.0	2.2	1.5	0.9	0.3	0.3	1.7	7.0
Sosiaalipolitiikka	0.8	0.3	0.3	0.4	0.0	0.5	0.2	2.6
Valtio-oppi, hallorg	2.3	0.2	4.3	0.1	0.4	3.9	1.2	12.4
Valtio-oppi, poltutk	3.2	0.8	**16.3+**	**5.7−**	0.7	0.1	0.0	**26.9**
Sos. ja kultt.antropologia	0.5	3.4	2.1	2.1	0.6	0.4	0.8	9.9
Kehitysmaatutkimus	0.9	0.5	2.7	1.1	0.0	0.2	0.7	6.1
Käytännöllinen filosofia	1.4	2.2	0.7	3.6	1.0	**31.4+**	0.4	**40.6**
Talous- ja sosiaalihistoria	1.6	0.5	0.4	0.1	0.8	0.2	0.5	4.1
Tilastotiede	0.2	0.3	1.3	1.5	0.1	3.9	1.3	8.6
Yht	16.6	**36.8**	**66.0**	27.0	13.2	**80.0**	13.2	252.8

Esimerkki: graduarvosanat valtsikassa 2006–2010



9. Tilastollinen hypoteesin testaus

Hypoteesit ovat perusjoukkoa koskevia väitteitä tai oletuksia, joita testataan otoksen avulla. Tilastollinen aineisto vastaa ''todisteita'', joiden tuella voidaan tehdä johtopäätöksiä *(vrt. rikostutkinta)*.

Johtopäätökset eivät ole ehdottoman selviä, koska niihin sisältyy erilaisia epävarmuuksia ja riskejä. Niiden todellista merkitystä on olennaista arvioida. Tilastollisessa testauksessa korostuu **otanta** ja siihen liittyvä epävarmuus. Toisen tyyppisiin tiedonkeruutapoihin perustuvissa asetelmissa on vielä enemmän epävarmuustekijöitä. Käytännön johtopäätöksiin vaikuttavat myös monet muut seikat.

Tilastollista testausta sovelletaan lähes kaikilla tieteenaloilla aivan **liian paljon**, **liian mekaanisesti** ja **liian kritiikittömästi**. Pahimmillaan (hyvissäkin lehdissä) varsinainen pointti saattaa hämärtyä *''tähdellisten''* testaustulosten loisteessa.

Sisällöllisen perustelun ja arvioinnin tulee aina olla etusijalla, koska pelkät tilastollisen testauksen tulokset eivät riitä todisteiksi eivätkä tutkimustuloksiksi.

Aihepiirin ongelmia, syitä ja seurauksia ovat ruotineet mm. Stephen T. Ziliak ja Deirdre N. McCloskey kirjassaan *''The Cult of Statistical Significance''* (2008).

Nollahypoteesi ja vaihtoehtoinen hypoteesi

Tutkimuskysymysten ja aiempien tutkimusten perusteella voidaan muotoilla **tutkimushypoteeseja** ja johtaa niistä tilastollisesti testattavia hypoteeseja, joita on yleensä kahden tyyppisiä:

Nollahypoteesi H_0:

- testattava oletus, aiempi tai vakiintunut käsitys
- tyypillisesti epäuskoinen kanta tutkimushypoteesiin
- sanamuoto tyyppiä ''ei eroa'', ''ei vaikutusta'' tms.

Vaihtoehtoinen hypoteesi H_1:

- vaihtoehto H_0:lle, vastakkainen tai uusi käsitys
- tyypillisesti tutkimushypoteesin mukainen väite
- sanamuoto tyyppiä ''eroa on'', ''vaikutusta on'' tms.

Hypoteesien sanamuodoista näkyy, että pelkkä H_0 vs H_1 painottaa vain sitä, **onko eroa vai ei**. Se harvemmin riittää, koska käytännössä kiinnostaa myös, *paljonko eroa on* tai *miten vahva vaikutus on*. Näihin kysymyksiin vastaaminen on olennaisen tärkeää, ja tilastollinen testaus on siinä vain yksi osa päättelyä. (Valitettavan usein näyttää siltä, kuin se olisi ainoa osa!)

Hypoteesin testauksen periaate ja kritiikki

Sanallisesti asetettuja hypoteeseja voidaan testata tilastollisesti erilaisilla *merkitsevyystesteillä*. Niiden periaate on seuraava:

1. Kootaan aineistopohjaiset ''todisteet'' **testisuureeksi**.
2. Tiivistetään testin tulos yhdeksi ainoaksi luvuksi, **p-arvoksi**.

P-arvo tarkoittaa *havaittua merkitsevyystasoa*, ts. miten vahvoja tilastollisia todisteita on H_0:aa vastaan ja siten H_1:n puolesta.

3. Tehdään testauksen perusteella *tilastollisia johtopäätöksiä*.

Jos todisteet riittävät, voidaan hylätä H_0 ja ajatella, että H_1 pätee. Jos todisteet eivät ole riittävät, H_1 ei saa tukea, ja H_0 jää voimaan.

Kohdat 1–3 vievät liian helposti kaiken huomion, mikä johtaa mekaaniseen päättelyyn ja laihoihin tutkimustuloksiin. **Ajatus on keskitettävä kohtaan 4:**

4. Perustellaan *(muttei vain testaustuloksilla)* tutkimusasetelman kannalta **olennaiset johtopäätökset** erojen, vaikutusten ym. **suuruudesta** ja arvioidaan niiden **käytännön merkitystä**.

Tässä tarvitaan **tutkimusalan asiantuntemusta**. Pelkkä tilastollinen testi, oli sen p-arvo mikä hyvänsä, ei ole pätevä todiste minkään tutkimusalan väitteille.

Esimerkki: jakaumien yhteensopivuuden testaus

Testaamalla voi arvioida, sopiiko jokin todennäköisyysjakauma tutkittavan ilmiön kuvaamiseen. Verrataan siis otoksesta havaitun ja teoreettisen jakauman yhteensopivuutta. Hypoteesit ovat:

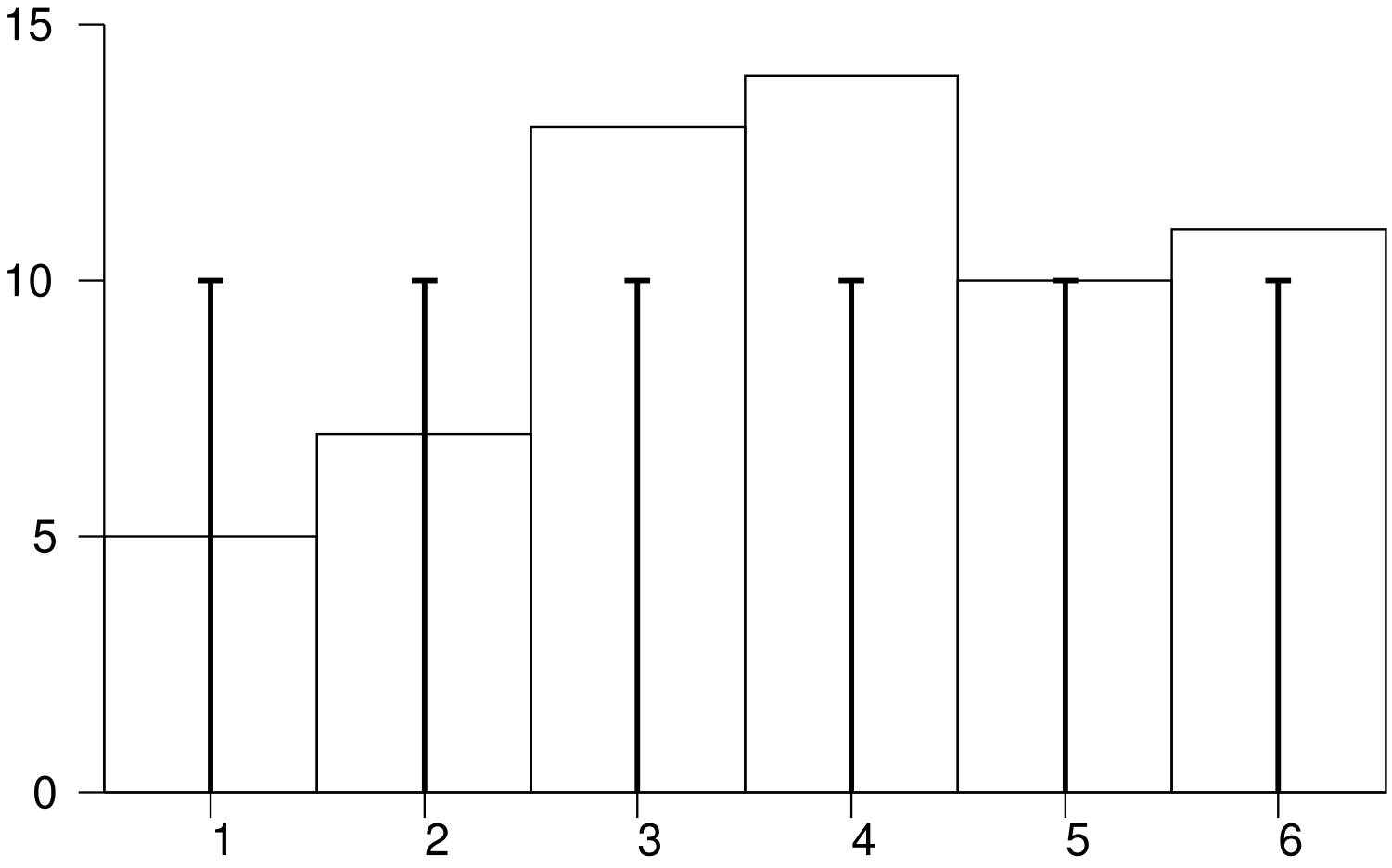
H_0 : Perusjoukon jakauma on diskreetti tasainen jakauma

$P(\text{"saadaan silmäluku } i\text{"}) = 1/6, \text{ kun } i = 1, 2, \ldots, 6.$

H_1 : Perusjoukon jakauma ei ole diskreetti tasainen jakauma.

Testataan tätä 60 nopanheittoa simuloivalla otoksella:

```
FILE MAKE NOPPA,1,60,X,1
VAR X=6*rand(2008)+1 TO NOPPA
GHISTO NOPPA,X,END+2 / X=0.5(1)6.5 XSCALE=0.5:?,1(1)6,6.5:?
pistetodennäköisyydet: FIT=MATRIX(NOPPA_H0)   YSCALE=0(5)15

Class midpoint  f     %  Sum     %      e    f  X^2
          1.0   5   8.3    5   8.3   10.0   5  2.5
          2.0   7  11.7   12  20.0   10.0   7  0.9
          3.0  13  21.7   25  41.7   10.0  13  0.9
          4.0  14  23.3   39  65.0   10.0  14  1.6
          5.0  10  16.7   49  81.7   10.0  10  0.0
          6.0  11  18.3   60 100.0   10.0  11  0.1
Fitted by MATRIX(NOPPA_H0) distribution
Chi-square=6.000 df=5 P=0.3062
```



Teoreettisen jakauman mukaan kunkin silmäluvun **odotettu** frekvenssi olisi 10. **Havaitut** vaihtelevat välillä [5,14]. Näiden poikkeamista muodostuu **testisuure**, jonka jakauman avulla saadaan testin **p-arvo** 0.3062. **Johtopäätös:** H_0 jää voimaan.

Johtopäätösten tekeminen p-arvon perusteella

Äskeisessä testauksessa saatu p-arvo ei tarjoa paljoa todisteita H_0:aa vastaan, joten voidaan pysyä sen mukaisessa oletuksessa. Otos on siis poimittu diskreetin tasaisen jakauman perusjoukosta.

Jos testausta **toistettaisiin**, esiintyisi nyt nähdyn kaltainen tilanne noin joka kolmas kerta (tähän viittaa 0.3:n luokkaa oleva p-arvo). Havaittu tilanne on siis tyypillinen. Poikkeamat teoreettisesta jakaumasta voidaan huoleti tulkita johtuvaksi satunnaisvaihtelusta (jota tähän on tuotettu [pseudo]satunnaislukugeneraattorilla).

P-arvoa kutsutaan siis **havaituksi merkitsevyystasoksi.** Tällä kertaa se jää niin korkeaksi, että testitulos **ei ole** niin sanotusti **tilastollisesti merkitsevä**, eli ''riittäviä todisteita'' H_0:aa vastaan ei otoksen perusteella ilmennyt, vaikka eroja havaittiinkin.

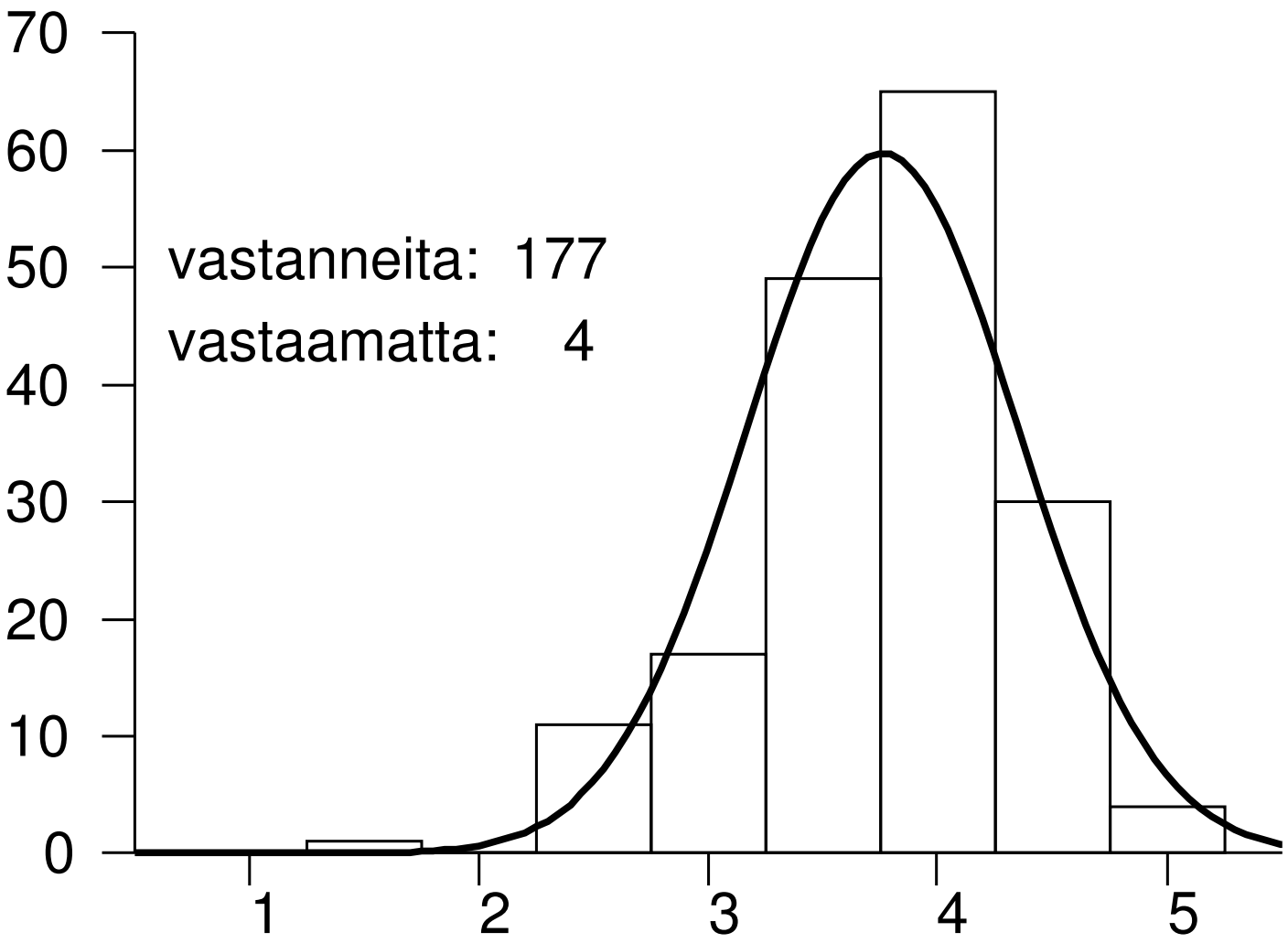
Toisinaan p-arvoa ajatellaan riskinä tehdä *hylkäämisvirhe* eli hylätä H_0 ''riittämättömin todistein''. Edellä tämä riski olisi liian suuri (30 %), mutta jälleen kysymys on vain otantaepävarmuuteen liittyvästä riskistä. Todelliset, päätöksentekotilanteisiin kytkeytyvät riskit ovat paljon monimutkaisempia eikä niitä pidä yksinkertaistaa tai yrittää hallita ainakaan pelkän p-arvon avulla.

