Tilastollisen päättelyn peruskurssi

Henri Pesonen

Sisältö

Luku 1

Johdanto tilastotieteeseen

Tilastotiede ei ole asioiden tilastointia, kuten usein virheellisesti luullaan. Tilastojen laatimisessa ja varsinkin tilastojen analysoinnissa kuitenkin käytetään hyödyksi tilastotieteellisiä menetelmiä. Tilastotiede onkin menetelmätiede, jossa kehitetään ja tutkitaan menetelmiä, joita käytetään hyväksi kaikkien havaintoja hyödyntävien soveltavien tieteiden parissa. Tilastotiedettä käytetään hyväksi esimerkiksi lääketieteessä, luonnontieteissä, yhteiskuntatieteissä, insinööritieteissä, taloustieteessä ja käyttäytymistieteissä. Usein tilastotiede on niin olennainen osa soveltavan tieteen tutkimusta, että siihen liittyviä tilastollisia menetelmiä kutsutaan omalla nimellään, kuten biologisten alojen biostatistiikka, taloustieteiden ekonometria, ja psykologian psykometria. Tilastotieteen avulla ei ainoastaan analysoida ja tulkita aineistoja, vaan se on myös keskeisessä osassa tutkimusten suunnittelussa ja havaintoaineistojen keräämisessa. Lisäksi havaintoaineistojen pohjalta tehdyt päätökset voidaan tehdä tilastotieteeseen pohjautuen.

1.1 Tilastotieteen peruskäsitteitä

Tilastotieteellinen tutkimus on luonteeltaan teoreettista tai empiiristä tutkimusta. Teoreettisen tutkimuksen avulla tarkastellaan tutkimuksen kohdetta ajatusrakennelmien ja mallien avulla, ja rakennetaan malleja empiiriselle tutkimukselle. Empiirisessä, eli kokemusperäisessä tutkimuksessa, tarkastelut perustetaan tutkimuskohteiden havainnointiin ja mittaamiseen. Tilastotieteellisessä tutkimuksessa tarkasteltavaa tutkimuskohteiden, eli **tilastoyksiköiden** joukkoa kutsutaan **populaatioksi** (perusjoukoksi).

Määritelmä 1 (Populaatio)**.** Konkreettinen tai hypoteettinen tutkimuskohteiden joukko, joka koostuu tilastoyksiköistä.

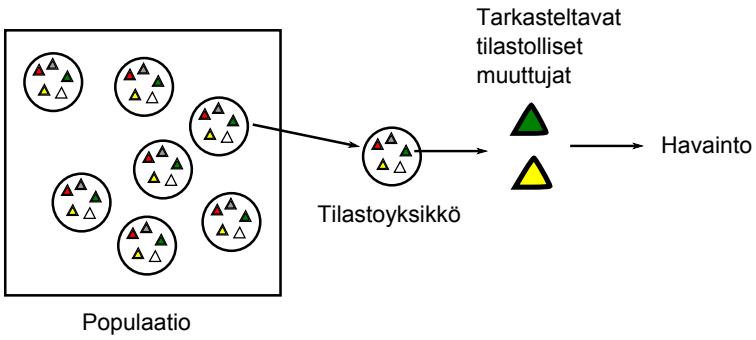
Populaatio voi siis koostua esimerkiksi tietyssä tehtaassa valmistetusta 1000 energiansäästölampusta. Tämä on konkreettinen populaatio, eli jokaisen yksilö voidaan tarvittaessa listata. Entä jos ajatellaankin populaationa kaikkia samanlaisia energiansäästölamppuja, jotka ovat valmistettu ja tullaan valmistamaan tulevaisuudessa samassa tehtaassa? Tämä populaatio on hypoteettinen joukko, jonka avulla voidaan ennustaa esimerkiksi tulevaisuudessa tehtaasta valmistuvien lamppujen eliniän tilastollisia ominaisuuksia.

Määritelmä 2 (Tilastoyksikkö ja tilastollinen muuttuja)**.** Populaation muodostavilta tilastoyksiköiltä tarkastellaan tilastollisia muuttujia, joita voidaan mitata tai havaita.

Tilastoyksiköiltä tarkastellaan niiden ominaisuuksia, eli tilastollisia muuttujia. Tilastoyksikköiden tilastollisilla muuttujilla on tietty mahdollisten arvojen joukko, ja näillä arvoilla on jokin **jakauma** populaatiossa. Kun tarkasteltavien tilastoyksikön tilastollisen muuttujien arvot havaitaan, kutsutaan näiden arvojen joukkoa **havainnoksi**. Eri tasojen suhdetta on selvitetty Kuvassa 1.1.

Määritelmä 3 (Havainto)**.** Havainto muodostuu tilastoyksikön tarkasteltavien tilastollisen muuttujien havaituista arvoista.

Kerättyjen havaintojen joukko muodostaa havaintoaineiston, eli datan, jonka pohjalta aloitetaan tilastollinen päättely populaatiosta.



Kuva 1.1: Populaatio koostuu tilastoyksiköistä, joilla on tilastollisia muuttujia. Tarkasteltavista tilastollisista muuttujista kerätään havaintoja, joiden pohjalta tutkitaan populaation ominaisuuksia.

Määritelmä 4 (Havaintoaineisto)**.** Havaintoaineisto on tilastoyksiköiden tilastollisista muuttujista kerätty havaintojen joukko.

Kaikki havaitut muuttujat eivät ole aina mielenkiintoisia. Mielenkiintoisia muuttujia tilastollisen tutkimuksen kannalta voidaan kutsua **tutkimusmuuttujiksi**, ja muita muuttujia **taustamuuttujiksi**.

Esimerkki 1. Tilastollisessa tutkimuksessa halutaan selvittää suomalaisten 30-40 vuotiaiden miesten keskimääräisen painoindeksin, joka muodostuu henkilön pituudesta ja painosta. Populaation muodostaa tässä tutkimuksessa kaikki suomalaiset 30-40 vuotiaat miehet. Tilastoyksikkö on yksi suomalainen 30-40 vuotias mies, jolla on tilastolliset muuttujat pituus ja paino. Henkilön punnittu paino kilogrammoissa sekä mitattu pituus metreissä muodostavat tilastoyksikön tilastollisten muuttujien havainnot. Kun kerätään yhteen kaikkien punnittujen sekä mitattujen henkilöiden pituudet ja painot yhteen, niin kyseinen kokonaisuus muodostaa havaintoaineiston.

Esimerkki 2. Tarkastellaan kurssin TILM3510 erääseen kymmenen henkilön harjoitusryhmään osallistuvan opiskelijan populaatiota, ja mielenkiintoisten tilastollisten muuttujien havaintoja. Populaatio on

Opiskelija	Sukupuoli	Pituus (cm)	Paino (kg)	TILM3509 suoritettu (k/e)
1	Mies	178	83	k
2	Mies	173	84	k
3	Mies	178	76	k
4	Nainen	168	60	e
5	Nainen	165	59	k
6	Mies	188	83	k
7	Nainen	158	62	k
8	Nainen	159	62	e
9	Nainen	170	60	k
10	Nainen	165	56	k

Tilastoyksikkö on tarkasteltavaan harjoitusryhmään osallistuva opiskelija, jonka tilastolliset muuttujat ovat Sukupuoli, Pituus, Paino sekä TILM3509:n suoritus.

Usein tilastollinen tutkimus on luonteeltaan selittävää tutkimusta, jossa pyritään havaintoaineiston avulla lisäämään tietämystä populaatiosta, tai jostain ilmiöstä populaatiossa, tai selittämään yleisiä tai syy-seuraus yhteyksiä tilastollisten muuttujien välillä.

Tilastollisia muuttujia, joiden esiintyminen ja arvot havaintoaineistossa tiedetään ennen havaintojen mittaamista, kutsutaan **tekijöiksi**. Vastaavasti tilastolliset muuttujat, joiden arvot havaitaan osana havaintoaineiston keräämistä, ovat **vasteita**. Selittävissä tutkimuksissa tutkitaan ns. selittävien muuttujien vaikutusta selitettäviin muuttujiin. Jos tarkastellaan selittävän tekijän vaikutusta selitettävään vasteeseen, kyseessä on tilastollisen yhteyden tarkastelu. Jos taas tarkastellaan selittävän vasteen vaikutusta selitettävään vasteeseen, tarkastellaan vasteiden välistä tilastollista riippuvuutta.

Määritelmä 5 (Tekijä)**.** Tekijä on tilastollinen muuttuja, jonka esiintyminen ja arvo havaintoaineistossa tiedetään ennen havaintojen mittaamista.

Määritelmä 6 (Vaste)**.** Vaste on tilastollinen muuttuja, jonka arvo havaitaan osana havaintoaineiston keräämistä.

Määritelmä 7 (Selitettävä muuttuja)**.** Selitettävä muuttuja on vaste, jonka jakauma on mielenkiinnon kohteena tilastollisessa tutkimuksessa.

Määritelmä 8 (Selittävä muuttuja)**.** Selittävä muuttuja on vaste tai tekijä, jonka avulla yritetään selittää selitettävän muuttujan vaihtelua.

Tärkeässä osassa tilastollisia tutkimuksia on satunnaistamisen käyttäminen osana tutkimusta. Satunnaistamisen ansiosta voidaan käyttää todennäköisyyslaskennan avulla johdettuja tilastollisia menetelmiä aineistojen analysoinnissa sekä laajentaa aineistojen pohjalta tehdyt päätelmät koko populaatiota koskeviksi.

1.2 Tilastolliset muuttujat

Olennaisena osana tilastollisten mallien rakentamista on jaotella tilastolliset muuttujat oikein jotta voidaan valita oikeat työkalut mallien käsittelyä varten. Usein työkalut ollaan kehitetty erilaisilla mitta-asteikoilla havaittaville muuttujille. Tilastolliset muuttujat voivat olla luokallisia (kategorisia) tai numeerisia (kvantitatiivisia) muuttujia, sen mukaisesti minkälaisilla mittaustasoilla niiden arvoja havaitaan. Voidaan myös puhua kvalitatiivisesta tai kvantitatiivisesta tutkimuksesta, sen mukaan tutkitaanko *minkä laatuista?*- tai *kuinka paljon?*-tyyppisiä kysymyksiä. Jos tilastollista muuttujaa ei voi mitata diskreetillä tai jatkuvalla numeerisella asteikolla, kyseinen muuttuja on luokallinen.

Määritelmä 9 (Luokallinen muuttuja)**.** Jos tilastollisen muuttujan havaittu arvo ei ole numeerinen luku, kyseessä on luokallinen muuttuja.

Määritelmä 10 (Numeerinen muuttuja)**.** Jos tilastollinen muuttujan havaittu arvo on jokin lukuarvo, kyseessä on numeerinen muuttuja.

Luokallisia muuttujia ovat mm. henkilön sukupuoli, auton kokoluokka (pikkuauto, perheauto, tila-auto). Luokalliset muuttujat voidaan jaotella edelleen kahteen luokkaan sen mukaisesti voidaanko ne laittaa luonnollisesti johonkin järjestykseen mitta-asteikkonsa perusteella. Jos muuttujia ei voida laittaa luonnolliseen järjestykseen, kuten henkilön sukupuolen mukaan ei voida laittaa henkilöitä järjestykseen, niin luokallisen muuttujan mitta-asteikko on nominaaliasteikko. Jos taas muuttujilla on luonnollinen järjestys, kuten auton kokoluokat voidaan järjestellä pienimmästä suurimpaan kokoluokkaan, niin luokallisen muuttujan mitta-asteikko on ordinaaliasteikollinen (järjestysasteikollinen).

Määritelmä 11 (Nominaaliasteikko)**.** Nominaaliasteikoilla mitattavat muuttujat voidaan jakaa ominaisuuksiensa perusteella ryhmiin, sen perusteella onko muuttujilla ominaisuutta vai ei.

Määritelmä 12 (Ordinaaliasteikko)**.** Ordinaaliasteikolla mitattavat muuttujat voidaan järjestellä joidenkin ominaisuuksiensa perusteella ryhmiin, joilla on jokin luonnollinen järjestys.

Numeeriset muuttujat, eli diskreettejä tai jatkuvia arvoja saavat muuttujat, voidaan myös jaotella aliluokkiin. Jos numeerisella muuttujan mitta-asteikolla on absoluuttinen nollakohta, niin numeerisen muuttujan mitta-asteikko on suhdeasteikko. Objektin paino kilogrammoissa on suhdeasteikolla mitattava numeerinen muuttuja, sillä objekteilla ei ole negatiivisia painoja. Suhdeasteikoksi tätä mitta-asteikkoa kutsutaan, koska voidaan esimerkiksi sanoa että 2 kg painava objekti on 2 kertaa painavampi kuin 1 kg painava objekti. Jos taas numeerisen muuttujan nollakohta on sopimuksenvarainen ja muuttuja voi saada negatiivisia arvoja, niin numeerisen muuttujan mitta-asteikko on intervalliasteikko. Lämpötila Celsius-asteikolla mitattuna on intervalli-asteikollinen muuttuja. Intervalliasteikolla voidaan verrata mitta-asteikon välien pituuksia toisiinsa, eli voidaan sanoa, että päivän minimi- ja maksimilämpötiloja kuvatessa vaihteluväli $[-25°\mathrm{C}, -5°\mathrm{C}]$ on kaksi kertaa pidempi kuin väli $[5°\mathrm{C}, 15°\mathrm{C}]$. Toisaalta Celsius-asteikolla ei voida kuvailla kuinka monta kertaa lämpimämpää $10°\mathrm{C}$ on kuin $-5°\mathrm{C}$.

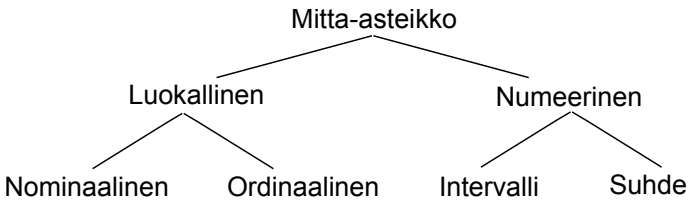
Määritelmä 13 (Intervalliasteikko)**.** Intervalliasteikolla mitattavien muuttujien havainnoista voidaan laskea erotus.

Määritelmä 14 (Suhdeasteikko)**.** Suhdeasteikolla mitattavilla muuttujilla on yksikäsitteinen nollapiste.

Esimerkki 3. Jonkin mielipiteen yhtymiseen liittyvää voimakkuutta mittaava Likert-asteikko on klassinen esimerkki ordinaaliasteikosta, vaikka joskus yksittäisiä Likert-asteikkoja on pidetty jossain tutkimuksissa intervalliasteikkona. Usein mielipide esitetään väitteenä kyselytutkimuksessa ja vastaaja valitsee esimerkiksi viiden tason Likert-asteikolta vaihtoehdon

1. vahvasti eri mieltä
2. eri mieltä
3. en osaa sanoa
4. samaa mieltä
5. vahvasti samaa mieltä

Vaikka ei olisi mahdollista puhua kvantitatiivisesti mielipiteen voimakkuudesta, voidaan kuitenkin aina järjestää vaihtoehdot mitta-asteikolla luonnolliseen järjestykseen.



Kuva 1.2: Tilastollisten muuttujien mitta-asteikkojen tyypit.

Havaintoaineiston keräämisen jälkeen täytyy aineisto koota muotoon, jossa sen käsittely on helppoa esimerkiksi tilastollisissa ohjelmistoissa kuten R, SPSS ja SAS. Kerätään aineisto taulukoksi, johon listaamme tilastoyksiköt riveille, ja tilastolliset muuttujat sarakkeisiin. Lopputulosta kutsutaan havaintomatriisiksi.

Määritelmä 15 (Havaintomatriisi)**.** Havaintoaineisto koostuu n tilastoyksiköstä, joista jokaisesta on kerätty m tilastollisesta muuttujasta havainnot. Havaintomatriisi on taulukon muotoon kirjoitettuna

	til. muuttuja 1	til. muuttuja 2	$\cdots$	til. muuttuja m
tilastoyksikkö 1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,m}$
tilastoyksikkö 2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,m}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
tilastoyksikkö n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,m}$,

ja matriisimuotoon kirjoitettuna

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix},$$

jossa rivillä i on i. tilastoyksikön havainto, ja j sarakkeessa on j. tilastollisesta muuttujasta havaitut arvot $x_{i,j}$. Jos tarkastellaan yhtä havaintoa i. tilastollisesta muuttujasta, niin se voidaan kirjoittaa matriisimuodossa

$$x_{i,:} = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}].$$

Jos tarkastellaan kaikkien tilastoyksiköiden havaittuja arvoja j. tilastollisesta muuttujasta, niin voimme kirjoittaa sen muodossa

$$x_{:,j} = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix},$$

tai tilaasäästäen muodossa

$$x_{:,j} = \begin{bmatrix} x_{1,j} & x_{2,j} & \cdots & x_{n,j} \end{bmatrix}^T.$$

Jos tarkastellaan ainoastaan yhtä tilastollista muuttujaa, niin voidaan toinen alaindeksi unohtaa merkinnöistä kokonaan ja merkitä havaintoja

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T.$$

Jos havainnot kerätään populaatiosta satunnaisotoksen avulla, niin voidaan havaintoja ajatella satunnaismuuttujina ennen niiden arvojen kiinnittämistä. Tällöin tilastollisten muuttujien voidaan ajatella jakautuneen populaatioon jollakin todennäköisyysjakaumalla. Havaintomatriisi, havainto sekä tilastollisen muuttuja voidaan kirjoittaa satunnaismuuttujina esimerkiksi merkitsemällä niitä isolla kirjaimella $X, X_{:,j}$ ja $X_{i,j}$. Jos $X_{i,j}$:llä on todennäköisyysjakauma $F_{X_{\cdot,j}}$, kirjoitamme $X_{i,j} \sim F_{X_{\cdot,j}}$.

Luku 2

Kuvaileva tilastotiede

Empiirisessä tilastollisessa tutkimuksessa havaintoaineiston keräämisen jälkeen aloitetaan aineiston analyysi. Analyysi voidaan jakaa kahteen osaan, eli kuvailevaan tilastotieteeseen sekä tilastolliseen päättelyyn. Tutustutaan ensin kuvailevan, eli deskriptiivisen tilastotieteen menetelmiin. Kuvailevan tilastotieteen avulla pyritään esittämään eri tavoin havaintoaineiston pääpiirteet tiivistetyssä muodossa. Tämä tarkoittaa siis että kuvailevan tilastotieteen menetelmillä ei pyritä tekemään yleistäviä päätelmiä aineiston ulkopuolelle, vaan pyritään tiivistämään kerätty havaintoaineisto informatiivisempaan muotoon.

2.1 Graafiset menetelmät

- Vertailu (comparison)
- Rakenne (Composition)
- Jakauma (Distribution)
- Asioiden suhteet (Relationship)

2.2 Frekvenssijakauma

Kerätyt havaintoaineistot ovat usein niin isoja, että havaintomatriisista ei suoraan tarkastelemalla käy ilmi aineiston pääpiirteet. Yksinkertainen tapa tiivistää aineistoa on muodostaa frekvenssijakauma aineistosta. Frekvenssijakauman koostamiseksi selvitetään tarkasteltavan tilastollisen muuttujan mahdolliset arvot. Frekvenssijakaumaa voidaan käyttää joko luokitellun muuttujan, tai diskreetin numeerisen muuttujan kuvaamiseen. Oletetaan, että mahdolliset arvot voidaan jakaa äärelliseen määrään (k kpl) luokkia, joita merkitään $E_1, \ldots, E_k$. Etsitään aineistosta niiden havaintojen määrät, eli **frekvenssit**, jotka kuuluvat näihin luokkiin ja merkitään näitä $f_1, \ldots, f_k$. Tilastollisen muuttujan frekvenssijakauma on $(E_i, f_i), i = 1, \ldots, k$. Frekvenssejä $f_1, \ldots, f_k$, jotka ilmaisevat havaintojen määriä, kutsutaan myös absoluuttisiksi frekvensseiksi. Absoluuttisten frekvenssien sijasta olemme usein kiinnostuneita eri luokkiin kuuluvien muuttujien suhteellisesta osuudesta. Tällöin muodostetaan suhteellisen frekvenssijakauman $(E_i, \frac{f_i}{n}), i = 1, \ldots, k$, jossa n on havaintoaineiston koko. Suhteellisten frekvenssijakaumien vertailu toisiinsa on helpompaa, koska periaatteessa ne eivät riipu otoskoosta n. Suhteellista frekvenssijakaumaa vastaa prosentuaalinen frekvenssijakauma $(E_i, \frac{f_i}{n} \cdot 100\%), i = 1, \ldots, k$.