Esimerkki: normaalijakauman yhteensopivuuden testaus

Useisiin tilastollisiin menetelmiin sisältyy normaalijakaumaoletus. Oletusta voidaan testata tilastollisesti. Tarkastellaan esimerkkinä tyytyväisyyttä asuinalueeseen (ks. Teema 4). Hypoteesit ovat:

H_0 : Perusjoukon jakauma on normaalijakauma.
H_1 : Perusjoukon jakauma ei ole normaalijakauma.

```
GHISTO KPT2001 TYYTALUE END+2 / TYYTALUE=0.75(0.5)5.25
               XSCALE=0.5:?,1(1)5,5.5:? YSCALE=0(10)70
tiheysfunktio (parametrit estimoidaan): FIT=NORMAL

Class midpoint  f     %  Sum     %       e     e   f   X^2
          1.0   0   0.0    0   0.0     0.0                
          1.5   1   0.6    1   0.6     0.1                
          2.0   0   0.0    1   0.6     1.0                
          2.5  11   6.2   12   6.8     7.0   8.1  12   1.9
          3.0  17   9.6   29  16.4    26.6  26.6  17   3.5
          3.5  49  27.7   78  44.1    52.2  52.2  49   0.2
          4.0  65  36.7  143  80.8    53.2  53.2  65   2.6
          4.5  30  16.9  173  97.7    28.3  28.3  30   0.1
          5.0   4   2.3  177 100.0     8.6   8.6   4   2.5
Fitted by NORMAL(3.7655,0.3504) distribution
Chi-square=10.77 df=3 P=0.0130
```



Testaustapa on sama kuin edellä, nyt vain jatkuva muuttuja on luokiteltu yhdeksään luokkaan. (Jakauman sovitus on yhdistänyt kolme ensimmäistä luokkaa neljänteen havaintojen pienen määrän vuoksi.) Odotetut ja havaitut frekvenssit poikkeavat sen verran, että p-arvo on 0.013. Todisteet tukevat sitä, että H_0 **hylätään**, eli perusjoukon jakauma ei ole normaalijakauma (H_1). Otos (jonka perusteella tämä siis päätellään) vaikuttaa kuvassakin melko vinolta.

Pohdintaa tilastollisesta merkitsevyydestä

Edellä johtopäätökset tehtiin havaitusta merkitsevyystasosta:
H_0 **jäi voimaan** $(p = 0.306) - H_0$ **hylättiin** $(p = 0.013)$.
Milloin todisteet riittävät? Onko tähän jokin yleinen sääntö?

Tilastollista *merkitsevyystestausta* kehittänyt *R.A.Fisher* esitti aikoinaan, että johtopäätökset testeistä tehtäisiin p-arvoista. Koska niiden laskeminen oli työlästä, Fisher tyytyi kompromissiin, jossa käytettiin valmiiksi laskettuja taulukoita *kolmelle merkitsevyystasolle*: $0.05, 0.01$ ja 0.001 (toisin sanoen 5 %, 1 % ja 0.1 %).

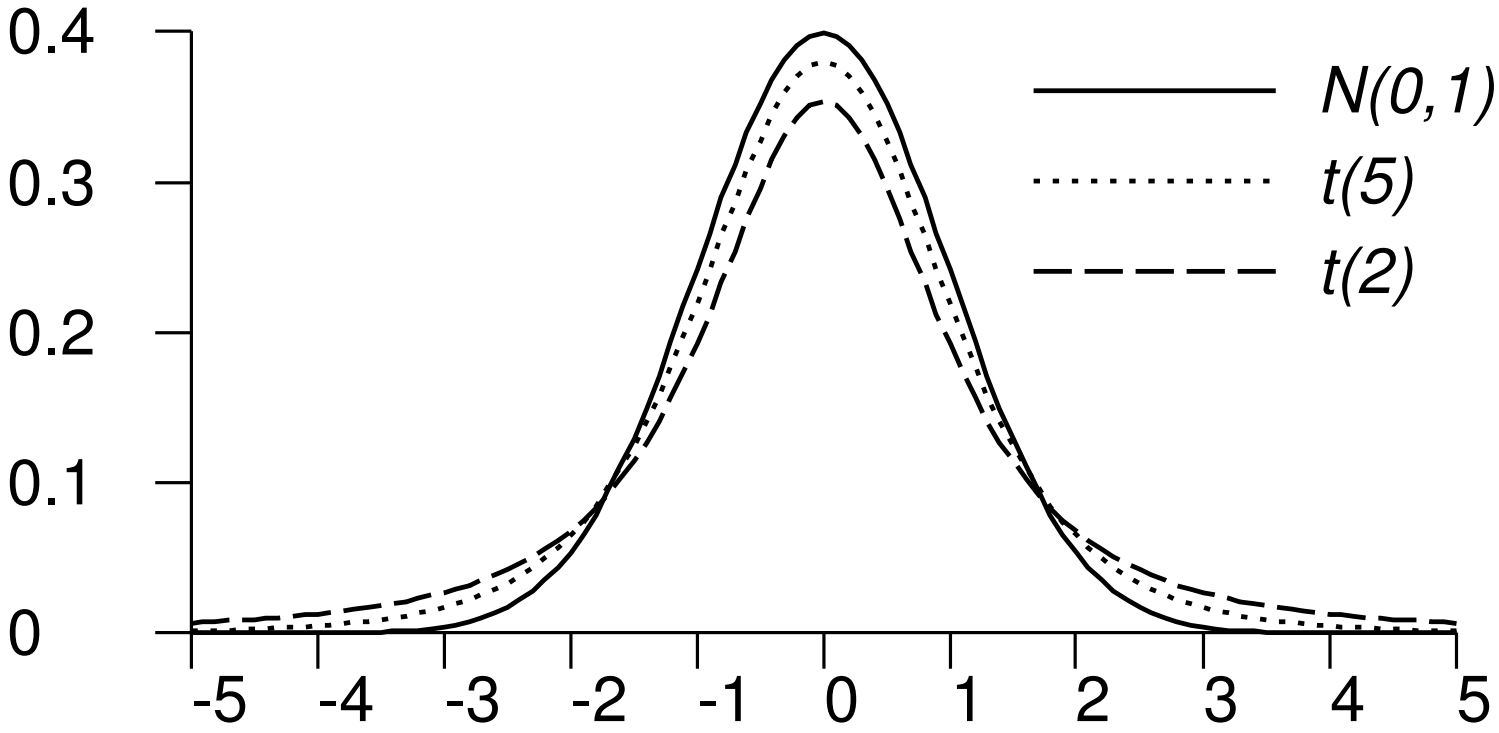
Osin samoihin aikoihin (1930-luvulla) *J.Neyman* ja *E.S.Pearson* kehittivät *kiinteään merkitsevyystasoon* nojaavaa *päätösteoriaa*, jossa valitaan etukäteen esim. $\alpha = 0.05$, ja hylätään H_0 mikäli α alittuu. (Vrt. luottamusvälit, joissa α valitaan samaan tapaan.)

Nykyisin ohjelmat kertovat p-arvot, mutta aikojen saatossa tasot (etenkin 5 %) ovat iskostuneet turhankin lujasti käytäntöön. Niitä sovelletaan usein **liian mekaanisesti**, esim. hylätään H_0 jos $p = 0.048$ muttei hylätä jos $p = 0.052$. Jälkimmäisen tyyppisissä tapauksissa tieteelliset lehdet saattavat jopa kieltäytyä julkaisemasta tutkimustuloksia vedoten siihen, ettei tilastollista näyttöä saatu riittävästi! *Onko tässä mitään järkeä?* Kyseessähän on olennaisesti sama luku, n. 5 % (ja pelkkä tilastollisen testin tulos). Lopulta **kaikki erot ovat tilastollisesti merkitseviä, jos otoskoko on riittävän suuri!**

Kokonaisuuden kannalta tärkeintä on **erojen todellinen suuruus** ja sen käytännön **merkitys** (relevanssi), ks. kohta 4 tämän teeman alussa olevassa luettelossa. Tilastollinen merkitsevyys *(eri sana!)* ei tätä takaa, ts. **pieni p-arvo ei ole tutkimustulos**. *Merkitsevä* ero ei ole siis käytännön kannalta *merkittävä*, ellei tutkija pysty perustelemaan sitä tutkimusalan sisällöllisistä lähtökohdista.

Entä jos otoskoko on hyvin pieni?

Tätä pohti yli 100 vuotta sitten Dublinissa panimomestari *W.S.Gosset*, joka vastasi **Guinnessin** olutpanimon laadunvalvonnasta. Hänen empiirinen tutkimuksensa johti (laadukkaan oluen ohella) merkittävään tulokseen, joka julkaistiin *Biometrika*-lehdessä (panimon sääntöjen vuoksi) salanimellä *Student* vuonna 1908. Kyseessä on t-testi *(käytännössä eniten sovellettu tilastollinen testi)* ja sen taustalla oleva t-jakauma, joka muistuttaa standardoitua normaalijakaumaa, mutta jolla on ''paksummat hännät'':



Kun normaalisti jakautuneen perusjoukon tuntematon hajonta σ korvataan otoshajonnalla s (vrt. Teema 8), saadaan t-testisuure, joka noudattaa t-jakaumaa *vapausastein* $n-1$. Vapausasteet liittyvät siis otoskokoon.

Kuvassa on t-jakaumia vapausastein 2, 5 ja ∞ (ääretön), joka on sama kuin $N(0,1)$. Käytännössä usein jo $n = 30$ riittää tämän ''äärettömyyden saavuttamiseen''.

Gossetin aineistot olivat usein paljon pienempiä, mikä oli hänen t-testinsä motivaatio.

Odotusarvoa koskevan hypoteesin testaus

Asetetaan nollahypoteesi

H_0 : Odotusarvo on jokin (ennalta määrätty) μ_0 (lyhyemmin: $\mu = \mu_0$).

Testaukseen on kaksi tapaa *(sama kaava, eri oletukset!)*:

- **Oletetaan** perusjoukon jakaumaksi normaalijakauma. Testisuure on tällöin $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, ja pienempikin n riittää. Havaittu merkitsevyystaso saadaan t-jakauman avulla (esim. sen taulukosta).
- **Ei oleteta** perusjoukon jakaumasta mitään. Testisuure on tällöin $z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$,mutta n on hyvä olla suurempi. Havaittu merkitsevyystaso saadaan $N(0,1)$-jakauman avulla (esim. sen

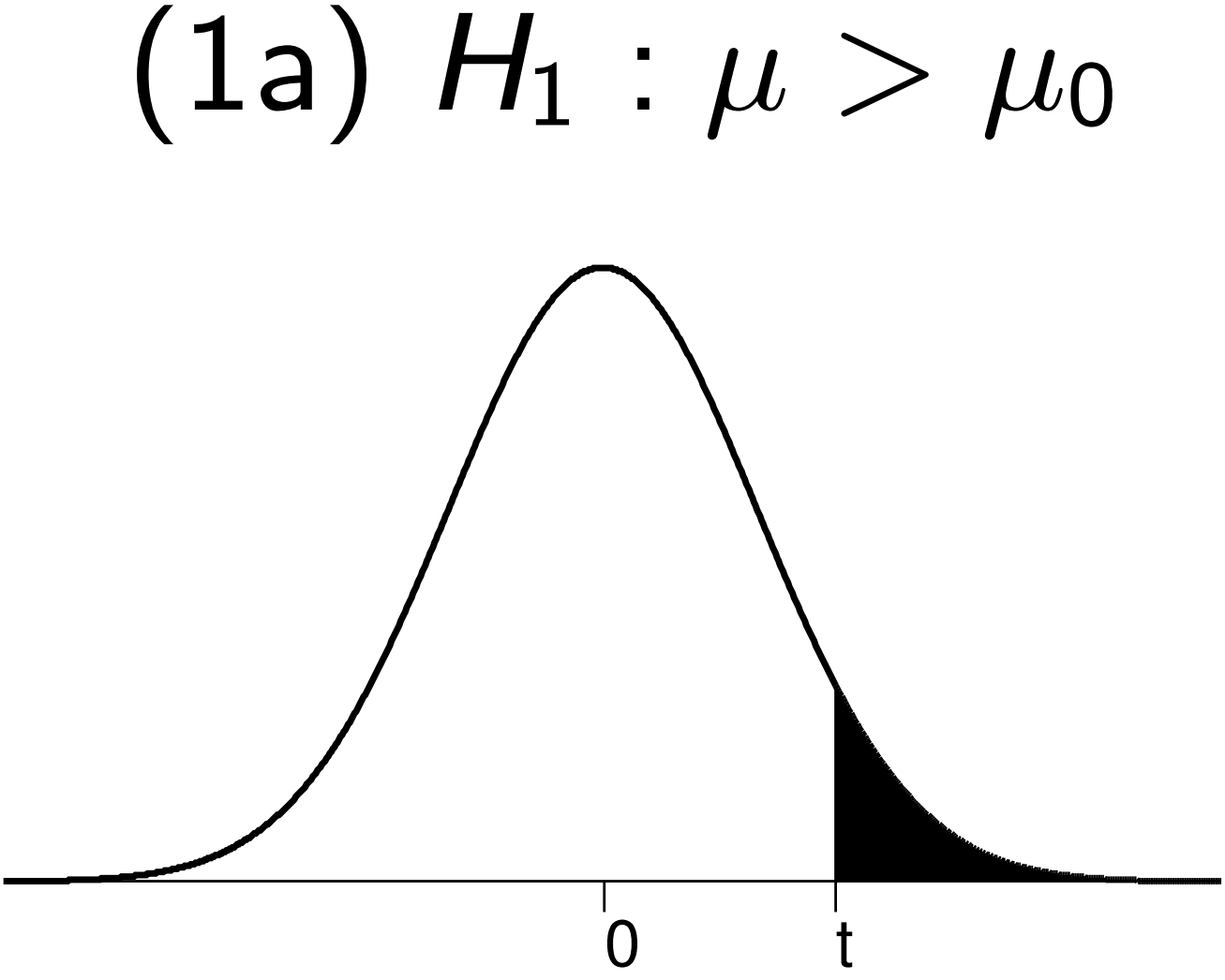
taulukosta).

Jos H_0 hylätään, niin johtopäätökset riippuvat testin **suunnasta**, jonka määrää vaihtoehtoinen hypoteesi H_1 (ks. seuraava sivu).

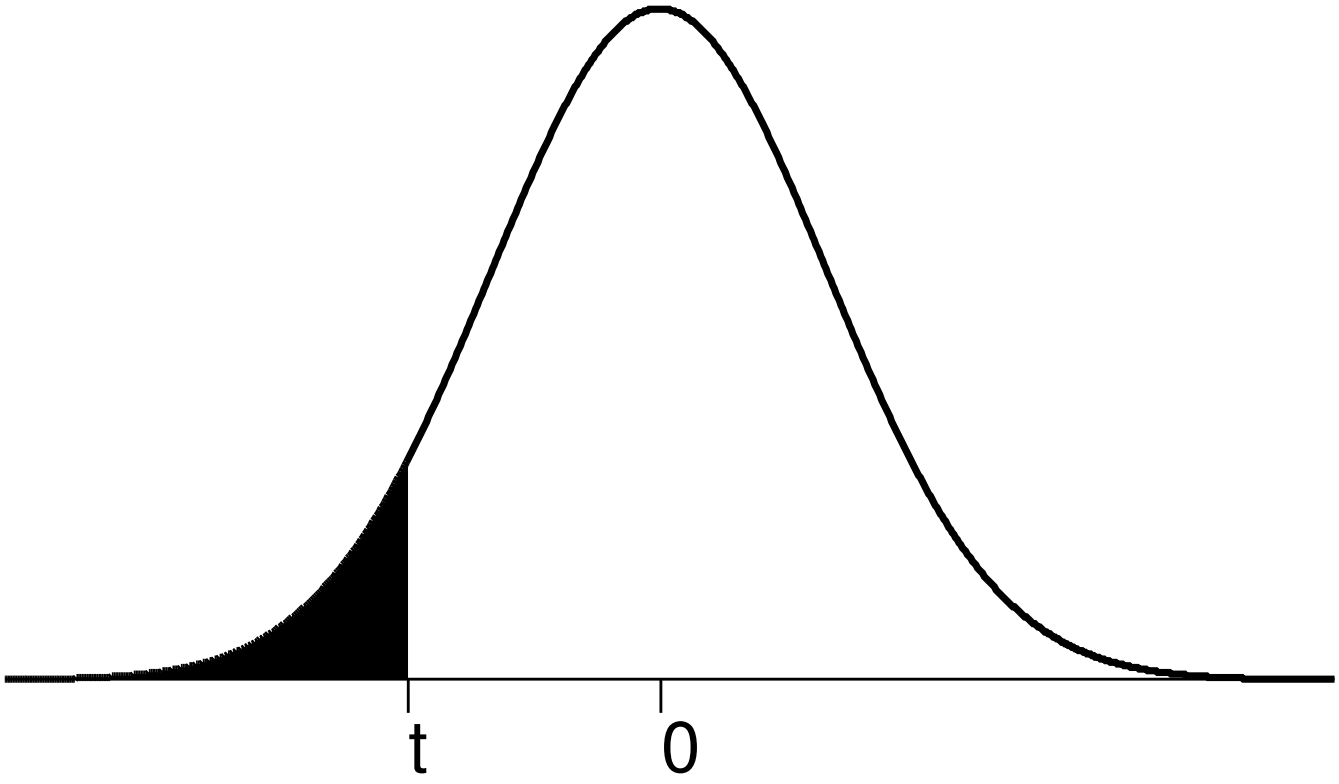
Vaihtoehtoinen hypoteesi ja testin suunta

Kun nollahypoteesi on edellä esitettyä tyyppiä $(\mu = \mu_0)$, niin vaihtoehtoinen hypoteesi H_1 on jokin kolmesta seuraavasta:

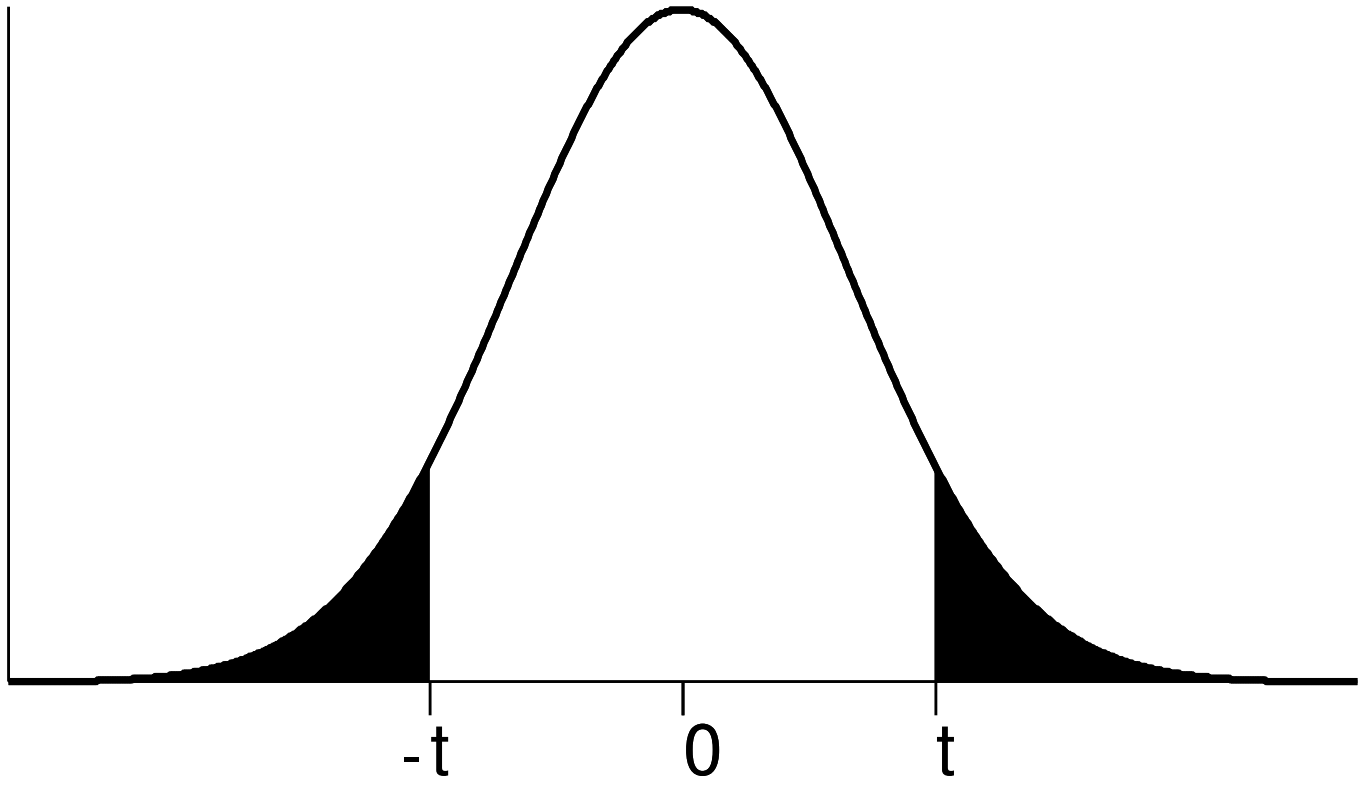
$(1\text{a})\ H_1 : \mu > \mu_0 \quad (1\text{b})\ H_1 : \mu < \mu_0 \quad (2)\ H_1 : \mu \neq \mu_0$



(1b) $H_1 : \mu < \mu_0$



$$(2)\ H_1 : \mu \neq \mu_0$$



Kuvat havainnollistavat testien p-arvoja näissä kolmessa tilanteessa.

Hypoteeseja ja niitä vastaavia testejä kutsutaan tämän mukaisesti
yksisuuntaisiksi (1a) ja (1b) tai **kaksisuuntaisiksi** (2).

Sekä yksi- että kaksisuuntaisia hypoteeseja voidaan yhtä hyvin asettaa muissakin kuin odotusarvoa koskevissa testaustilanteissa.

Käytännössä yleisempiä ovat kaksisuuntaiset hypoteesit (ja testit). Yksisuuntainen edellyttää enemmän esitietoja tutkimuskohteesta.

Suhteellista osuutta koskevan hypoteesin testaus

Vastaavasti asetetaan suhteellista osuutta koskeva hypoteesi
H_0 : Suhteellinen osuus p on jokin p_0 (lyhyemmin: $p = p_0$).

Hypoteesin testaukseen voidaan käyttää testisuuretta

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}},$$

jossa otoskokoa koskevat samat huomautukset kuin Teemassa 8. Myös lausekkeen nimittäjä, suhteellisen frekvenssin keskivirhe, on tuttu luottamusvälien yhteydestä.

Testin havaittu merkitsevyystaso saadaan $N(0, 1)$-jakaumasta. Vaihtoehtoinen hypoteesi H_1 voi olla yksi- tai kaksisuuntainen.

Muita testausasetelmia

Edellä esitetyt testit yleistyvät kahden odotusarvon tai kahden suhteellisen osuuden vertailuun. Näissä tilanteissa otokset voivat olla toisistaan riippumattomia tai ne voivat riippua toisistaan (kuten esim. toistomittauksissa), jolloin testit ovat vähän erilaisia.

Kaikkiaan tässä esitetyt testit ovat luonteeltaan sellaisia, että niissä tehdään oletuksia perusjoukon parametreista tai todennäköisyysjakaumista. On helppo tehdä oletuksia, mutta niiden toteutumistakin on syytä testata. **Jos oletukset eivät päde, testit eivät välttämättä toimi luotettavasti.**

Yksi tapa vähentää oletuksia ovat ns. ei-parametriset testit. Esimerkiksi t-testin ei-parametrinen vastine tunnetaan nimellä *Mannin ja Whitneyn (ja Wilcoxonin) testi.* Ei-parametrisiin testeihin ei kuitenkaan perehdytä tällä kurssilla.

Katsaus tähänastiseen sisältöön (Teema 9)

Tähän mennessä on käyty läpi testaukseen liittyviä perusasioita

- **yleisellä tasolla:** tilastollisen merkitsevyystestauksen idea ja periaatteet (esimerkkinä empiirisen ja teoreettisen jakauman yhteensopivuuden testaus), käytännön ohjeita sekä vähän historiallista taustaa ja lisätietoa sekä *paljon kritiikkiä*
- **teknisellä tasolla:** testisuureita tilanteisiin, joissa testataan perusjoukon jotakin hypoteettista arvoa — joko odotusarvoa tai suhteellista osuutta.

Käytännössä testaus (kuten muutkin tilastolliset menetelmät) suoritetaan **tilastollisilla ohjelmistoilla**, mutta ohjelmien järkevä käyttö edellyttää **menetelmien perusteiden** ymmärtämistä. Siksi osassa 2 kaivaudutaan myös jonkin verran pintaa syvemmälle.

Ilman todennäköisyyslaskennan ja tilastollisen päättelyn perusteita käytännön menetelmäosaaminen jää helposti liian pinnalliseksi.

Katsaus kurssin loppuosan sisältöihin

Tästä eteenpäin asiat menisivät teknisellä tasolla vaikeammiksi, eikä niiden tarkastelu olisi tällä kurssilla tarkoituksenmukaista.

Sen sijaan tarkasteluja jatketaan käytännön tasolla, vastaavasti
kuin kurssin osassa 1, mutta nyt hyödyntäen osassa 2 opittuja teoriaperusteita. Seuraava vaihe olisi harjoitella näitä asioita itse tilastollisilla ohjelmistoilla, mutta se jätetään muille kursseille (esim. *Sosiaalitutkimuksen tilastolliset menetelmät*).

Kolme eniten käytettyä tilastollista menetelmää lienevät (jo Teeman 8 lopussa esitellyn χ^2-testin lisäksi)

- t-testi,
- regressioanalyysi ja
- varianssianalyysi.

Kaikkiin näihin tutustutaan vielä kurssin kuluessa, mutta nyt lähestymistavan painotus muuttuu: **keskitytään ohjelmien tulosteiden analysointiin**. Seuraavien sivujen taulukoihin ja tulosteisiin perehdytään sekä luennoilla että harjoituksissa.

Kahden odotusarvon vertailu (riippumattomat otokset)

Tutkitaan KPT2001-aineiston avulla, ovatko helsinkiläiset miehet ja naiset yhtä tyytyväisiä asuinalueeseensa. Käytetään aiemmin esillä ollutta summamuuttujaa TYYTALUE.

H_0 : Helsinkiläiset miehet ja naiset ovat yhtä tyytyväisiä $(\mu_1 = \mu_2)$.
H_1 : Tyytyväisyydessä on eroa sukupuolten välillä $(\mu_1 \neq \mu_2)$.

H_1 on siis kaksisuuntainen. Testataan tätä SPSS:n t-testiproseduurilla:
(Analyze – Compare Means – Independent-Samples T Test)

Group Statistics

	[k03] Vastaajan sukupuoli	N	Mean	Std. Deviation	Std. Error Mean
TYYTALUE	Mies	543	35.8435	5.74676	.24662
	Nainen	858	36.8846	5.78268	.19742

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
TYYTALUE	Equal variances assumed	.000	.983
	Equal variances not assumed		

Aluksi saadaan otoksesta lasketut tunnusluvut, esim. keskiarvot. Sitä seuraa *Levenen testi* ryhmien varianssien yhtäsuuruudelle, joka on t-testin oletus.