Esimerkki 4. Helsinki-Vantaan lentokentällä valittiin satunnaisiin turvatarkastuksiin 10 matkustajaa. Näistä matkustajista kirjattiin ylös heidän kansalaisuutensa, ja saatiin havaintoaineisto

$$\begin{bmatrix} \text{Suomi} & \text{Suomi} & \text{Ruotsi} & \text{Saksa} & \text{Suomi} & \text{Tanska} & \text{Saksa} & \text{Italia} & \text{Venäjä} & \text{Venäjä} \end{bmatrix}^T$$

Havaintojen frekvenssijakauma on

E_i	f_i
Suomi	3
Ruotsi	1
Saksa	2
Tanska	1
Italia	1
Venäjä	2

Esimerkki 5. Tarkastellaan tilannetta, jossa ollaan kysytty 30 pankin asiakkaan tyytyväisyyttä asiakaspalveluun. Mahdolliset vastausvaihtoehdot kyselyssä ovat (Tyytymätön = 1, Melko tyytymätön = 2, En osaa sanoa = 3, Melko tyytyväinen = 4, Hyvin tyytyväinen = 5), jossa vaihtoehdot ovat koodattu valmiiksi numeroiksi $1, 2, 3, 4, 5$.

Havainnot tyytyväisyydestä ovat

$$\left[\begin{array}{cccccccccccccccccccccccccccc} 3&4&3&3&2&5&2&1&4&3&4&3&3&2&2&2&1&3&1&3&4&3&3&3&3&5&4&1&2&1 \end{array}\right]^T.$$

Havaintojen frekvenssijakauma on

E_i	f_i
1	5
2	6
3	12
4	5
5	2

Silloin kun tilastollisen muuttujan mitta-asteikko on luokallinen, tai diskreetti numeerinen, niin luokkajaon tekeminen on selkeää. Jos luokkia on hyvin suuri määrä, siten että niiden käsittely on hankalaa tai jokaiseen luokkaan kuuluu vain pieni määrä havaintoja, niin luokkia voidaan mahdollisesti yhdistellä. Tällöin luokkia tulee pienempi määrä ja luokkiin kuuluu enemmän havaintoja.

Jos muuttujan mitta-asteikko on jatkuva numeerinen, niin frekvenssijakaumaa ei voida suoraan käyttää. Tällöin muuttujan mahdollisten arvojen joukko jaetaan toisensa poissulkeviin väleihin, jolloin jokainen mahdollinen arvo kuuluu täsmälleen yhteen luokkaan. Tämän jälkeen frekvenssijakauma muodostetaan väleistä muodostettujen luokkien avulla, ja tätä kutsutaan luokitelluksi frekvenssijakaumaksi. On huomion arvoista, että vaikka muuttujan mitta-asteikko olisi jatkuva numeerinen, niin käytännössä aina tapahtuu havaintojen luokittelua mittaustarkkuuden vuoksi. Esimerkiksi kahden desimaalin arvolla havaintoja antava mittauslaite voi antaa kaikille luvuille väliltä $[0.005, 0.015)$ havainnon 0.01.

Esimerkki 6. Tarkastellaan erään opiskelijoiden joukon pituutta senttimetreissä. Joukosta kerätään havaintoaineisto 40 tilastostoyksiköltä. Saadaan havaintojen arvot

174.72	182.29	192.11	173.09	180.44	181.93	185.96	179.32	194.89	180.03
183.92	187.87	178.25	173.72	193.48	164.82	187.15	181.25	188.09	184.03
182.14	159.10	178.63	181.18	167.53	150.34	170.84	163.32	173.05	169.53
172.67	169.73	175.03	165.51	162.06	163.33	155.42	161.18	163.59	165.77

Muodostetaan 5cm levyiset luokat sekä frekvenssijakauma

E_i	f_i
$(150, 155]$	1
$(155, 160]$	2
$(160, 165]$	6
$(165, 170]$	5
$(170, 175]$	6
$(175, 180]$	4
$(180, 185]$	9
$(185, 190]$	4
$(190, 195]$	3

Jos tilastollinen muuttuja on mitattu vähintään järjestysasteikolla, niin voidaan havaintoja kuvata summafrekvenssillä F_i, tai vastaavasti luokitellulla summafrekvenssillä. Summafrekvenssillä kuvataan kuinka monta havaintoa kuuluu luokkaan E_i **tai sitä edeltäviin luokkiin**. Summafrekvenssit saadaan laskettua frekvenssien avulla

$$\begin{aligned} F_1 &= f_1 \\ F_2 &= f_1 + f_2 && = F_1 + f_2 \\ F_3 &= f_1 + f_2 + f_3 && = F_2 + f_3 \\ &\vdots \\ F_k &= f_1 + f_2 + \cdots + f_k && = F_{k-1} + f_k \end{aligned}$$

Summafrekvenssijakauma on siis muotoa $(E_i, F_i), i = 1, \ldots, k$. Jälleen määritellään suhteellinen summafrekvenssijakauma $(E_i, \frac{F_i}{n}), i = 1, \ldots, k$, sekä prosentuaalinen summafrekvenssijakauma $(E_i, \frac{F_i}{n} \cdot 100\%), i = 1, \ldots, k$.

Esimerkki 7. Esimerkissä 5 muodostettiin frekvenssijakauma asiakastyytyväisyysaineistosta. Summafrekvenssijakauma samalle aineistolle on

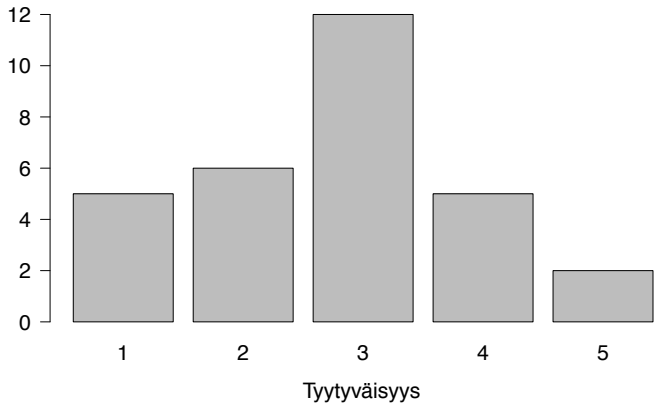
E_i	F_i
1	5
2	$5 + 6 = 11$
3	$11 + 12 = 23$
4	$23 + 5 = 28$
5	$28 + 2 = 30$

Frekvenssi- että summafrekvenssijakauma tiivistävät informaatiota havaintoaineistosta hallittavampaan muotoon sillä kustannuksella, että havaintojen keräämisjärjestyksen informaatio kadotetaan.

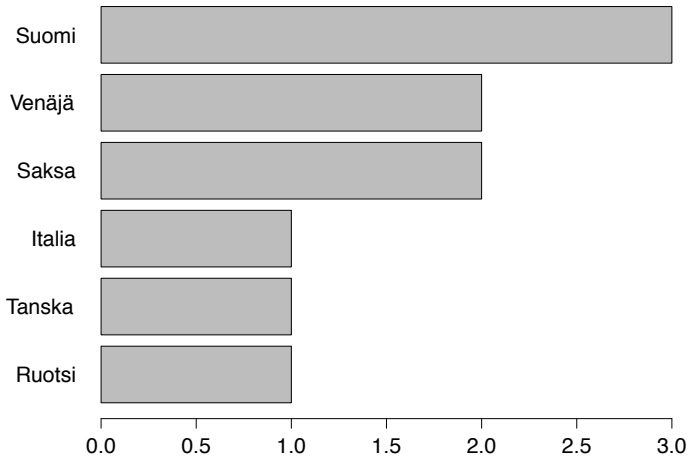
2.2.1 Pylväskuviot ja ympyrädiagrammi

Graafisia menetelmiä kannattaa käyttää aina aineiston havainnollistamisessa. Kuvien avulla aineisto saadaan usein nopeasti tulkittavaan muotoon. Usein kuvien avulla saadaan lisäksi välittömästi informaatiota, jos käytössä olevat tilastolliset mallit eivät ole riittäviä selittämään aineistoa.

Yksinkertaisia luokitellun aineiston havainnollistamiseen käytettäviä kuvatyyppejä ovat erilaiset pylväskuviot. Pylväskuvioissa **pylvään korkeus** ilmaisee määrää, ja sitä voidaan käyttää kun havainnot ovat luokallisia tai diskreettejä numeerisia. Usein käytetty pylväskuvio on **pystypylväskuvio**, etenkin jos kuvataan jonkin asian muuttumista ajassa vaaka-akselilla vasemmalta oikealle, tai mitta-asteikko on ordinaalinen ja järjestys kasvaa vasemmalta oikealle. Pylväät voidaan esittää myös **vaakapylväskuviona**, jolloin pylväiden nimet voidaan kirjoittaa selvemmin kuvaan. Jos luokilla ei



Kuva 2.1: Esimerkin 4 aineisto pystypylväskuvion muotoon piirrettynä.