Kahden odotusarvon vertailu (jatkoa)

Seuraavaksi tulee varsinainen t-testin osuus. Tuloste sisältää testisuureen ym. tiedot sekä samojen että eri varianssien tilanteissa:

		t-test for Equality of Means			
		t	df	Sig. (2-tailed)	Mean Difference
TYYTALUE	Equal variances assumed	-3.291	1399	.001	-1.04115
	Equal variances not assumed	-3.296	1158.367	.001	-1.04115

Lopuksi tulostuu 95 % luottamusväli edellisen taulukon reunassa raportoidulle havaitulle erolle:

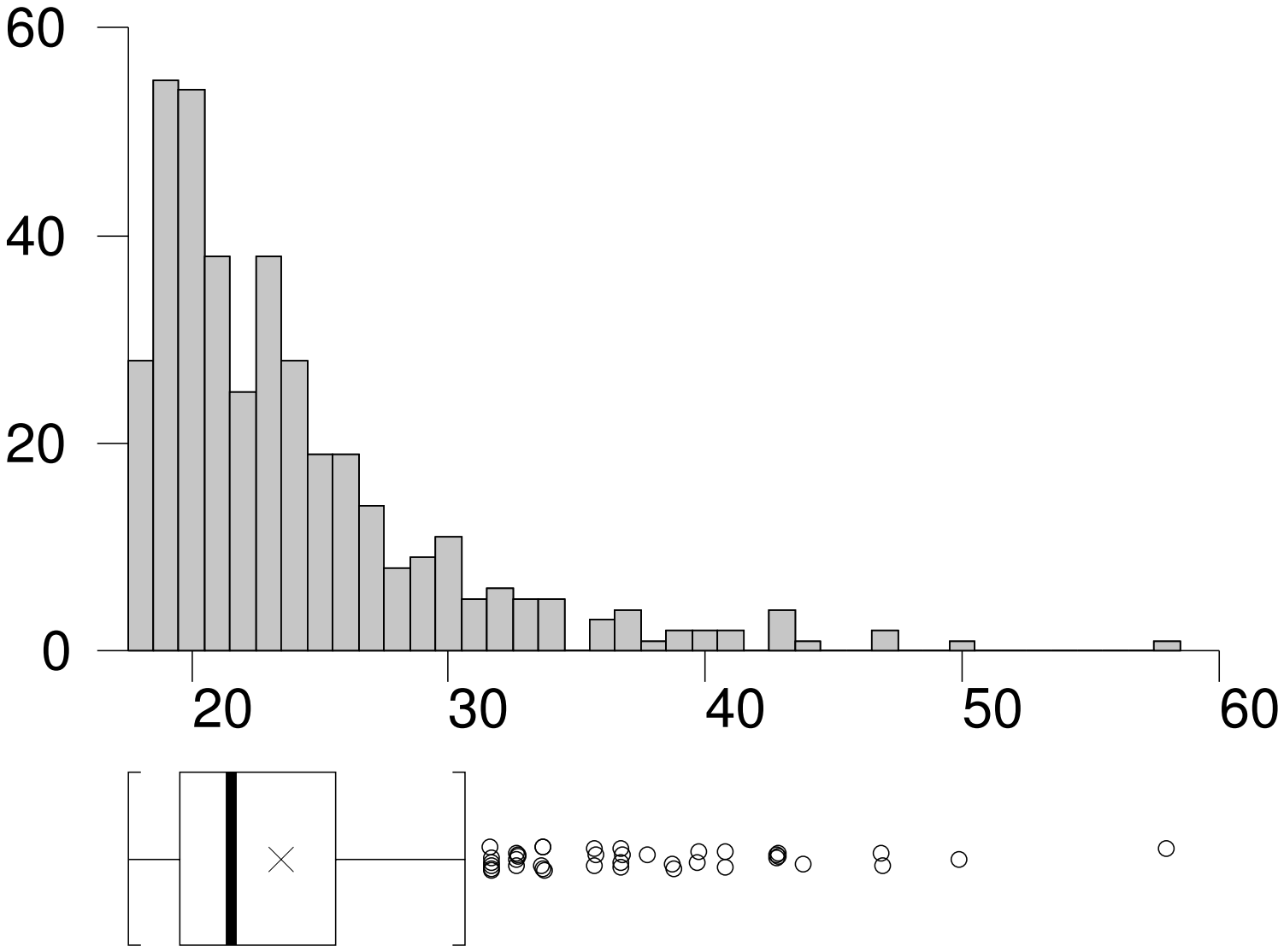
		t-test for Equality of Means		
		Std. Error Difference	95% Confidence Interval of the Difference	
			Lower	Upper
TYYTALUE	Equal variances assumed	.31634	-1.66171	-.42059
	Equal variances not assumed	.31590	-1.66096	-.42135

Odotusarvon vertailu ennalta määrättyyn arvoon

Syksyn 2009 kurssilaisten ikä (vuoden 2009 alussa):

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
ikä2009	390	23.89	5.957	.302



One-Sample Test

	Test Value = 23					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
ikä2009	2.967	389	.003	.895	.30	1.49

	Test Value = 24					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
ikä2009	-.349	389	.728	-.105	-.70	.49

	Test Value = 25					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
ikä2009	-3.664	389	.000	-1.105	-1.70	-.51

Kahden odotusarvon vertailu t-testillä

Syksyn 2009 kurssilaisten ikä (vuoden 2009 alussa), VT ja ML:

Group Statistics

	tdk	N	Mean	Std. Deviation	Std. Error Mean
ikä2009	VT	188	23.07	6.056	.442
	ML	131	24.40	5.301	.463

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
ikä2009	Equal variances assumed	.083	.773
	Equal variances not assumed		

t-test for Equality of Means			
t	df	Sig. (2-tailed)	Mean Difference
-2.018	317	.044	-1.322
-2.066	300.942	.040	-1.322

Independent Samples Test

		t-test for Equality of Means		
		Std. Error Difference	95% Confidence Interval of the Difference	
			Lower	Upper
ikä2009	Equal variances assumed	.655	-2.612	-.033
	Equal variances not assumed	.640	-2.582	-.063

Kahden odotusarvon vertailuja eri vuosina

Lomapäivät (*Prices and Earnings* 2006 ja 2009):

2006:

```
Independent samples            Non-OECD(Vacdays)   OECD(Vacdays)
Sample size                    30                  40
Mean                            18.16667            21.22500
Standard deviation              5.965899            5.040795
Student's t=-2.321 df=68 (P=0.0116 one-sided test)
Sum of ranks (R)               858                 1627
Mann-Whitney (U)               393                 807
(P=0.0070 one-sided Mann-Whitney, normal approximation)
Critical levels by simulation:
                Mean    R or U
Critical level 0.01232 0.00658  N=100000
Standard error 0.00035 0.00026
```

2009:

```
Independent samples            Non-OECD(Vacdays)   OECD(Vacdays)
Sample size                    33                  40
Mean                            18.75758            21.10000
Standard deviation              6.874001            5.592119
Student's t=-1.606 df=71 (P=0.0564 one-sided test)
Sum of ranks (R)               1085.5              1615.5
Mann-Whitney (U)               524.5               795.5
(P=0.0666 one-sided Mann-Whitney, normal approximation)
Critical levels by simulation:
                Mean    R or U
Critical level 0.05952 0.06742  N=100000
Standard error 0.00075 0.00079
```

Kahden odotusarvon vertailuja eri vuosina

Lomapäivät (*Prices and Earnings* 2006 ja 2009):

Oikea tapa (parittainen t-testi):

Paired comparisons:	Non-OECD	Non-OECD	
Samples: N=30	Vacdays(2006)	Vacdays(2009)	Difference
Mean	18.16667	18.80000	0.633333
Standard deviation	5.965899	7.038809	3.337182

Paired t=1.039 (P=0.8464 one-sided t test df=29)
Wilcoxon signed ranks test=0.734 (P=0.7686 normal approximation)
Critical levels by simulation:

	Differences	Signed rank	
Critical level	0.85890	0.76715	N=100000
Standard error	0.00110	0.00134	

Väärä tapa (miksi?):

Independent samples	Non-OECD Vacdays(2006)	Non-OECD Vacdays(2009)
Sample size	30	30
Mean	18.16667	18.80000
Standard deviation	5.965899	7.038809
Student's t=-0.376 df=58 (P=0.3542 one-sided test)		
Sum of ranks (R)	888.5	941.5
Mann-Whitney (U)	423.5	476.5

(P=0.3476 one-sided Mann-Whitney, normal approximation)
Critical levels by simulation:

	Mean	R or U	
Critical level	0.35943	0.34760	N=100000
Standard error	0.00152	0.00151	

10. Regressio- ja varianssianalyysi

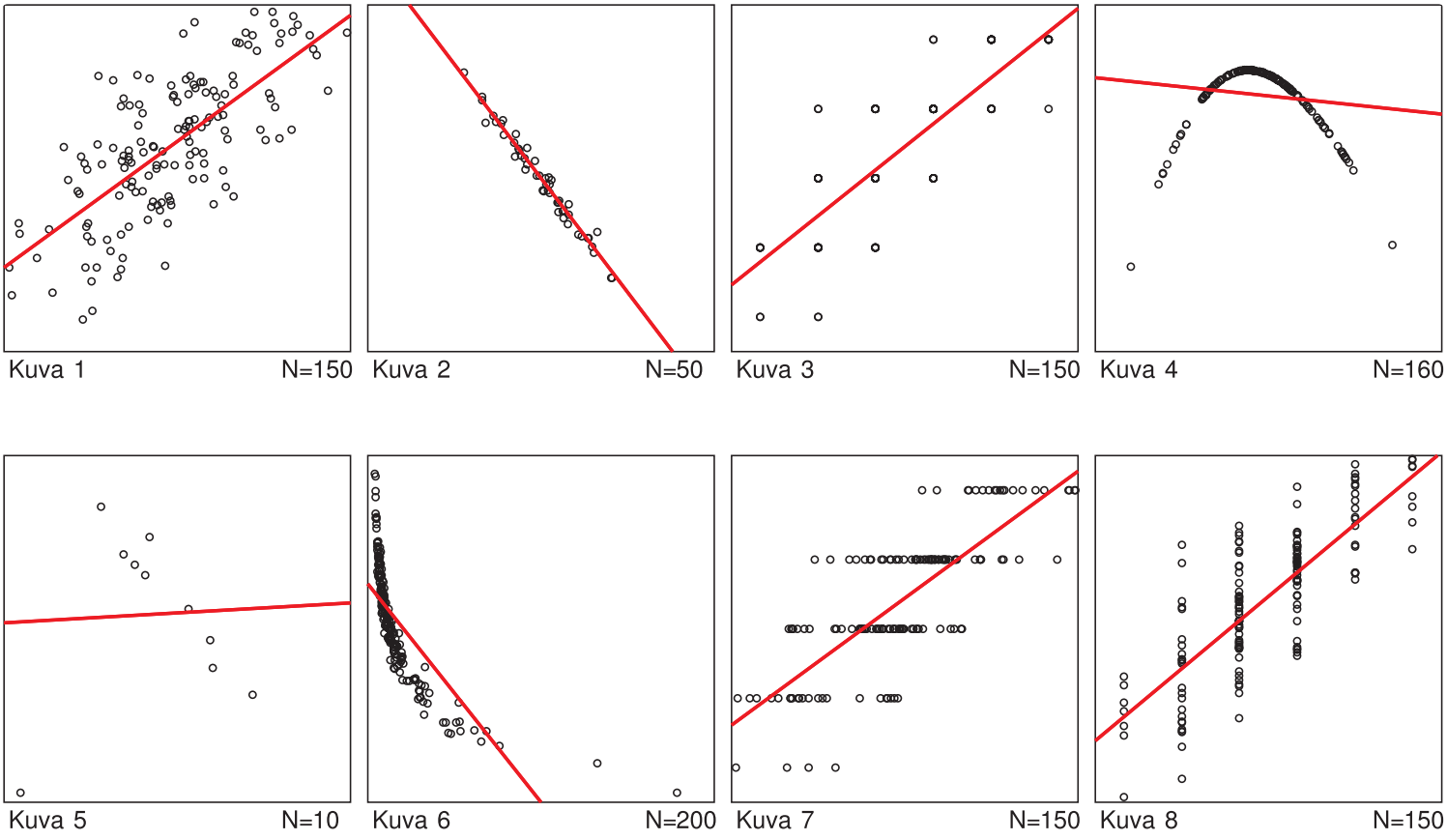
Kurssin päätteeksi tarkastellaan kahden yhteiskuntatieteissä tyypillisen tilastollisen menetelmän käytännön soveltamista.

Regressioanalyysi lienee t-testin ohella maailman eniten käytetty tilastollinen menetelmä. Sitä sivuttiin jo alustavasti Teemassa 4.

Varianssianalyysi liittyy useallakin tavalla regressioanalyysiin, ja on toisaalta t-testin yleistys silloin, kun vertaillaan useampaa kuin kahta ryhmää. Menetelmät kietoutuvat siis vahvasti toisiinsa.

Esimerkkejä hajontakuvista simuloiduilla aineistoilla

Mieleenpalautus Teemasta 4, nyt mukana regressiosuorat:



Miltei kaikissa kuvissa ilmenee jotain ongelmallisuuksia – mitä?

Regressioanalyysin asetelma

Regressioanalyysissa on kysymys riippuvuuden mallintamisesta, kun riippuvuus on **lineaarista**. Kuten Teema 4:n yhteydessä todettiin, äskeisistä vain kuvat 1 ja 2 ovat selvästi tämän oletuksen mukaisia.

Näissä yksinkertaisissa tilanteissa vastaavissa regressiomalleissa on vain **yksi selittäjä** (kuvissa vaaka-akselille piirretty muuttuja, jota kutsutaan usein x-muuttujaksi). Yleisesti mallissa voi olla useita selittäjiä, mutta vain **yksi selitettävä** (y-muuttuja).

Kuvassa 8 selittäjä on diskreetti, mikä ei sinänsä haittaa, kunhan sen luokilla on selvä järjestys. Muut kuvat ovat regressioanalyysin kannalta enemmän tai vähemmän kyseenalaisia:

- kuvat 3 ja 7: selitettävä muuttuja liian **diskreetti**
- kuvat 4 ja 6: riippuvuus aivan **epälineaarista**
- kuva 5: **poikkeava havainto** rikkoo lineaarisuuden

Perehdytään nyt regressioanalyysin käsitteistöön vähän tarkemmin.

Regressioanalyysin perusteet

Lineaarinen regressiomalli voidaan esittää lyhyesti muodossa
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$
jossa y on **selitettävä** (jatkuva) muuttuja, $x_1, x_2, \ldots, x_k$ ovat **selittäjiä**, ε on satunnainen **mallivirhe** ja $\beta_0, \beta_1, \ldots, \beta_k$ ovat **regressiokertoimia**, joiden arvot estimoidaan otoksen perusteella.

Joka selittäjällä on oma regressiokerroin. Niiden estimaatit toimivat painokertoimina, kun x:istä muodostetaan **sovite** (tai **ennuste**) $\hat{y}$. Tavoitteena on siis *selittää* (tai *ennustaa*) y vain x:ien avulla.

Kerrointa β_0 kutsutaan mallin **vakiotermin** kertoimeksi, jonka käytännön merkitys on yleensä melko vähäinen. Tulkinnat koskevat pääosin estimoituja regressiokertoimia $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.

Satunnaisvaihtelusta johtuen $\hat{y}$ (x:ien painotettu summa) ei ole (käytännössä koskaan) aivan sama kuin y. Mitä lähempänä se on, sitä korkeampi on **selitysaste** (ks. Teema 4).

Regressioanalyysin perusteet (jatkoa)

Regressioanalyysia käytetään paljon – valitettavasti myös väärin. Elleivät menetelmän oletukset päde (edes jollain tavalla), eivät tulokset tai johtopäätöksetkään voi olla kovin vakuuttavia. Sama pätee toki kaikkiin tilastollisiin menetelmiin (vrt. Teema 9).

Regressiomallin tärkeimmät oletukset koskevat mallivirhettä ε.
Siitä oletetaan yleensä, että $\varepsilon \sim N(0, \sigma^2)$, mikä tarkoittaa että mallivirhe tuo regressiomalliin vain satunnaista, samansuuruista, normaalisti jakautunutta vaihtelua nollan molemmin puolin.

Diskreeteille y-muuttujille soveltuva **logistinen regressioanalyysi** perustuu puolestaan binomijakaumaoletukseen (menetelmää ei käsitellä tällä kurssilla).

Regressioanalyysiin liittyy monenlaisia testaustilanteita, joissa käytetään mm. t-testiä. Jos normaalijakaumaoletus ei päde, testeistä tehdyt johtopäätökset menettävät luotettavuuttaan.