Kuva 2.2: Esimerkin 4 aineisto vaakapylväskuvion muotoon piirrettynä.

ole luonnollista järjestystä, on hyvä piirtää pylväät suuruusjärjestyksessä joko vasemmalta oikealle tai ylhäältä alas. Esimerkin 5 pystypylväskuvio on piirretty Kuvaan 2.1, ja Kuvassa 2.2 on vaakapylväskuvio Esimerkin 4 aineistolle.

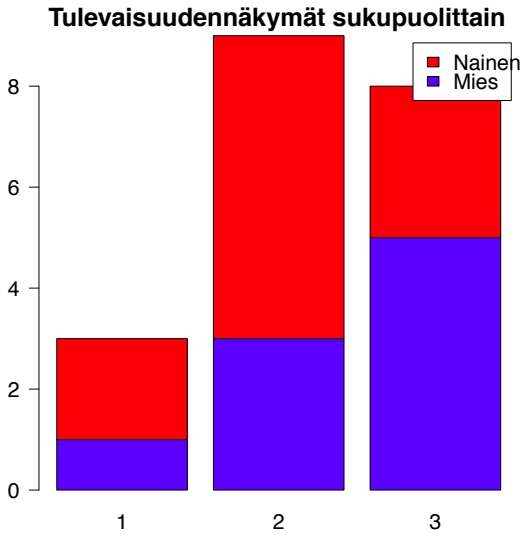
Jos kuvissa mielenkiinnon kohteena on useita tilastollisia muuttujia, ja halutaan tarkastella onko jakaumissa eroja ryhmien välillä, voidaan käyttää hyödyksi **summapylväskuvioita** ja **pylväsryhmäkuvioita**.

Esimerkki 8. Tarkastellaan kyselytutkimuksen aineistoa, jossa opiskelijoilta kysyttiin heidän mielipidettään oman alansa työllisyystilanteesta. Mielipide työllisyystilanteesta pyydettiin järjestysasteikolla

huonot näkymät = 1, *ok näkymät* = 2, ja *hyvät näkymät* = 3. Kerättiin havaintoaineisto

Opiskelija	Sukupuoli	Mielipide	Opiskelija	Sukupuoli	Mielipide
1	Mies	3	11	Mies	1
2	Nainen	1	12	Mies	3
3	Nainen	2	13	Nainen	2
4	Mies	2	14	Nainen	3
5	Mies	2	15	Nainen	1
6	Mies	3	16	Nainen	2
7	Nainen	3	17	Nainen	2
8	Mies	3	18	Mies	3
9	Nainen	2	19	Mies	2
10	Nainen	2	20	Nainen	3

Havaintoaineisto ollaan piirretty Kuvaan 2.3 summapylväskuviona sekä Kuvaan 2.4 ryhmäpylväskuviona

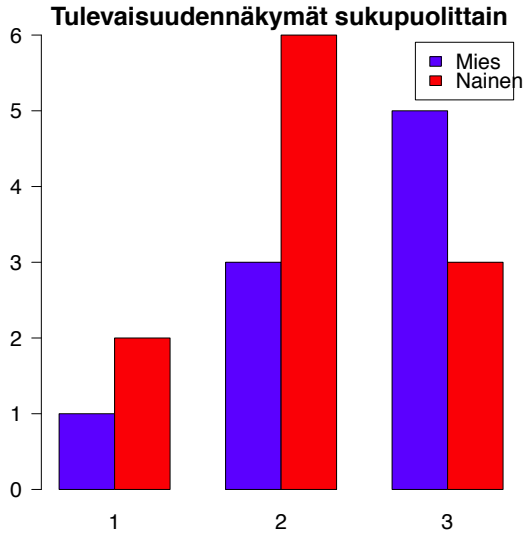


Kuva 2.3: Esimerkin 8 aineisto summapylväskuvion muotoon piirrettynä.

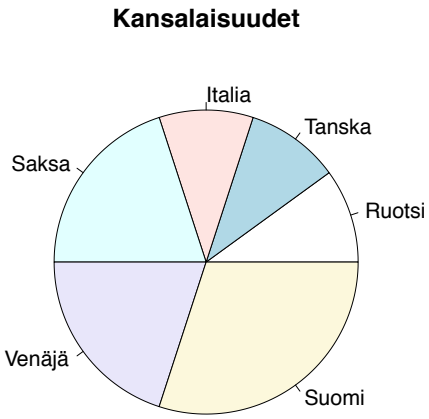
Usein mediassa käytetty graafinen menetelmä on **ympyrädiagrammi** (piirakkakuvio), jossa havaintoaineisto piirretään ympyrän muotoon. Ympyrä jaetaan sektoreihin jokaista luokkaa vastaten ja sektorin koko vastaa luokan suhteellista osuutta jakaumassa. Joskus sektorit suositellaan piirtämään suuruusjärjestyksessä myötäpäivään kiertäen, siten että aloitetaan kello kolmesta. Esimerkin 4 aineisto on piirretty piirakkakuvion muotoon Kuvassa 2.5.

2.2.2 Histogrammi ja frekvenssimonikulmio

Histogrammi on pylväskuvio luokitellulle frekvenssijakaumalle, jossa **pylvään ala** kuvaa luokkien frekvenssiä. Usein käytäntönä on luokitella aineisto pääosin yhtä leveisiin luokkiin lukuunottamatta mahdollisesti äärimmäisiä luokkia, ettei histogrammilla kuvattavan jakauman muotoa päästä vääristämään. Histogrammin pylväät piirretään kiinni toisiinsa havainnollistamaan mitta-asteikon jatkuvuutta. Histogrammin muoto saattaa riippua hieman luokittelun määritelmästä, eli määritelmästä mihin luokkiin kuuluvat päätepisteitä vastaavat havainnot. Tämä käytäntö saattaa vaihdella tilastollisesta ohjelmasta toiseen. Kuvaan 2.6 on piirretty Esimerkin 6 aineisto.



Kuva 2.4: Esimerkin 8 aineisto ryhmäpylväskuvion muotoon piirrettynä.



Kuva 2.5: Esimerkin 4 aineisto piirakkakuvion muotoon piirrettynä.

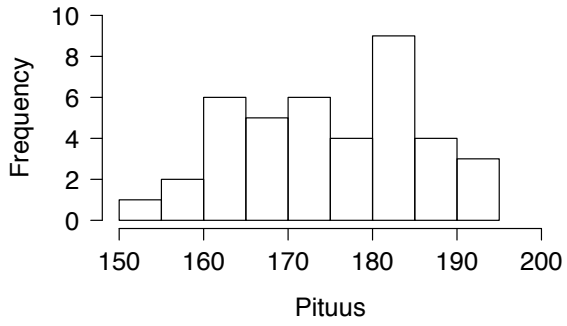
Frekvenssimonikulmio on eräänlainen approksimointi populaatiojakaumalle, jossa luokiteltu frekvenssijakauma piirretään yhdistämällä luokkien keskipisteet ja luokkien frekvenssiarvot murtoviivalla. Frekvenssimonikulmion päätepisteet voidaan asettaa ylimääräisten luokkien luokkakeskipisteisiin, joissa se saa arvon 0. Kuvaan 2.7 on piirretty Esimerkin 6 aineiston frekvenssimonikulmio histogrammin lisäksi.

2.3 Jakauman tunnusluvut

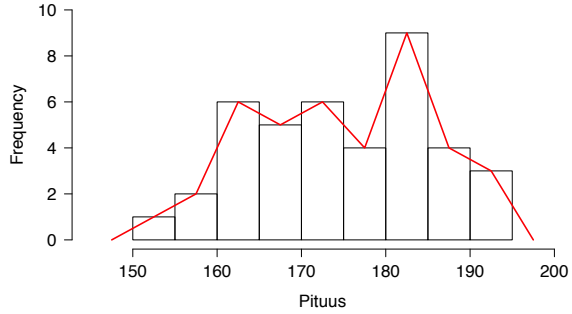
Joskus halutaan tiivistää havaintoaineiston informaatio frekvenssijakaumaakin pienemmäksi. Tällöin otetaan käyttöön tunnusluvut. Jakauman tunnusluvuilla kuvataan jotain jakauman piirteitä. Tunnuslukuja ovat esimerkiksi jakauman **keskusmomenteista** johdetut luvut.

Määritelmä 16 (Keskusmomentti)**.** Satunnaismuuttujan X jakauman r. keskusmomentti on

$$\mu_r := \mathsf{E}\left((X - \mathsf{E}\left(X\right))^r\right), r = 1, 2, 3, \ldots,$$



Kuva 2.6: Esimerkin 6 aineisto histogrammin muotoon piirrettynä.



Kuva 2.7: Esimerkin 6 aineisto frekvenssimonikulmiolla esitettynä.

jossa $\mathrm{E}\,(X)$ on satunnaismuuttujan odotusarvo.

Määritelmän mukaisesti ensimmäinen keskusmomentti on 0. Toisesta, kolmannesta ja neljännestä keskusmomentista voidaan laskea jakauman **varianssi**, **vinouskerroin** ja **huipukkuuskerroin**, joilla voidaan kuvailla erilaisia jakauman ominaisuuksia.

Havaintoaineistosta laskettua keskusmomentteja merkitään

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r, r = 1, 2, 3, \ldots,$$

jossa $\bar{x}$ on otoskeskiarvo

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Jakauman kuvailemiseksi kannattaa käyttää useita tunnuslukuja sekä graafisia esityksiä, sillä kaikissa esityksissä ollaan pyritty tiivistämään informaatioita ja mahdollisesti jotain tärkeitä piirteitä ollaan näin menetetty.

2.3.1 Sijainti

Jakauman sijaintia muuttuja-akselilla kuvataan tunnusluvuilla, joita voidaan kutsua keskiluvuiksi. Eri tilanteisiin ja erilaisille tilastollisille muuttujille on määritelty monenlaisia keskilukuja, joista käydään tärkeimpiä läpi.