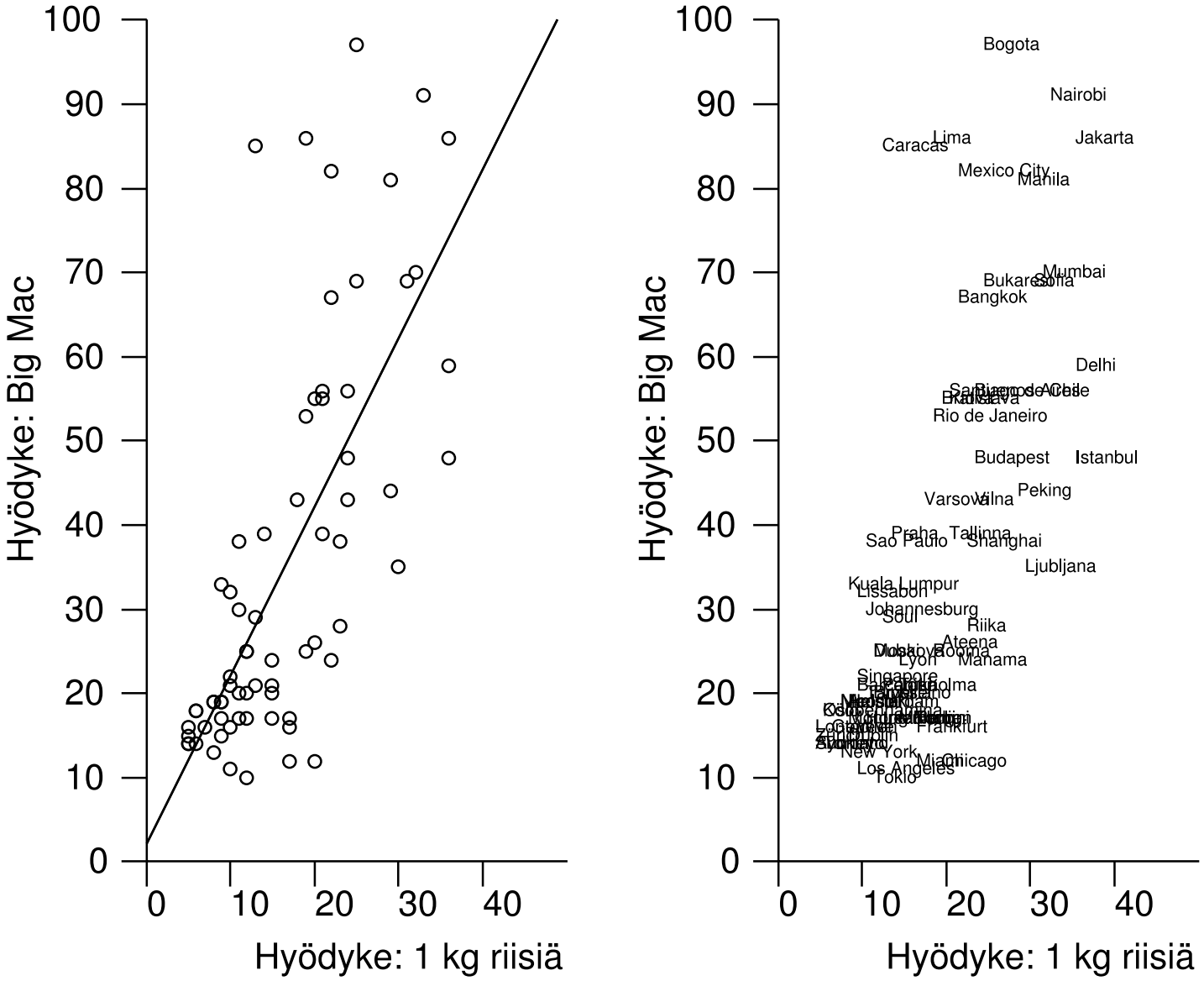
Mallivirhe on satunnaismuuttuja eikä sellaisenaan havaittavissa, mutta oletusten voimassaoloa tutkitaan mallin jäännöstermin eli **residuaalin** (y:n ja $\hat{y}$:n erotuksen) avulla.

Esimerkki: Prices and Earnings, BigMac

Tietoja 70 maasta ja kaupungista vuodelta 2006 (ks. Teema 4)

Kuva: ostovoima eri puolilla maailmaa

Muuttujat: *hyödykkeen ostoon vaadittu työaika minuutteina*



Selitetään nyt BigMac-indeksiä riisi-indeksillä (Rice):

```
Linear regression analysis: Data BIGMAC06, Regressand BigMac    N=70
Variable Regr.coeff.    Std.dev.    t     p
Rice      2.002298       0.237603  8.427  0.000
constant  2.204914       4.367544  0.505  0.308
Variance of regressand BigMac=557.2730849 df=69
Residual variance=276.6006701 df=68
R=0.7147 R^2=0.5108
```

Esimerkki jatkuu: tulkitaan tulostusta

Tulosteesta nähdään, että BigMac-indeksin vaihtelusta noin puolet (51 %) selittyy riisi-indeksillä:

```
Variable Regr.coeff.    Std.dev.    t      p
Rice      2.002298       0.237603  8.427  0.000
constant  2.204914       4.367544  0.505  0.308
Residual variance=276.6006701 df=68
R=0.7147 R^2=0.5108
```

Rice-muuttujan regressiokertoimen estimaatti on $\hat{\beta}_1 = 2$, eli kun Rice kasvaa yhden yksikön, BigMac kasvaa kaksi yksikköä. Tämä yhteys on nähtävissä myös kuvasta. Yksikkönähän molemmissa muuttujissa on minuutti (työaikaa).

Kertoimen $\hat{\beta}_1$ estimoitu hajonta (eli keskivirhe) on 0.237. Kerroin jaettuna keskivirheellään on t-testisuure (vrt. Teema 9) 8.427 (vapausasteita on 68). Tämän perusteella $H_0 : \beta_1 = 0$ voidaan hylätä, joten kerroin poikkeaa nollasta tilastollisesti merkitsevästi. Sisällöllistä tulkintaa olisi hyvä pohtia tarkemmin.

Vakiotermin kerroin on $\hat{\beta}_0 = 2.2$. Sillä on mallissa oma tehtävänsä, vaikkei se olekaan tilastollisesti merkitsevä (p=0.308).

Esimerkki jatkuu: lisätään toinen selittäjä

Lisätään malliin toiseksi selittäjäksi ruokakorin keskimääräistä hintaa kuvaava Basket (kotitalouden ruokamenot/kk, US\$):

```
Variable Regr.coeff.    Std.dev.    t      p
Rice      1.322144       0.269971  4.897  0.000
Basket   -0.074029       0.017947 -4.125  0.000
constant  42.84683       10.54302  4.064  0.000
Residual variance=223.8777762 df=67
R=0.7810 R^2=0.6099
```

Selitysaste paranee ja on nyt 61 %. Molemmat selittäjät ovat merkitseviä, mutta Rice menettää painoarvoaan ($\hat{\beta}_1 = 1.32$). *(Mistä arvelisit tämän johtuvan?)*

Basketin estimoitu kerroin $\hat{\beta}_2 = -0.07$ on pieni, koska muuttujan arvot ovat vastaavasti suurempia lukuja. Muuttujat voivat regressiomalleissa olla eri mittayksiköissä ja eri suuruusluokkaa.

Vapausasteet vähenevät, kun estimoidaan yksi parametri lisää (β_2).

Vakiotermikin muuttuu tilastollisesti merkitseväksi, vaikkei sillä seikalla edelleenkään ole mallin kannalta ihmeempää merkitystä.

Esimerkki jatkuu: lisätään kolmas selittäjä

Lisätään malliin vielä lomapäivien määrää kuvaava Vacdays, jota on tarkasteltu aiemmissa harjoituksissa:

```
Variable Regr.coeff.    Std.dev.    t      p
Rice      1.216663       0.279237  4.357  0.000
Basket   -0.077837       0.018052 -4.312  0.000
Vacdays  -0.450943       0.331290 -1.361  0.178
constant  55.08178       13.76280  4.002  0.000
Residual variance=221.0640444 df=66
R=0.7878 R^2=0.6206
```

Selitysaste paranee edelleen, mutta se ei ole ihme, sillä näin käy vaikka malliin lisättäisiin *mitä tahansa selittäjiä*. **Selitysaste ei sellaisenaan ole riittävä peruste hyvälle mallille.**

Huomaa, että Rice menettää taas painoarvoaan ja vapausasteet vähenevät yhdellä, kun β_3 estimoidaan aineistosta.

Vacdays-selittäjän ottamiselle malliin ei kuitenkaan ole yhtä selviä tilastollisia perusteluja, sillä sen estimoitu kerroin $\hat{\beta}_3 = -0.45$ ei ole tilastollisesti merkitsevä (p=0.178). $H_0 : \beta_3 = 0$ jää voimaan, Vacdays voidaan jättää pois ja palata aiempaan, suppeampaan malliin. **Sisällölliset perusteet voisivat kuitenkin puoltaa Vacdays-muuttujaa** (p-arvo ei tässäkään ratkaise kaikkea).

Esimerkki jatkuu: vaihdetaan kolmas selittäjä

Otetaan Vacdays:in tilalle malliin muuttuja Bread, joka kuvaa sitä, kuinka kauan pitää työskennellä voidakseen ostaa kilon leipää:

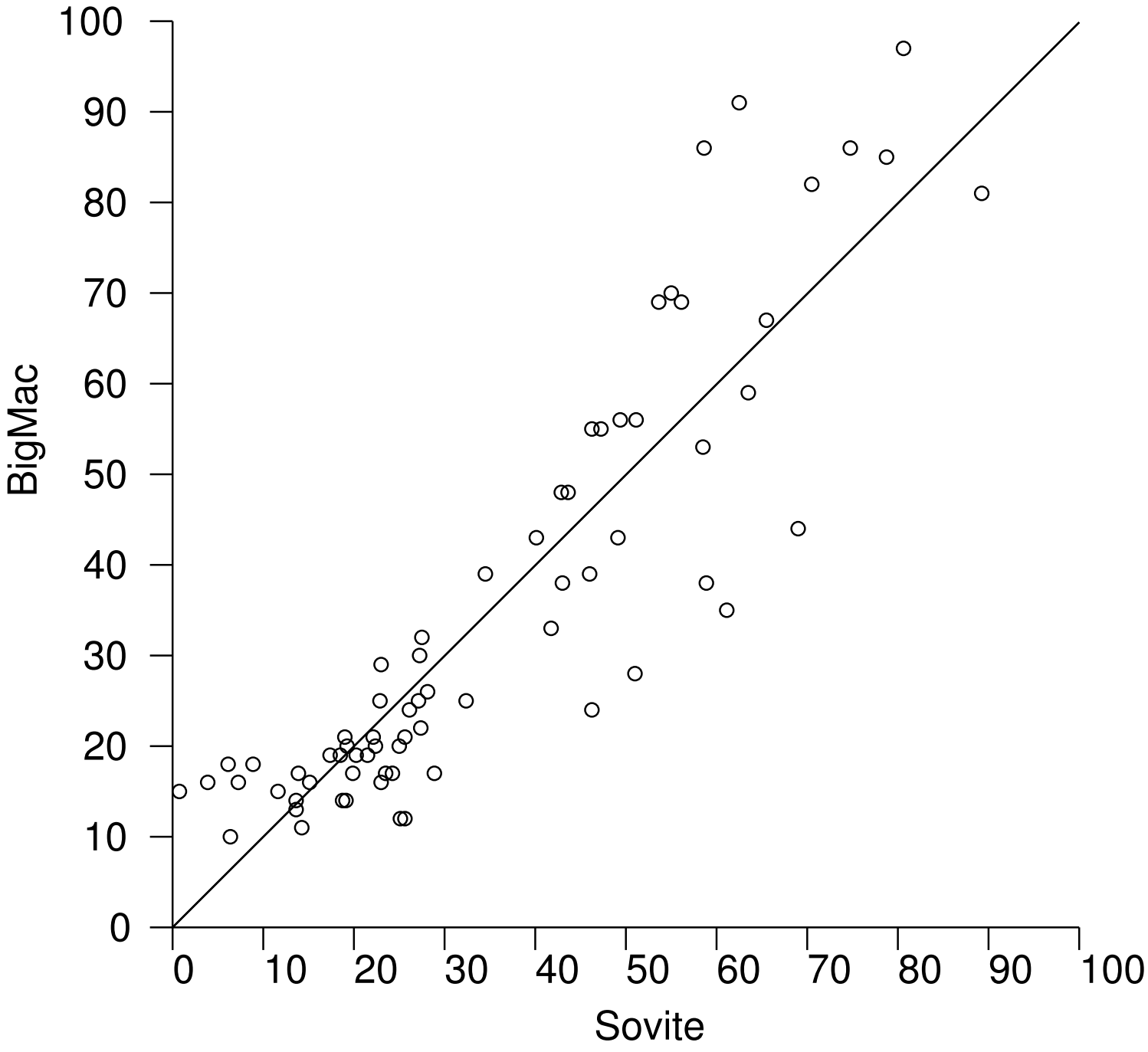
```
Variable Regr.coeff.    Std.dev.    t      p
Rice      0.802165       0.206413  3.886  0.000
Basket   -0.062346       0.013083 -4.766  0.000
Bread     0.826217       0.105148  7.858  0.000
constant  28.61589       7.844347  3.648  0.000
Residual variance=117.4217887 df=66
R=0.8936 R^2=0.7985
```

Ainakin selitysaste nousee selvästi: se on nyt lähes 80 %. Mallin rakentamista voisi jatkaa paljon huolellisemmin, mutta siirrytään tässä vaiheessa tutkimaan sitä hieman pintaa syvemmältä.