Usein keskiluvut kuvaavat havaittujen arvojen keskimääräistä suuruutta, mutta jakauman otoskvantiilien (otosfraktiilien) avulla kuvataan jakauman suhteellisten osien sijoittumista muuttuja-akselille. Otoskvantiileja voidaan käyttää järjestys-, intervalli- ja suhdeasteikollisten muuttujien sijainnin kuvailemiseen.

Määritelmä 17 (Otoskvantiili)**.** Jakauman otoskvantiili $q(p)$ on sellainen luku, että enintään $100p\%$ havainnoista on pienempiä kuin $q(p)$ ja enintään $100(1-p)\%$ havainnoista on suurempia kuin $q(p)$. Kiinnitetään lisäksi luvut $q(0) = \min_i(x_i)$ ja $q(1) = \max_i(x_i)$, eli pienin ja suurin otoskvantiili.

Yleisimmät otoskvantiilien joukot tunnetaan omilla nimillään.

Otoskvantiili	p
Mediaani	0.5
Kvartiilit	$0.25, 0, 50, 0.75$
Kvintiilit	$0.2, 0.4, 0.6, 0.8$
Desiilit	$0.1, 0.2, 0.3, \ldots, 0.9$
Sentiilit	$0.01, 0.02, \ldots, 0.99$

Havaittujen arvojen keskimääräistä suuruutta kuvaa mediaani, eli luku $q(0.5)$, joka on suurempi tai yhtäsuuri kuin puolet havaintoaineiston arvoista.

Otoskvantiilin määritelmästä huomataan, että luvut eivät ole aina yksikäsitteisesti määriteltyjä. Tarkastellaan esimerkiksi tilannetta, jossa meillä on neljän havainnon $1, 2, 3, 4$ aineisto. Nythän mikä tahansa luku väliltä $[2, 3)$ toteuttaa mediaanin määritelmän. Tällä kurssilla käytetään ns. osuusajattelua, jotta otoskvantiilit saadaan määrättyä systemaattisesti. Tarkastellaan osuusajattelua kvartiilien laskemisessa. Tarkastellaan ensin tilannetta, jossa havaintoja on n kappaletta siten, että $n/2$ ei ole tasaluku. Jos havainnot ovat järjestetty suuruusjärjestykseen, niin $\lceil n/2 \rceil$. alkio on suurempi tai yhtäsuuri kuin $(n-1)/2$ havainnoista, ja pienempi tai yhtäsuuri kuin $(n-1)/2$ havainnoista. Merkintä $\lceil x \rceil$ tarkoittaa lukua x pyöristettynä ylöspäin seuraavaan kokonaislukuun. Samoin toimitaan $0.25-$ ja 0.75-kvartiileja (ala- ja ylä-kvartiili) hakiessa. Hakiessa järjestettyjen havaintojen joukosta havaintoa, jota pienempiä olisi $n/4$ osuus kaikista havainnoista, toteutuu tämä ehto indeksillä $\lceil n/4 \rceil$. Samoin 0.75-kvartiili löydetään indeksillä $\lceil 3n/4 \rceil$.

Jos luvut $\frac{n}{4}, \frac{n}{2}, \frac{3n}{4}$ ovat tasalukuja, niin vastaavat otosmediaani, ala- ja yläotoskvartiilit määritellään kyseisen luvun ja seuraavan luvun keskiarvoiksi. Esimerkiksi jos $n/2 = 9$, niin mediaaniksi otetaan $9.$ ja $10.$ lukujen keskiarvo.

Esimerkki 9. Ollaan kerätty havaintoaineisto

14	14	14	29	29	43	43	43
71	71	71	86	86	100	114	143
171	300	400					

Koska havaintoaineisto koostuu yhdeksästätoista havainnosta, löydetään mediaani suuruusjärjestyksessä $\lceil 19/2 \rceil. = 10.$ suurimpana lukuna, $q(0.5) = 71$. Samoin alakvartiili löytyy suuruusjärjestyksessä $\lceil 19/4 \rceil. = 5.$ suurimpana lukuna, $q(0.25) = 29$, ja yläkvartiili suuruusjärjestyksessä $\lceil 3 \cdot 19/4 \rceil. = 15.$ suurimpana lukuna, $q(0.75) = 114$.

Esimerkki 10. Tarkastellaan Esimerkin 6 havaintoaineistoa, joka koostui 40 pituusmittauksesta. Suuruusjärjestyksessä havainnot ovat

150.34	155.42	159.10	161.18	162.06	163.32	163.33	163.59	164.82	165.51
165.77	167.54	169.53	169.73	170.84	172.67	173.05	173.09	173.72	174.72
175.03	178.25	178.63	179.32	180.03	180.44	181.18	181.25	181.93	182.14
182.29	183.92	184.03	185.96	187.15	187.87	188.09	192.11	193.48	194.89

Koska nyt $n = 40$ ja $n/4 = 10, n/2 = 20$ ja $3n/4 = 30$, niin otoskvartiilit ovat

$$q(0.25) = \frac{165.51 + 165.77}{2} = 165.64$$
$$q(0.50) = \frac{174.72 + 175.03}{2} = 174.875$$
$$q(0.75) = \frac{182.14 + 182.29}{2} = 182.215$$

Yleisin keskiluku on havaintoaineistosta intervalli- tai suhdeasteikollisille muuttujille laskettu aritmeettinen keskiarvo. Kun tarkastellaan otosta jostain populaatiosta, kutsutaan tästä aineistosta laskettua aritmeettista keskiarvoa otoskeskiarvoksi.

Määritelmä 18 (Otoskeskiarvo)**.**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Jos numeerinen havaintoaineisto on luokiteltu k luokkaan, voidaan frekvenssijakauman frekvensseistä f_i ja luokkien E_i luokkakeskipisteistä x_i laskea aritmeettinen keskiarvo, eli frekvenssijakauman otoskeskiarvo.

Määritelmä 19 (Frekvenssijakauman otoskeskiarvo)**.**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

Esimerkki 11. Eräälle kurssille osallistui 20 opiskelijaa. Opiskelijat saivat kurssista arvosanat

4	5	3	3	2	5	5	2	2	1
3	4	4	5	5	5	5	3	2	4

Arvosanojen aritmeettinen keskiarvo

$$\bar{x} = \frac{4+5+3+3+2+5+5+2+2+1+3+4+4+5+5+5+5+3+2+4}{20} = 3.6$$

Esimerkki 12. Muodostetaan edellisen esimerkin frekvenssijakauma arvosanoille

E_i	f_i
1	1
2	4
3	4
4	4
5	7

Nyt frekvenssijakauman luokkakeskipisteet ovat $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5$ ja frekvenssijakauman otoskeskiarvo on

$$\bar{x} = \frac{1}{20}(1 \cdot 1 + 4 \cdot 2 + 4 \cdot 3 + 4 \cdot 4 + 7 \cdot 5) = \frac{72}{20} = 3.6.$$

Otoskeskiarvo on hyvin herkkä yksittäisten havaintojen arvoille. Kaavaa tarkastelemalla nähdään, että yhden havainnon x_i vaikutus keskiarvoon on x_i/n. Tämän vuoksi joskus on tarpeen jättää pois joitain pienimpiä sekä suurimpia arvoja, jolloin saadaan ääriarvojen vaikutus poistettua aritmeettisesta keskiarvosta.

Määritelmä 20 (Leikattu otoskeskiarvo)**.** Jätetään havainnoista pois $100p\%$ pienimmistä ja $100p\%$ suurimmista arvosta, ja lasketaan otoskeskiarvo jäljelle jäävistä arvoista. Tätä keskilukua kutsutaan leikatuksi otoskeskiarvoksi $\bar{x}_p$.

Esimerkki 13. Tarkastellaan havaintoaineistoa

$$\begin{array}{cccccccc} 14 & 14 & 14 & 29 & 29 & 43 & 43 & 43 \\ 71 & 71 & 71 & 86 & 86 & 100 & 114 & 143 \\ 171 & 300 & 400 & & & & & \end{array}$$

Tästä aineistosta laskettu otoskeskiarvo on $\bar{x} = 96.94737$. Lasketaan leikattu otoskeskiarvo $\bar{x}_{0.15}$. Koska $0.15 \cdot 19 = 2.85$, jätetään aineistosta pois 2 pienintä ja 2 suurinta havaintoa, ja lasketaan otoskeskiarvo.

$$\begin{aligned} \bar{x}_{0.15} &= \frac{14+29+29+43+43+43+71+71+71+86+86+100+114+143+171}{15} \\ &= 74.26667 \approx 74. \end{aligned}$$

Kaikille mitta-asteikoille sopiva keskiluku on moodi, joka on suurimman frekvenssin omaava muuttujan arvo tai luokka. Huomaa, että jos useammalla arvolla tai luokalla on sama frekvenssi, niin moodi ei ole yksikäsitteinen.

Määritelmä 21 (Tyyppiarvo eli moodi)**.** Havaintoaineiston moodi mode on aineiston suurimman frekvenssin omaava muuttujan arvo tai luokka.

Esimerkki 14. Esimerkin 4 aineiston frekvenssijakauma on

E_i	f_i
Suomi	3
Ruotsi	1
Saksa	2
Tanska	1
Italia	1
Venäjä	2,

jolloin tämän aineiston moodi on mode = Suomi. Esimerkin 12 jakauma on

E_i	f_i
1	1
2	4
3	4
4	4
5	7,

jolloin tämän aineiston moodi on mode $= 5$.

Kun havaintoarvot ovat positiivisia lukuja, voidaan aineistosta laskea geometrinen keskiarvo tai harmoninen keskiarvo. Geometrinen keskiarvo soveltuu tilanteisiin, joissa halutaan etsiä usein suhteellisen muutoksen keskiarvo.