Piirretään muutama **diagnostiikkakuva**, joilla voidaan tutkia, miten regressiomallin oletukset toteutuvat tämän mallin osalta:

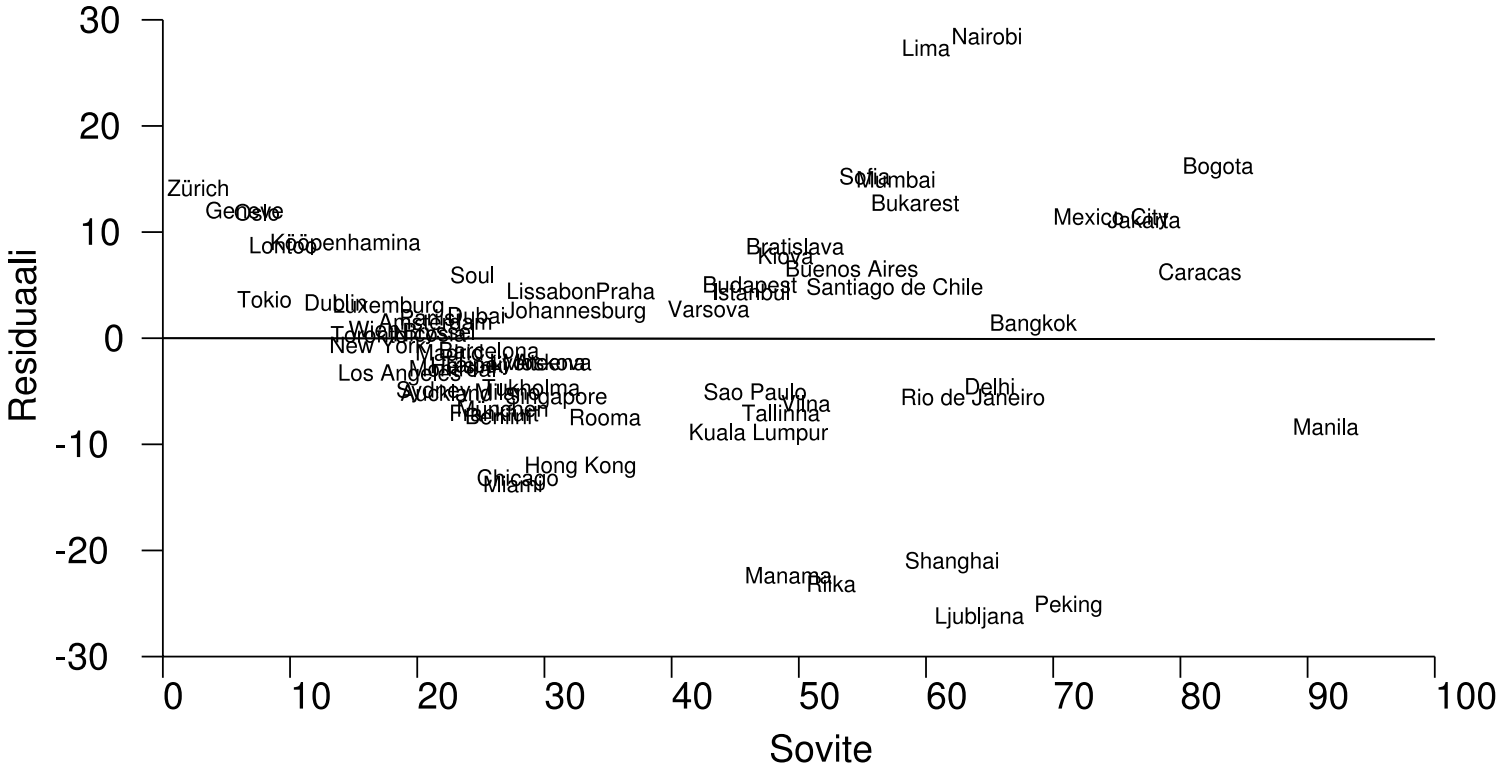
- sovite ja selitettävä muuttuja
- sovite ja residuaali (*jäännösvaihteludiagrammi*)
- residuaalin normaalisuus (*todennäköisyyspaperikuva*)
- residuaalin normaalisuus (histogrammi, normaalijakauma)

Diagnostiikkakuva 1: sovite ja selitettävä muuttuja



Mitä paremmin pisteet asettuvat suoralle, sitä lähempänä mallin perusteella muodostettu sovite on selitettävää muuttujaa.

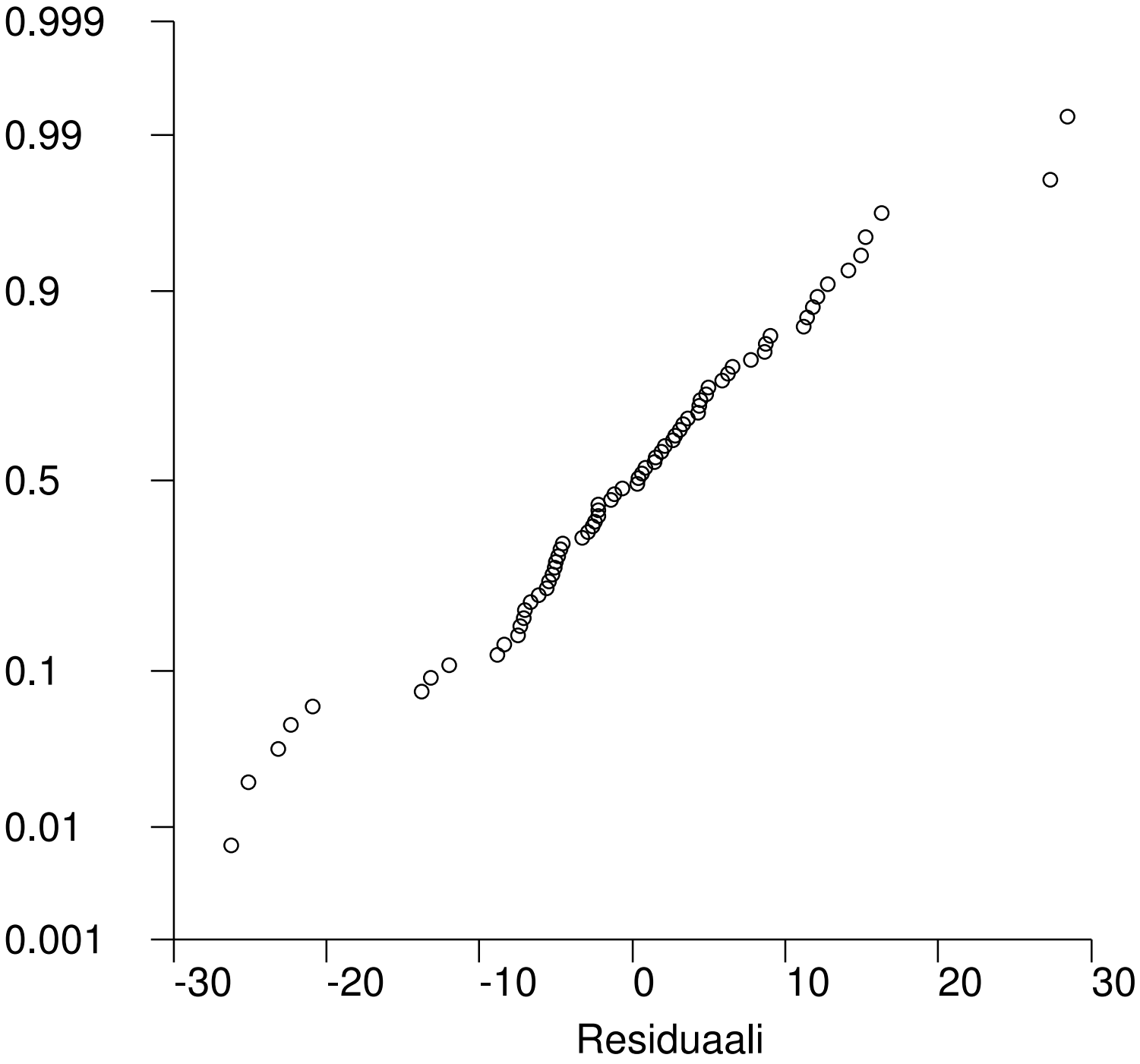
Diagnostiikkakuva 2: sovite ja residuaali (jäännösvaihtelu)



Tässä ei saisi näkyä mitään systemaattista, mutta nyt ilmenee, että mallissa on vaikeuksia. Aineiston heterogeenisuus tuo ongelmia.

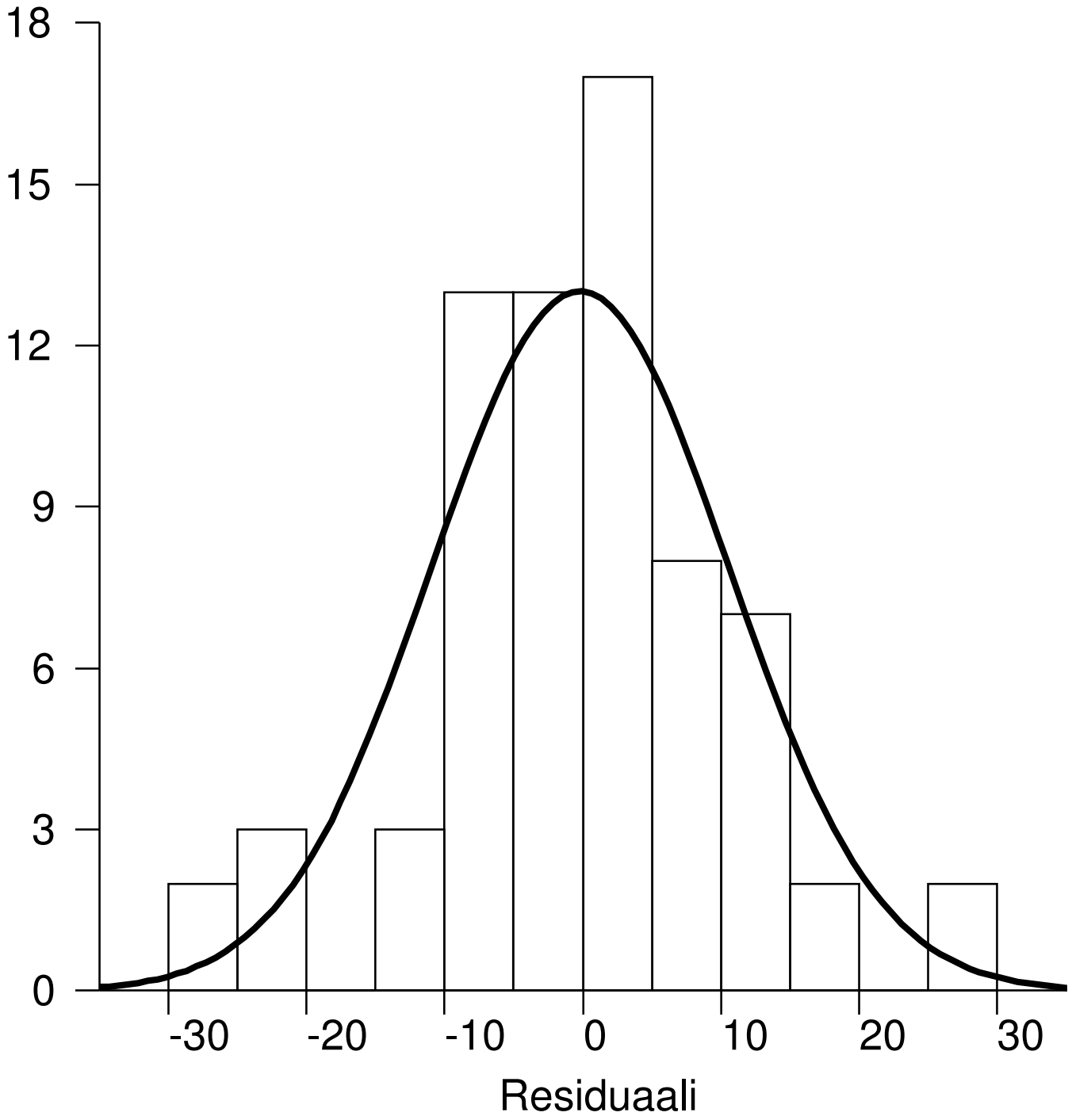
Jäännösvaihtelu ei ole satunnaista. Malli on lievästi harhainen. Näistä tarkemmin esim. MOOCilla mainituissa (ja osin tai kokonaan saatavilla olevissa) kirjoissa.

Diagnostiikkakuva 3: residuaali todennäköisyyspaperilla



Todennäköisyyspaperikuvassa pystyakselilla ovat normaalijakauman kertymäfunktion arvot. Mitä paremmin pisteet kuvautuvat suoraksi, sitä paremmin empiirinen jakauma vastaa normaalijakaumaa.

Diagnostiikkakuva 4: residuaalin histogrammi



Teema 9:stä tutussa χ^2-yhteensopivuustestissä p=0.37. Residuaalia voitaneen siis pitää normaalisti jakautuneena. Oletus mallivirheen normaalisuudesta näyttää pätevän.

Esimerkki päättyy: johtopäätökset

Näin rakennettu malli on kohtuullisen hyvä, mutta lievä harhaisuus olisi hyvä saada korjattua, joko paremmilla selittäjävalinnoilla tai mahdollisilla regressiodiagnostisilla keinoilla kuten muunnoksilla.

Malliin liittyvä normaalijakaumaoletus näyttäisi pitävän paikkansa, joten t-testien tulokset ovat sinänsä luotettavia.

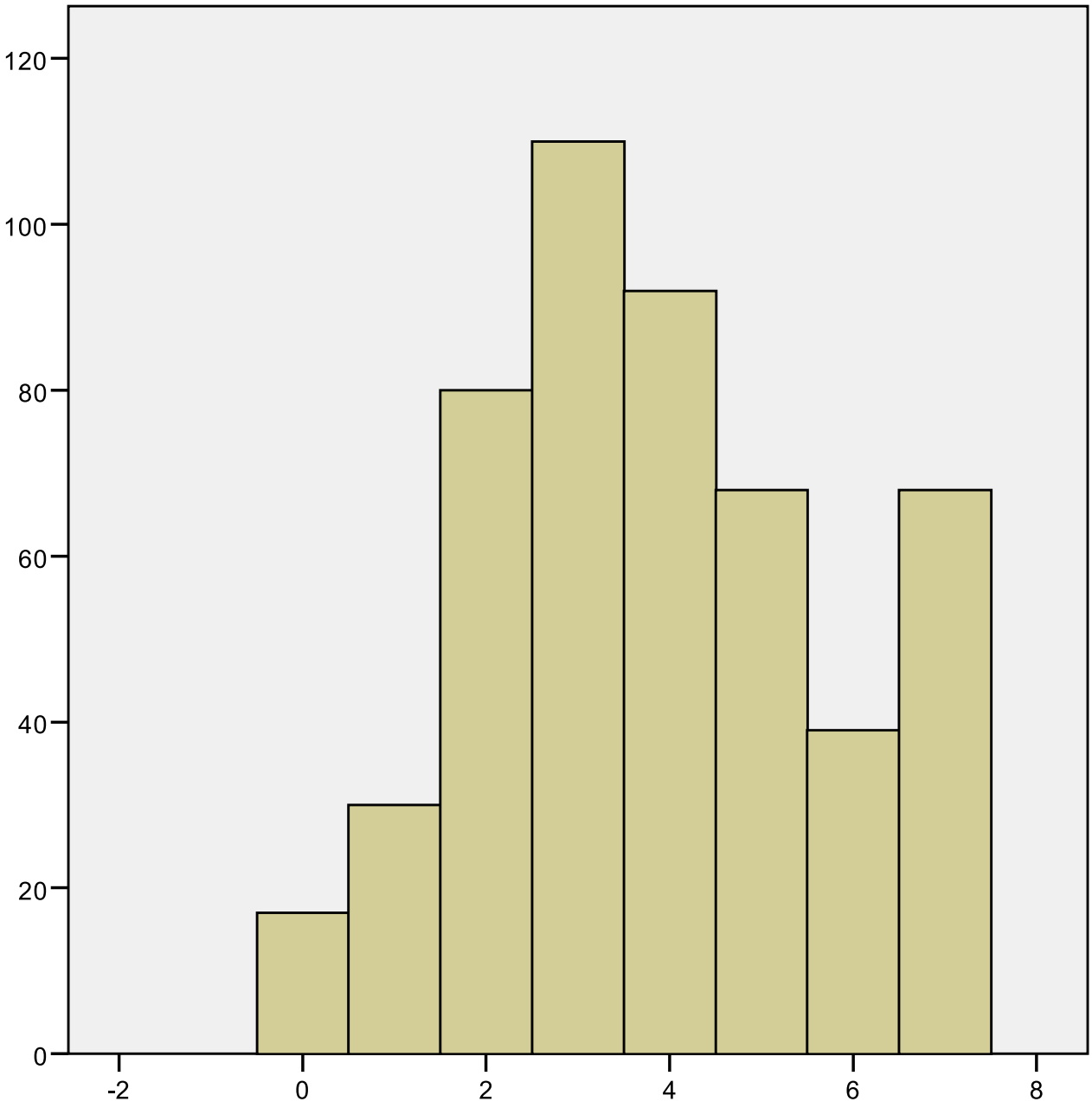
Sisällölliset tulkinnat on tässä jätetty vähemmälle huomiolle, ja saatuun malliin onkin syytä suhtautua enemmänkin teknisenä esimerkkinä regressiomallin laatimisesta.