Määritelmä 22 (Geometrinen keskiarvo)**.**

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Esimerkki 15. Erään tuotteen vuosittainen hinnanmuutos neljänä peräkkäisenä vuotena on ollut $+30\%$, $+15\%$, $+9\%$,-6%. Otoskeskiarvo vuosittaiselle hinnanmuutokselle on

$$\bar{x} = \frac{1}{4}(30 + 15 + 9 - 6)\% = 12\%.$$

Geometrinen keskiarvo vuosittaiselle hinnanmuutokselle on

$$\bar{x}_g = \sqrt[4]{1.3 \cdot 1.15 \cdot 1.09 \cdot 0.94} = \approx 1.1125.$$

Olkoon tarkasteltavan tuoteen hinta alussa 1000 EUR. Kun tuotteen hinta muuttuu edelläolevalla tavalla, niin neljän vuoden jälkeen tuotteen hinta on

$$1000 \cdot 1.3 \cdot 1.15 \cdot 1.09 \cdot 0.94 \,\text{EUR} \;= 1531.78 \,\text{EUR}$$

Käyttäen otoskeskiarvoa vuosittaisena hinnanmuutoksena, tuotteen hinta neljän vuoden jälkeen olisi

$$1000 \cdot 1.12^4 \,\text{EUR} \;= 1573.52 \,\text{EUR}\,,$$

kun taas geometrista otoskeskiarvoa käyttäen saadaan

$$1000 \cdot (\sqrt[4]{1.3 \cdot 1.15 \cdot 1.09 \cdot 0.94})^4 \,\text{EUR} \;= 1531.78 \,\text{EUR}\,.$$

On hyvä huomata myös, että geometrisen keskiarvon logaritmi on otoskeskiarvo logaritmimuunnetulle havaintoaineistolle.

$$\log(\bar{x}_g) = \log(\sqrt[n]{x_1 \cdot x_2 \cdots x_n}) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i).$$

Harmoninen keskiarvo soveltuu tilanteisiin, joissa mittaukset ovat lukujen suhteita, kuten nopeuksia (matka/aika). Harmoninen keskiarvo painottaa havaintoaineiston pieniä arvoja enemmän kuin suuria, joten voidaan ajatella, että harmoninen keskiarvo on otoskeskiarvoa robustimpi sijainnin tunnusluku.

Määritelmä 23 (Harmoninen keskiarvo)**.**

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Esimerkki 16. Tarkastellaan 100 km pituisen välin ajamista edestakaisin. Ajetaan matka toiseen suuntaan nopeudella 80 km/h ja palataan takaisin nopeudella 100 km/h. Matkaan kuluu aikaa $(100/80 + 100/100)$ h $= 9/4$ h. Otoskeskiarvo keskinopeudelle on $\bar{x} = (80 + 100)/2$ km/h $= 90$ km/h. Jos kuljettaisiin tällä keskinopeudella 100 km matka edestakaisin, matkaan kuluu $200/90$ h $= 20/9$ h. Harmoninen keskiarvo nopeuksille on

$$\bar{x}_h = \frac{2}{\frac{1}{80} + \frac{1}{100}} = \frac{800}{9}.$$

Keskinopeudella $\bar{x}_h$ kuljettuna edestakaiseen matkaan kuluu $(200/(800/9))$ h $= 9/4$ h.

On hyvä huomata, että harmoninen keskiarvo on havaintoaineiston käänteislukujen otoskeskiarvon käänteisluku

$$(\bar{x}_h)^{-1} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}.$$

2.3.2 Hajonta

Tärkeä jakaumien peruspiirre on jakauman hajonta. Hajontaa mittaavat tunnusluvut ilmentävät havaintojen vaihtelun suuruutta havaintoaineistossa. Jakaumaa ajatellen hajonnan tunnusluvut kuvaavat kuinka tiukasti tai laveasti mahdolliset arvot ovat jakautuneet. Mitä lähempänä hajonnan tunnusluvut ovat nollaa, sitä vähemmän arvot vaihtelevat.

Yksinkertaisin aineistosta laskettava hajonnan tunnusluku on vaihteluvälin pituus (range), joka kuvaa kuinka suurelta väliltä kaikki aineiston arvot löytyvät. Tämä tunnusluku riippuu ainoastaan pienimmästä ja suurimmasta havainnosta, joka määrittävät vaihteluvälin päätepisteet, eikä kerro mitään muiden havaintojen jakautumisesta tälle välille. Vaihteluvälin pituus on käyttökelpoinen vain, jos havaintoaineiston koostuu hyvin pienestä määrästä, esimerkiksi alle viidestä havainnosta.

Määritelmä 24 (Vaihteluvälin pituus)**.** Olkoon $q(0)$ havaintoaineiston pienin arvo ja $q(1)$ suurin arvo. Havaintoaineiston vaihteluvälin pituus on

$$r = q(1) - q(0)$$

Esimerkki 17. Havaintoaineiston

$$\begin{array}{cccccccc} 14 & 14 & 14 & 29 & 29 & 43 & 43 & 43 \\ 71 & 71 & 71 & 86 & 86 & 100 & 114 & 143 \\ 171 & 300 & 400 & & & & & \end{array}$$

vaihteluvälin pituus on

$$r = q(1) - q(0) = 400 - 14 = 386.$$

Havaintoaineiston kvartiileista voidaan laskea kvartiilivälin pituus (interquartile range), joka on puolet havainnoista aineiston keskeltä sisältävän välin pituus.

Määritelmä 25 (Kvartiilivälin pituus)**.** Olkoon $q(0.25)$ ja $q(0.75)$ aineiston ala- ja yläkvartiilit. Havaintoaineiston kvartiilivälin pituus on

$$q_{\mathrm{r}} = q(0.75) - q(0.25)$$

Esimerkki 18. Edellisen esimerkin havaintoaineiston ala- ja yläkvartiilit ovat

$$\begin{aligned} q(0.25) &= 29 \\ q(0.75) &= 114, \end{aligned}$$

joten kvartiiliväli on

$$q_{\mathrm{r}} = q(0.75) - q(0.25) = 85.$$

Kvartiilivälin avulla voidaan laskea havaintoaineiston kvartiilipoikkeama tai pseudokeskihajonta. Kvartiilipoikkeamaa käytetään arvioimaan populaation keskihajontaa varsinkin jos populaation sijaintia arvioidaan mediaanilla. Kvartiilipoikkeama määritellään

$$q_{\mathrm{rd}} = \frac{q_{\mathrm{r}}}{2}.$$

Pseudokeskihajonnassa kvartiiliväli suhteutetaan standardinormaalijakauman kvartiiliväliin. Standardinormaalijakauman (normaalijakauma, jolla on odotusarvo 0 ja keskihajonta 1) kvartiilivälin pituus on 1.34898. Pseudokeskihajonta saadaan jakamalla kvartiiliväli pyöristetyllä luvulla 1.35, ja käytännössä se siis ilmoittaa *montako kertaa havaintoaineiston kvartiiliväli on standardinormaalijakauman kvartiiliväli*.

Määritelmä 26 (Pseudokeskihajonta)**.** Olkoon q_{r} havaintoaineistosta laskettu kvartiiliväli. Aineiston pseudokeskihajonta on

$$s_{\mathrm{pseudo}} = \frac{q_{\mathrm{r}}}{1.35}.$$

Esimerkki 19. Edellisen esimerkin havaintoaineistolle kvartiilipoikkeama ja pseudokeskihajonta ovat

$$\begin{aligned} q_{\mathrm{rd}} &= 42.5 \\ s_{\mathrm{pseudo}} &= 62.96296 \end{aligned}$$

Yleisin käytetty hajonnan tunnusluku on otosvarianssi, joka lasketaan samantapaisesti kuin toinen keskusmomentti m_2, ainoana erona on tärkeää huomata luvulla $(n-1)$ jakaminen havaintoaineiston koon n sijaan.

Määritelmä 27 (Otosvarianssi ja otoskeskihajonta)**.** Olkoon $\bar{x}$ havantoaineistosta $x_1, \ldots, x_n$ laskettu aritmeettinen keskiarvo. Havaintoaineistosta laskettu otosvarianssi on

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Otosvarianssista voidaan laskea otoskeskihajonta $s = \sqrt{s^2}$.

Esimerkki 20. Tarkastellaan arvosana-esimerkkiä, jossa havaintoaineistona oli

$$\begin{array}{cccccccccc} 4 & 5 & 3 & 3 & 2 & 5 & 5 & 2 & 2 & 1 \\ 3 & 4 & 4 & 5 & 5 & 5 & 5 & 3 & 2 & 4 \end{array}$$

Arvosanojen aritmeettinen keskiarvo oli $\bar{x} = 3.6$. Otosvarianssi on

$$\begin{aligned} s^2 = \quad & \frac{1}{20-1}\Big((4-3.6)^2 + (5-3.6)^2 + (3-3.6)^2 + (3-3.6)^2 + (2-3.6)^2 \\ & + (5-3.6)^2 + (5-3.6)^2 + (2-3.6)^2 + (2-3.6)^2 + (1-3.6)^2 \\ & + (3-3.6)^2 + (4-3.6)^2 + (4-3.6)^2 + (5-3.6)^2 + (5-3.6)^2 \\ & + (5-3.6)^2 + (5-3.6)^2 + (3-3.6)^2 + (2-3.6)^2 + (4-3.6)^2\Big) \\ = 1.726316 \approx 1.7. \end{aligned}$$

Otoskeskihajonta on $s = \sqrt{1.726316} \approx 1.3$

Lause 1. Otosvarianssi voidaan laskea kaavalla

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right].$$

Todistus. Harjoitustehtävä. □

Jos tilastolliset muuttuja on määritelty suhdeasteikolla, niin sille voidaan laskea mittausyksiköistä riippumaton variaatiokerroin (coefficient of variation), joka ilmaisee hajonnan määrää suhteessa populaation odotusarvoon. Havaintoainestosta lasketun otosvariaatiokertoimen avulla voidaan verrata erilaisilla mittayksiköillä kerättyjä aineistoja.