Huomaa jälleen, miten kurssilla aiemmin opitut asiat ja käsitteet (jakaumat, parametrit, estimointi, testit, p-arvot jne.) sekä erilaiset merkintätavat tulivat tässä yhteydessä käyttöön.

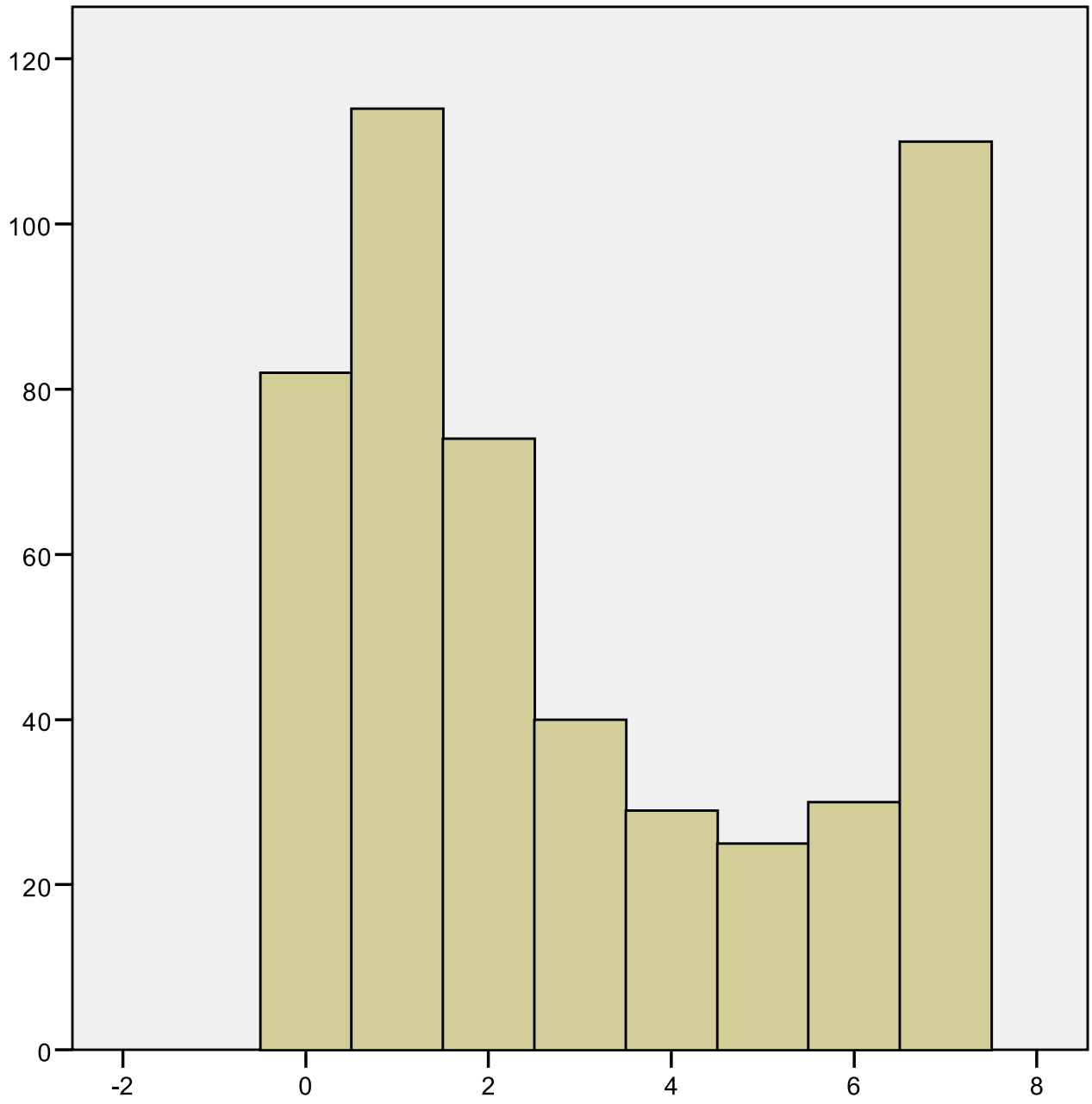
Tästä olisi hyvä jatkaa pidemmällekin, mutta sitä ei tämän kurssin puitteissa tehdä. Siirrytään sen sijaan toiseen esimerkkiin.

Esimerkki: European Social Survey, mediaseuranta

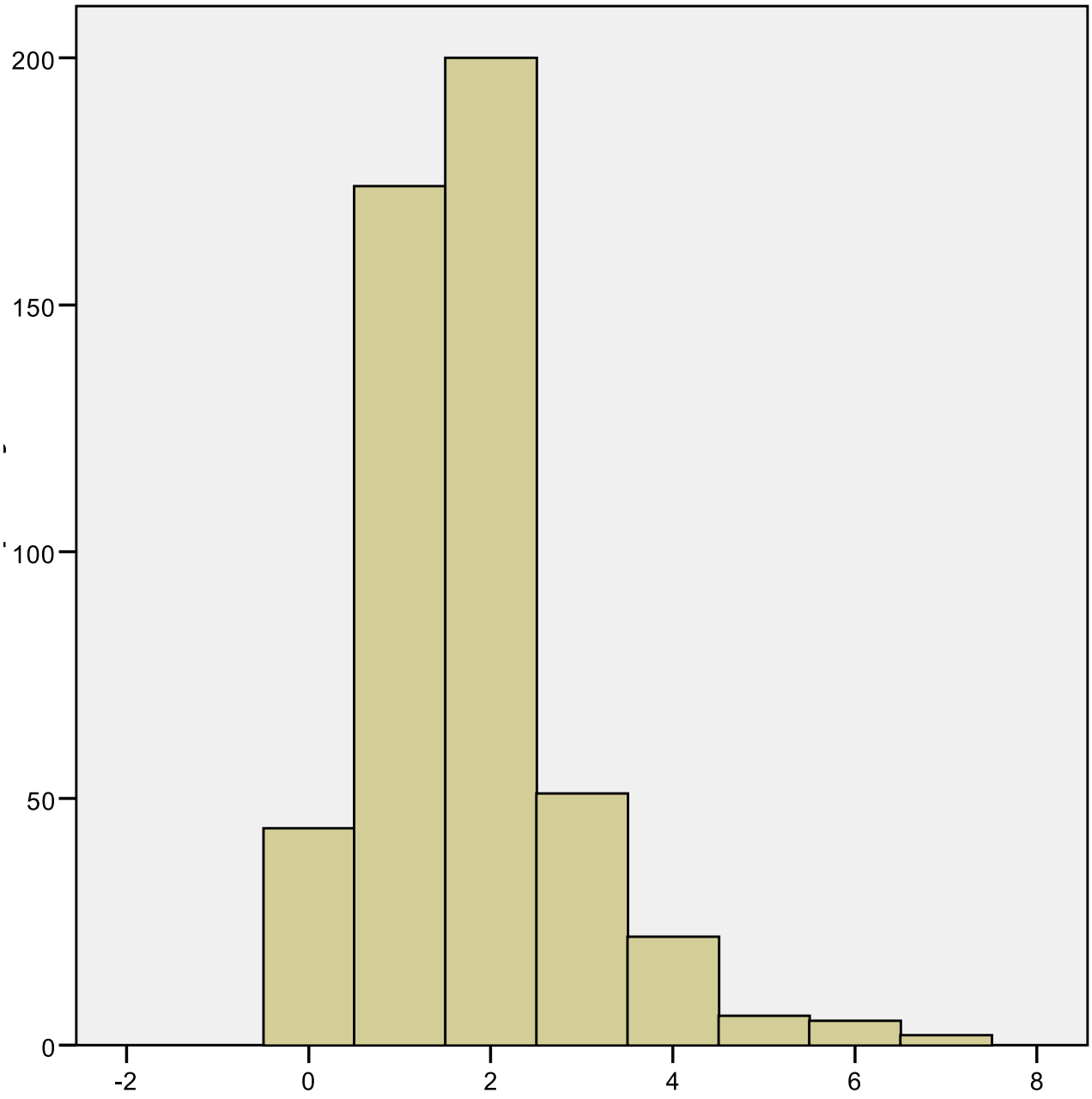
Tarkastellaan ESS-aineiston kolmea muuttujaa: TV, radio ja lehdet:



[a1] Kuinka paljon yhteensä käytätte aikaa television katsomiseen tavallisina arkipäivinä



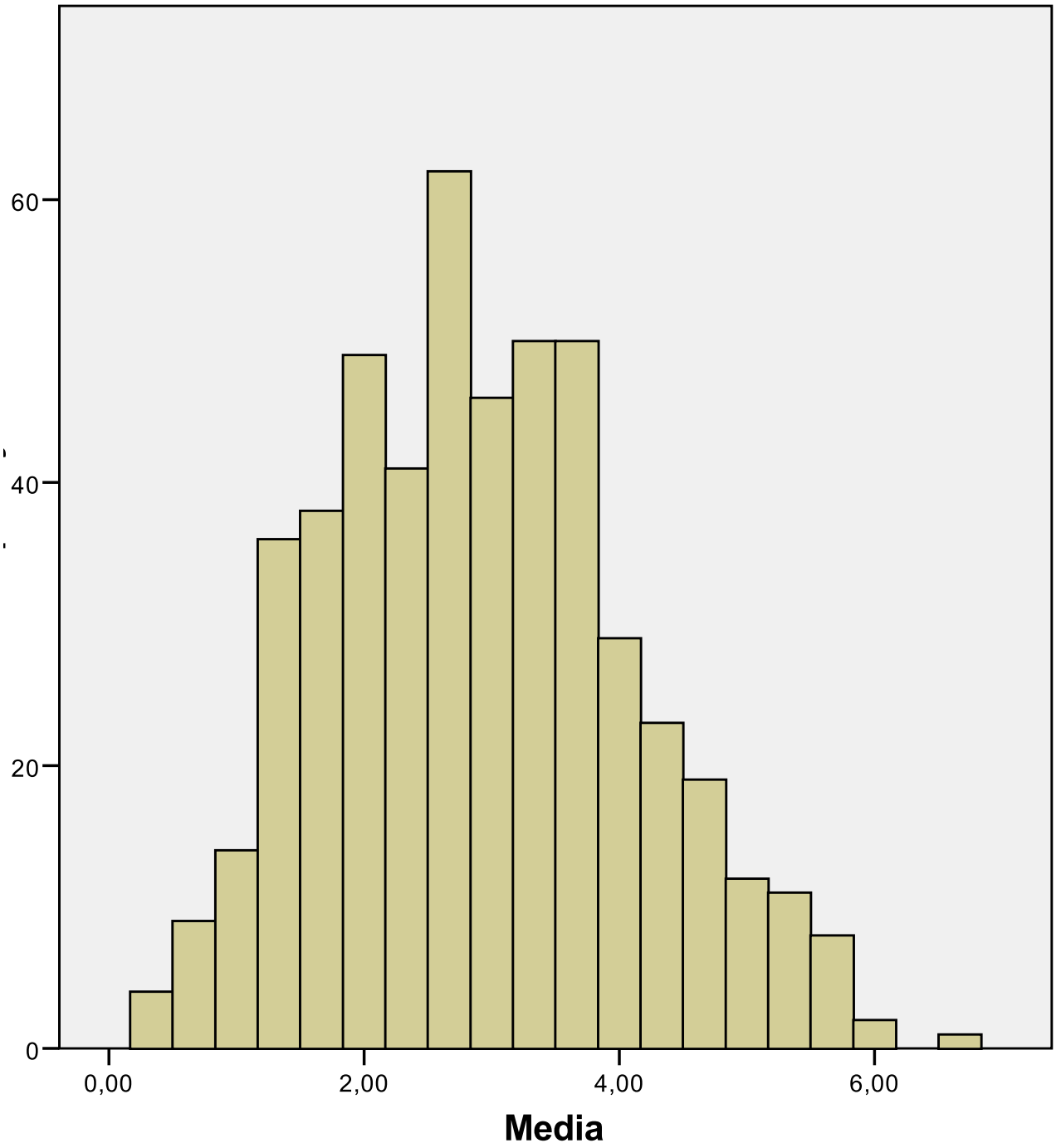
[a3] Kuinka paljon yhteensä käytätte aikaa radion kuunteluun tavallisina arkipäivinä



Muodostetaan summamuuttuja Media(seurannan aktiivisuus):

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
[a1] Kuinka paljon yhteensä käytätte aikaa television katsomiseen tavallisina arkipäivinä	504	0	7	3.85	1.890
[a3] Kuinka paljon yhteensä käytätte aikaa radion kuunteluun tavallisina arkipäivinä	504	0	7	3.12	2.607
[a5] Kuinka paljon yhteensä käytätte aikaa sanomalehtien lukemiseen tavallisina arkipäivinä	504	0	7	1.76	1.132
Media	504	.33	6.67	2.9101	1.19450
Valid N (listwise)	504				



Esimerkki jatkuu: ESS ja mediaseuranta

Laaditaan regressiomalli, jonka avulla selitetään mediaseurannan aktiivisuutta vastaajan sukupuolella ja iällä. Ikää (vuosina) käytetään mallissa jatkuvana muuttujana, ja sukupuoli on koodattu 1=mies, 2=nainen. Tulostus alkaa yhteenvedolla:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.327[a]	.107	.103	1.13127

a. Predictors: (Constant), Sukupuoli (1=mies, 2=nainen), Ikä (vuosina)

b. Dependent Variable: Media

Selitysaste ei ole kovin kaksinen (n. 11 %). Samaa luokkaa on myös korjattu selitysaste, joka laajemmissa malleissa rankaisee liioista selittäjistä ja on siksi hieman parempi mallin hyvyyden mitta.

Selitysasteen R^2 neliöjuurta merkitään usein (myös edellä olleissa tulosteissa) symbolilla R ja kutsutaan *yhteiskorrelaatiokertoimeksi*. Nimi tulee siitä, että kyseessä on y:n ja $\hat{y}$:n korrelaatiokerroin. Jos selittäjiä on vain yksi,
R on sama kuin y:n ja x:n tavallinen korrelaatiokerroin.

Esimerkki jatkuu: ESS ja mediaseuranta

Seuraava tuloslaatikko on otsikoitu lyhenteellä ANOVA, joka tulee **varianssianalyysia** tarkoittavista sanoista ANalysis Of VAriance.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	76.538	2	38.269	29.903	.000[a]
	Residual	641.162	501	1.280		
	Total	717.700	503			

Varianssianalyysi liittyy läheisesti regressioanalyysiin (nämä menetelmät ovat oikeastaan saman *lineaarisen mallin* ilmentymiä). Tässä kohtaa on kysymys siitä, miten regressioanalyysi on yleisesti onnistunut: paljonko y-muuttujan (Media) vaihtelusta selittyy mallin avulla ja paljonko jää selittämättä.