Määritelmä 28 (Otosvariaatiokerroin)**.** Olkoon $\bar{x}$ ja s tilastollisen muuttujan havainnoista $x_1, \ldots, x_n$ lasketut otoskeskiarvo sekä otoskeskihajonta. Otosvariaatiokerroin on määritelty

$$c_v = \frac{s}{\bar{x}}.$$

Esimerkki 21. Arvosanat ovat mitattu suhdeasteikolla. Mahdolliset arvosanat ovat $(0, 1, 2, \ldots, 5)$ Edellisen esimerkin arvosana-aineistolle otosvariaatiokerroin on

$$c_v = \frac{\sqrt{1.726316}}{3.6} = 0.3649703 \approx 0.36.$$

Toisessa koulussa on käytössä arvosteluasteikko $(0, 1, 2, \ldots, 10)$. Samasta kurssista kyseissä koulussa 20 opiskelijaa keräsivät arvosanat

$$\begin{array}{cccccccccc} 8 & 7 & 9 & 10 & 8 & 10 & 4 & 6 & 9 & 8 \\ 8 & 8 & 6 & 7 & 5 & 9 & 6 & 1 & 7 & 5 \end{array}$$

josta lasketut otoskeskiarvo, otoskeskihajonta sekä otosvariaatiokerroin ovat

$$\begin{aligned} \bar{x} &= 7.05 \\ s &= 2.187885 \\ c_v &= 0.3103383 \approx 0.3, \end{aligned}$$

joten jälkimmäisessä koulussa oli otosvariaatiokertoimen mukaan suhteessa vähemmän vaihtelua arvosanoissa.

2.3.3 Vinous

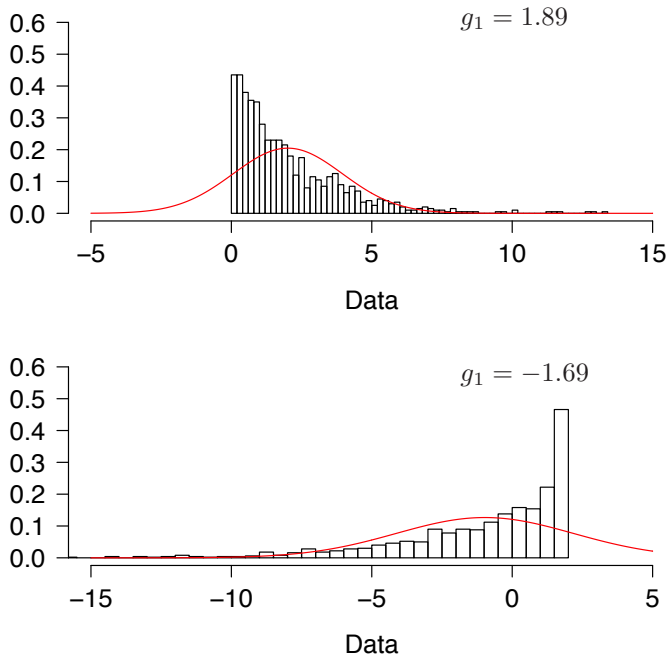
Jos jakauma ei ole symmetrinen jonkin pisteen suhteen, on jakauma **vino**. Jakauman vinous on ikävä piirre kun tarkastellaan jakauman sijaintia. Satunnaismuuttujan X vinous γ_1 on määritelty sen kolmantena standardoituna momenttina

$$\gamma_1 := \mathsf{E}\left(\left(\frac{X-\mu}{\sigma}\right)^3\right),$$

jossa μ ja σ ovat X:n odotusarvo ja keskihajonta. Havaintoaineiston avulla laskettava otosvinouskerroin määritellään 2. ja 3. keskitetyn otosmomentin avulla

$$g_1 := \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2\right)^{3/2}}. \tag{2.1}$$

Jos otosvinouskerroin $g_1 < 0$, on jakauma **vasemmalle vino** ja jos $g_1 > 0$, on jakauma **oikealle vino**. Käytännössä jos aineisto on vasemmalle vino, niin sen jakauman vasen häntä on pidempi. Jos taas aineisto on oikealle vino, niin sen jakauman oikea häntä on pidempi. Muutamasta havainnosta koostuvista aineistoista ei ole mielekästä laskea vinouskertoimen arvoa. Kuvassa 2.8 on havainnoillistettu oikealle sekä vasemmalle vinoja jakaumia.



Kuva 2.8: Oikealle sekä vasemmalle vino jakauma verrattuna standardinormaalijakaumaan.

Esimerkki 22. Tarkastellaan havaintoaineistoa

14	14	14	29	29	43	43	43
71	71	71	86	86	100	114	143
171	300	400					

Aineistosta lasketut aritmeettinen keskiarvo, toinen keskusmomentti ja kolmas keskusmomentti ovat

$$\begin{aligned} \bar{x} &= 96.94737 \\ m_2 &= 9570.576 \\ m_3 &= 1781461, \end{aligned}$$

joiden avulla laskettu otosvinouskerroin on

$$g_1 = 1.902695,$$

joten jakauma on oikealle vino.

2.3.4 Huipukkuus

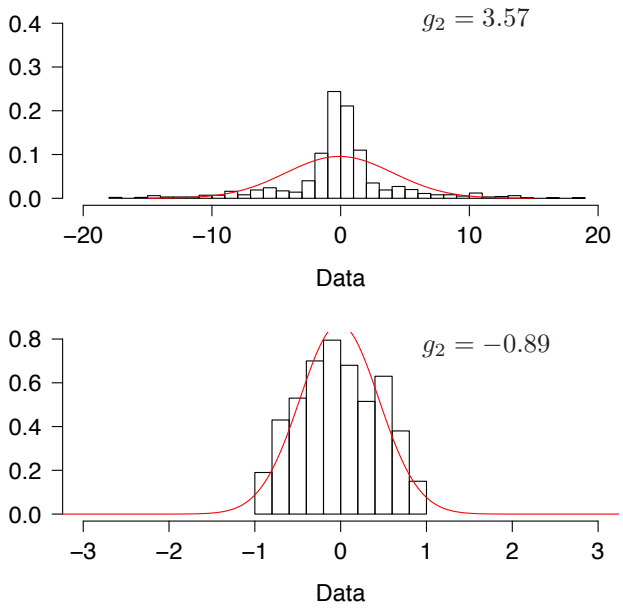
Jakauman muotoa voidaan kuvailla sen *huipukkuutta* kuvaavalla tunnusluvulla. Satunnaismuuttujan X huipukkuus voidaan laskea neljännen standardoitun momentin avulla

$$\gamma_2 := \mathsf{E}\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) - 3,$$

jossa μ ja σ ovat satunnaismuuttujan X odotusarvo sekä keskihajonta. Näin määriteltynä huipukkuus lasketaan suhteessa normaalijakaumaan, jolle huipukkuus on 0. Tämä selittää siis miksi huipukkuudessa vähennetään luku 3. Jakauman huipukkuuden tulkinta on hankalaa, määritelty γ_2 kuvaakin oikeastaan jakauman huipukkuutta normaalijakaumaan verrattavien jakauman häntien paksuuksien avulla. Eli mitä paksummat hännät jakaumalla on verrattuna normaalijakaumaan, niin sitä suurempi on kerroin γ_2. Havaintoaineistosta laskettava otoshuipukkuuskerroin voidaan laskea aineistosta neljännen ja toisen keskitetyn otosmomentin avulla

$$g_2 := \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3.$$

Jos $g_2 > 0$, niin jakauma on **huipukas** ja $g_2 < 0$, niin jakauma on **litteä**. Muutamasta havainnosta koostuvista aineistoista ei ole mielekästä laskea huipukkuuskertoimen arvoa. Kuvassa 2.9 on kuvattu kaksi havaintoaineistoa, joista toinen on huipukkaasta ja toinen on litteästä jakaumasta.



Kuva 2.9: Huipukas sekä litteä jakauma verrattuna standardinormaalijakaumaan.

Esimerkki 23. Tarkastellaan havaintoaineistoa

14	14	14	29	29	43	43	43
71	71	71	86	86	100	114	143
171	300	400					

Aineistosta lasketut aritmeettinen keskiarvo, toinen keskusmomentti ja neljäs keskusmomentti ovat

$$\bar{x} = 96.94737$$
$$m_2 = 9570.576$$
$$m_4 = 546357458,$$

joiden avulla laskettu otoshuipukkuuskerroin on

$$g_2 = 2.964866,$$

joten jakauma on huipukas.

2.4 Tukeyn laatikko-janakuvio

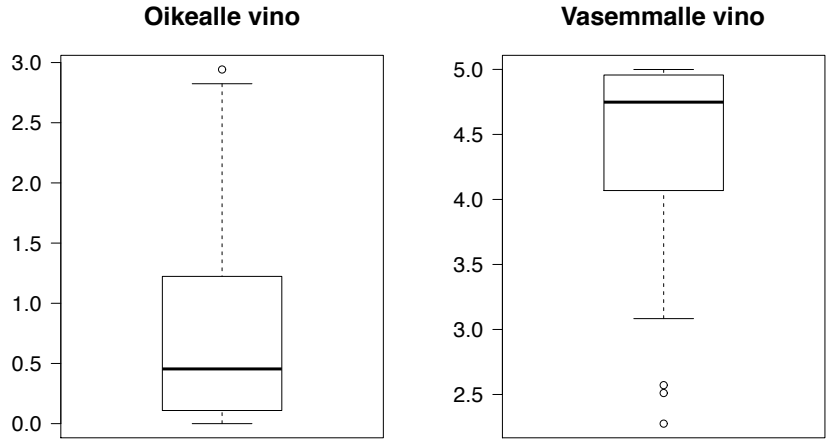
Tukeyn laatikko-janakuvio (box and whisker plot) on havaintoaineiston otoskvartiileista koostettu kuvaaja, josta saadaan välittömästi informaatiota jakauman sijainnin lisäksi jakauman muodosta, sekä poikkeavista havainnoista.