Nämä tiedot esitetään perinteisesti tässä näkyvänä *varianssitauluna*, joka koostuu kumpaakin osuutta kuvaavista neliösummista, vapausasteista ja näiden osamäärinä saatavista varianssitermeistä (Mean Square). Varianssien suhteena saadaan F-testisuure, joka testaa *(todella skeptistä!)* hypoteesia
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$
siis että kaikkien varsinaisten selittäjien regressiokertoimet olisivat nollia.

Testin p-arvon perusteella *ainakin yksi* kertoimista poikkeaa nollasta, ts. mallin selittäjävalinta ei ole aivan pielessä (vaikka selitysaste onkin melko vähäinen).

Esimerkki jatkuu: ESS ja mediaseuranta

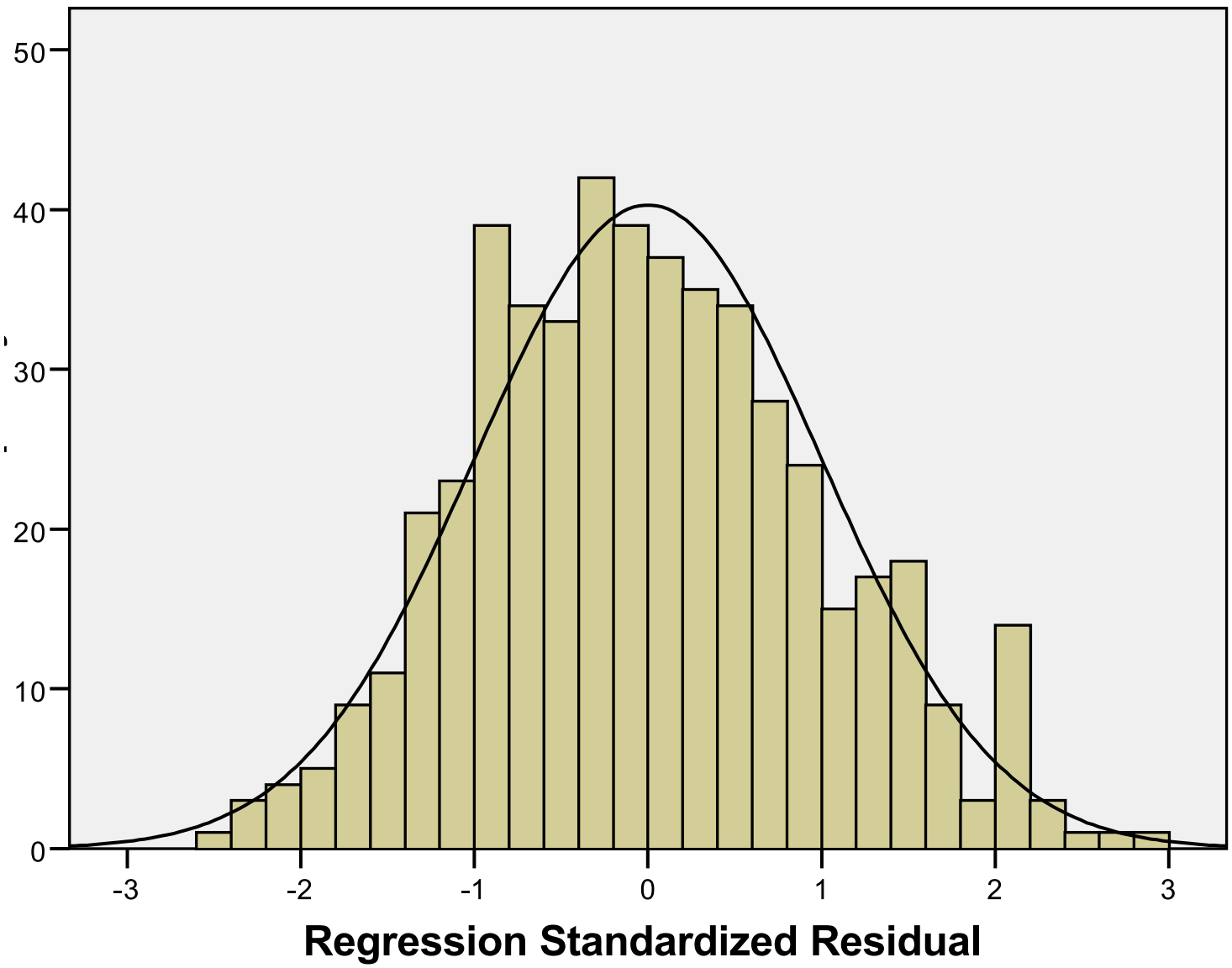
Kiinnostavin osa tulostuksesta on yleensä seuraava laatikko:

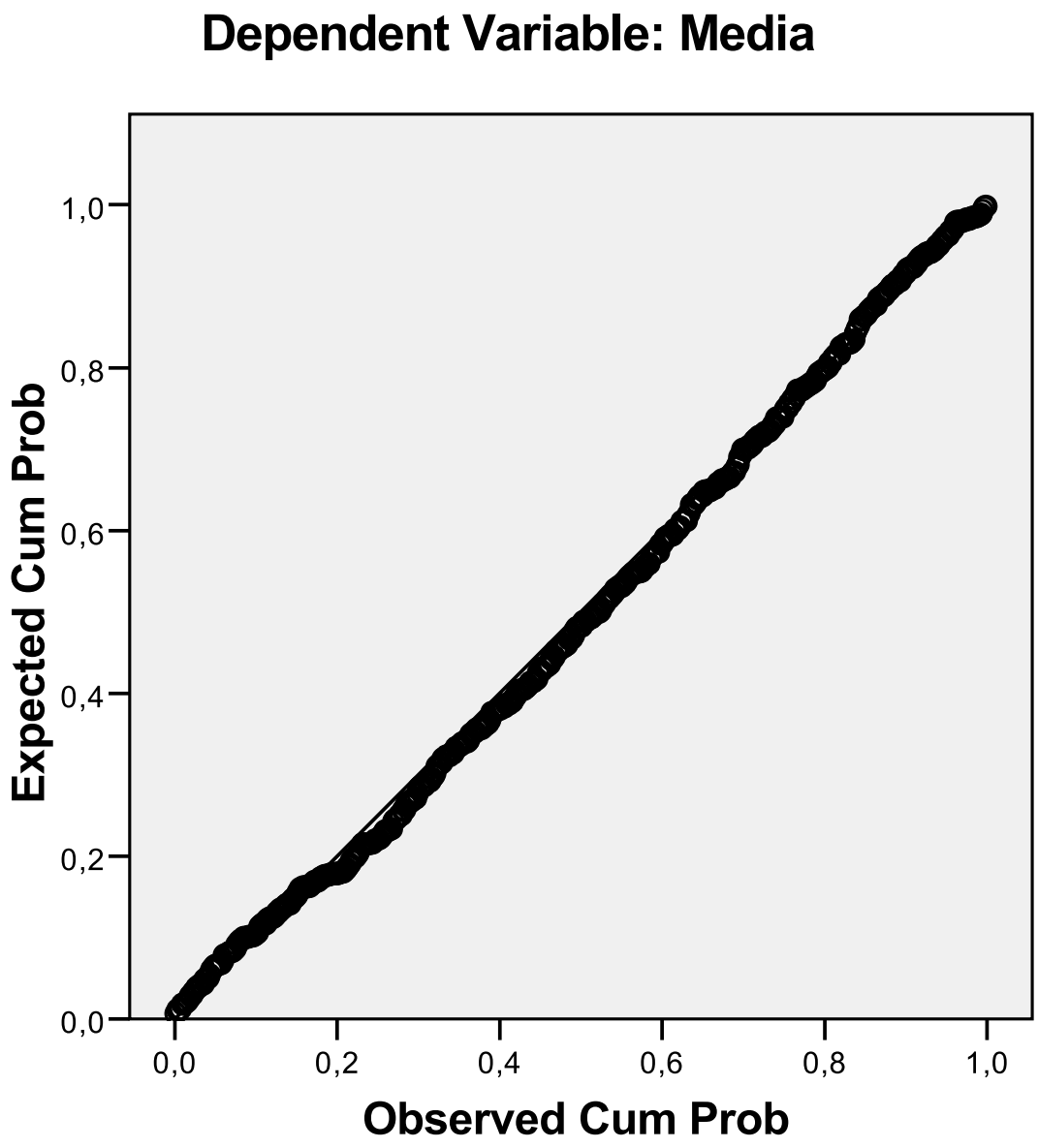
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.270	.203		11.159	.000
	Ikä (vuosina)	.020	.003	.317	7.518	.000
	Sukupuoli (1=mies, 2=nainen)	-.190	.101	-.079	-1.878	.061

a. Dependent Variable: Media

Analyysia täydentävät kaksi diagnostista kuvaa (vrt. edellä):


Dependent Variable: Media
50
40
30
20
10
0
-3
-2
-1
0
1
2
3
Regression Standardized Residual



Tulkintoihin palataan tarkemmin harjoituksissa.

Esimerkki jatkuu: ESS ja mediaseuranta

Äskeisen kaltaisia tilanteita on yhteiskuntatieteissä tyypillistä tutkia myös varianssianalyysilla. Luokitellaan seuraavassa ikä (karkeasti) ryhmiin 15–34 -vuotiaat, 35–54 -vuotiaat ja yli 54-vuotiaat.

Jos ikäryhmiä olisi vain kaksi, niitä voitaisiin vertailla t-testillä.
Sen yleistys useamman ryhmän vertailuun tunnetaan nimellä **yksisuuntainen varianssianalyysi**.
Tutkittavana muuttujana on edelleen mediaseurannan aktiivisuus.

Aluksi saadaan ryhmiä kuvaavia tunnuslukuja ja luottamusvälejä:

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
15-34	158	2.5865	1.07499	.08552	2.4176	2.7554
35-54	176	2.6913	1.13865	.08583	2.5219	2.8607
55-	170	3.4373	1.18463	.09086	3.2579	3.6166
Total	504	2.9101	1.19450	.05321	2.8055	3.0146

Esimerkki jatkuu: ESS ja mediaseuranta

Seuraavaksi tulee varianssitaulu, joka muistuttaa hyvin paljon regressioanalyysin yhteydessä nähtyä vastaavaa esitystä. Tässä kysymys on ryhmien välisen ja ryhmien sisäisen vaihtelun suhteesta. Kiinnostuksen kohteena on nimenomaan ryhmien välinen vaihtelu, ts. miten suuria eroja ikäryhmien välillä on havaittavissa.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	72.214	2	36.107	28.025	.000
Within Groups	645.486	501	1.288		
Total	717.700	503			

F-testin perusteella eroja on selvästi havaittavissa. Tarkemmin erot paljastuvat vasta **monivertailutesteillä**:

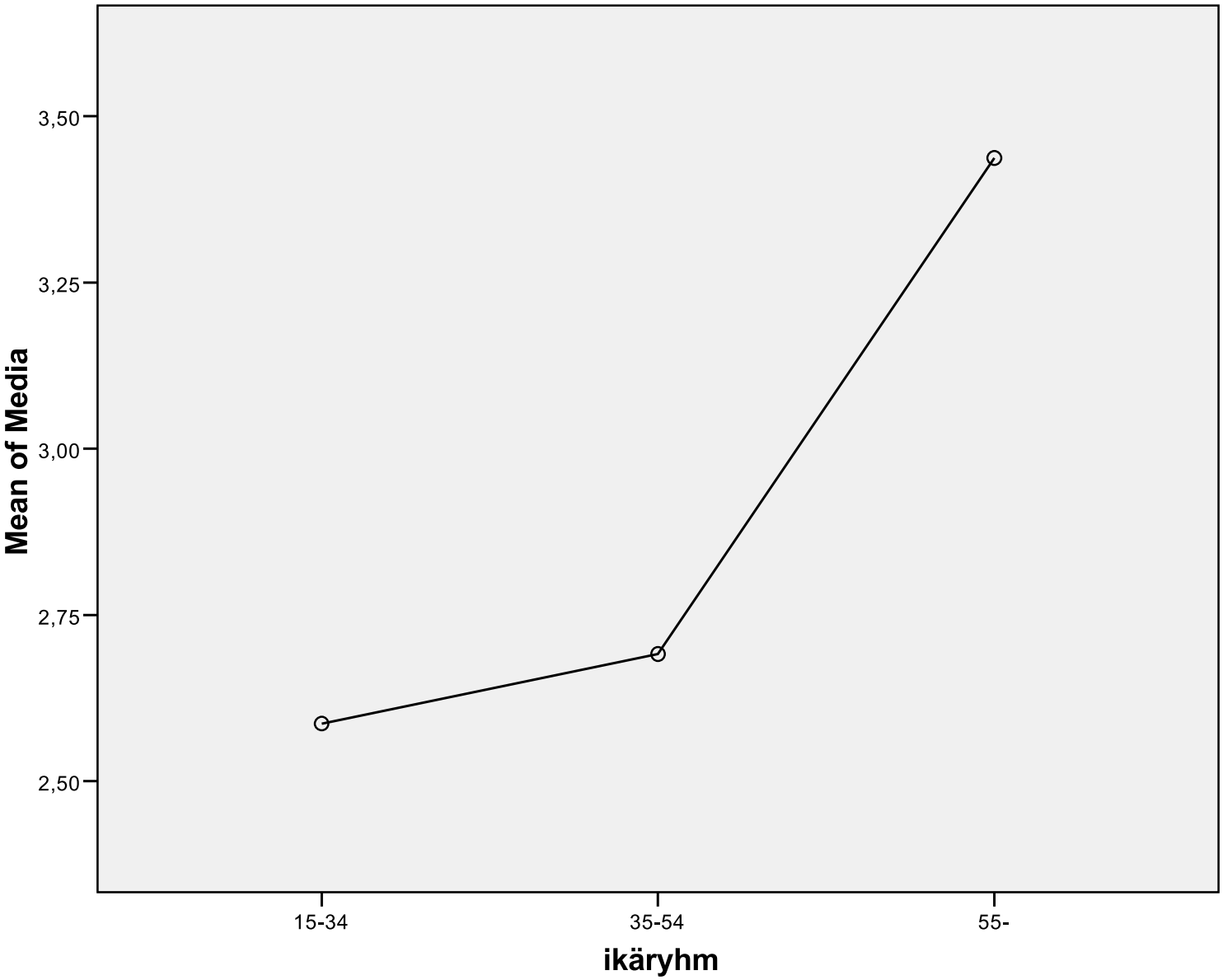
(I) ikäryhm	(J) ikäryhm	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
15-34	35-54	-.10479	.12440	.677	-.3972	.1876
	55-	-.85076*	.12543	.000	-1.1456	-.5559
35-54	15-34	.10479	.12440	.677	-.1876	.3972
	55-	-.74597*	.12206	.000	-1.0329	-.4590
55-	15-34	.85076*	.12543	.000	.5559	1.1456
	35-54	.74597*	.12206	.000	.4590	1.0329

*. The mean difference is significant at the .05 level.

Esimerkki (ja kurssi!) päättyy: ESS ja mediaseuranta

Varianssianalyysin (joskus hieman harhaanjohtavalta tuntuva) nimi tulee siis y-muuttujan **varianssin** hajottamisesta erilaisiin osiin, mutta mielenkiinto keskittyy **ryhmien keskiarvojen vertailuun**.

Tuloksia on usein tapana visualisoida **keskiarvoprofiileilla**:



Varianssianalyysillakin voidaan tutkia useita muuttujia sekä niiden **yhdysvaikutuksia**, mutta niihin perehdytään muilla kursseilla.