Laatikko-janakuvio koostetaan laskemalla ensin aineistosta otoskvartiilit $q(0.25), q(0.50)$ ja $q(0.75)$ sekä kvartiilivälin pituus $q_{\mathrm{r}} = q(0.75) - q(0.25)$. Tämän lisäksi lasketaan janojen ala- ja ylärajat

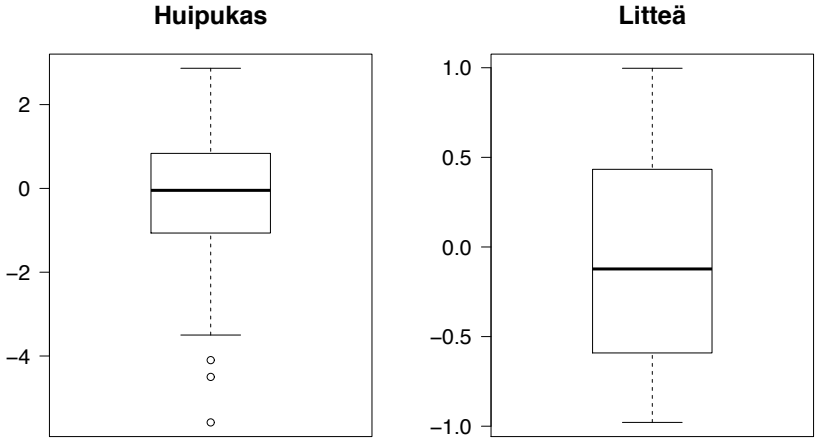
$$q_{\mathrm{d}} = q(0.25) - 1.5q_{\mathrm{r}}$$
$$q_{\mathrm{u}} = q(0.75) + 1.5q_{\mathrm{r}}.$$

Laatikko-janakuvioon piirretään ensin laatikko jonka ala- ja ylärajana ovat $q(0.25)$ ja $q(0.75)$, ja laatikon sisälle piirretään viiva mediaanin kohdalle. Tämän jälkeen etsitään havainnot, jotka ovat joko pienempiä kuin q_{d} tai suurempia kuin q_{u}. Näitä havaintoja pidetään poikkeavina havaintoina, ja nämä merkitään laatikko-janakuvioon erikseen esimerkiksi palloilla. Nyt alajana piirretään pienimmän ei-poikkeavan havainnon sekä alakvartiilin $q(0.25)$ välille, sekä yläjana piirretään yläkvartiilin $q(0.75)$ sekä suurimman ei-poikkeavan havainnon välille.

Laatikko-janakuviolla voidaan tutkia havaintoaineiston jakaumien muotoja tarkastelemalla kvartiilivälilaatikon, mediaanin ja janojen suhteita toisiinsa, sekä tutkimalla poikkeavia havaintoja. Kuvaan 2.14 on piirretty vasemmalle vinosta sekä oikealle vinosta jakaumasta generoitujen havaintoaineistojen laatikko-janakuviot, sekä Kuvaan 2.11 on piirretty huipukkaasta sekä litteästä jakaumasta generoitujen havaintoaineistojen laatikko-janakuvio.



Kuva 2.10: Vasemmalle ja oikealle vinoista jakaumista generoitujen havaintoaineistojen laatikko-janakuviot.



Kuva 2.11: Huipukkaasta ja litteästä jakaumasta generoitujen havaintoaineistojen laatikko-janakuviot.

Esimerkki 24. Ollaan kerätty havaintoaineisto

$$\begin{array}{cccccccc} 14 & 14 & 14 & 29 & 29 & 43 & 43 & 43 \\ 71 & 71 & 71 & 86 & 86 & 100 & 114 & 143 \\ 171 & 300 & 400 & & & & & \end{array}$$

Aiemmin laskettiin kvartiilit

$$\begin{aligned} q(0.25) &= 29 \\ q(0.50) &= 71 \\ q(0.75) &= 114, \end{aligned}$$

joista laskettu kvartiilivälin pituus on

$$q_{\mathrm{r}} = 114 - 29 = 85.$$

Kvartiilivälin pituuden avulla lasketut laatikko-janakuvion janojen rajat ovat

$$\begin{aligned} q_{\mathrm{d}} &= q(0.25) - 1.5q_{\mathrm{r}} = 29 - 1.5 \cdot 85 = -107 \\ q_{\mathrm{u}} &= q(0.75) + 1.5q_{\mathrm{r}} = 114 + 1.5 \cdot 85 = 241.5. \end{aligned}$$

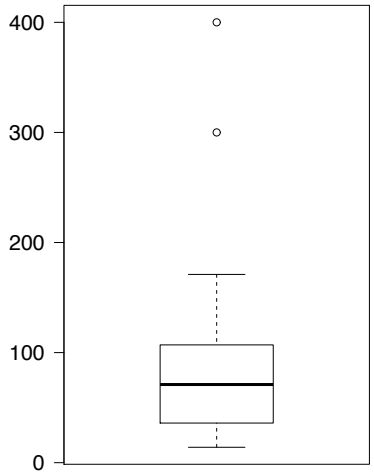
Koska -107 on pienempi kuin kaikki havainnot, piirretään alajana havaintoon 14 asti. Koska havainnot 300 sekä 400 ovat suurempia kuin 241.5, piirretään yläjana pisteeseen 171 asti ja liian suuret havainnot merkitään kuviin palloilla.

2.5 Kvantiilikuvaajat

Havaintoaineiston jakauman muotoa voidaan tutkia graafisesti esimerkiksi pylväskuvioiden ja laatikko-janakuvioiden avulla. Jakaumaa voidaan tutkia myös erilaisten **kvantiilikuvaajien** avulla. Kvantiilikuvan piirtämiseksi järjestetään ensin havaintoaineisto suuruusjärjestykseen. Jos havaintoaineiston koko on n, niin järjestetty havaintoaineisto on

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

Alaindeksiä suluissa käytetään merkitsemään järjestystä, eli $x_{(i)}$ on i. pienin arvo.



Kuva 2.12: Esimerkin 24 havaintoaineiston laatikko-janakuvio.

Tämän jälkeen lasketaan jokaiselle havainnolle $x_{(i)}$, jota vastaavaa arvioitua kertynyttä suhteellista osuutta merkitään f_i. Arvioidaan siis, että $x_{(i)}$ on likimääräisesti f_i-kvantiili, joka voidaan laskea

$$f_i = \frac{i - 0.5}{n}.$$

Esimerkiksi ensimmäinen luku $x_{(1)}$ kymmenen havainnon aineistossa on likimääräisesti 0.05-kvantiili, sillä

$$f_1 = \frac{1 - 0.5}{10} = 0.05.$$

Kvantiilikuva on pisteet $(f_i, x_{(i)}), i = 1, \ldots, n$ piirrettynä pistekuviona (tai porraskuviona). Jos piirrettäisiinkin pisteet $(x_{(i)}, f_i)$ porraskuviona, niin voitaisiin kuvaajaa pitää aineiston empiirisenä kertymäfunktiona.

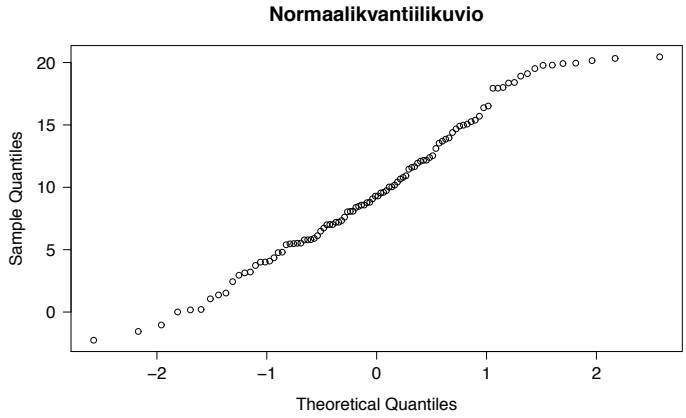
Usein halutaan tarkastella kuinka hyvin havaintoaineisto noudattaa normaalijakaumaa. Tämä tehdään piirtämällä havaintoaineistosta lasketut arvioidut f_i-kvantiilit suhteessa normaalijakauman vastaaviin teoreettisiin kvantiileihin. Mielivaltaisen normaalijakauman $\mathsf{Normal}\left(\mu, \sigma^2\right)$ f-kvantiilille $q(f \mid \mu, \sigma^2)$ pätee

$$\begin{aligned} & q(f \mid 0, 1) = \frac{q(f \mid \mu, \sigma^2) - \mu}{\sigma} \\ \Leftrightarrow \quad & q(f \mid \mu, \sigma^2) = \mu + \sigma q(f \mid 0, 1). \end{aligned}$$

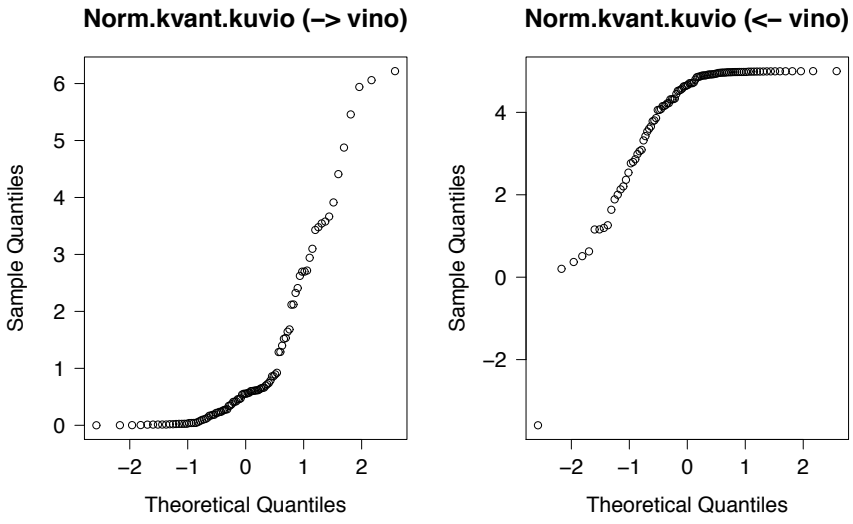
Oletetaan, että havainnot ovat jakaumasta $\mathsf{Normal}\left(\mu, \sigma^2\right)$. Nyt pitäisi päteä likimääräisesti suoran yhtälö

$$x_{(i)} \approx \mu + \sigma q(f_i \mid 0, 1),$$

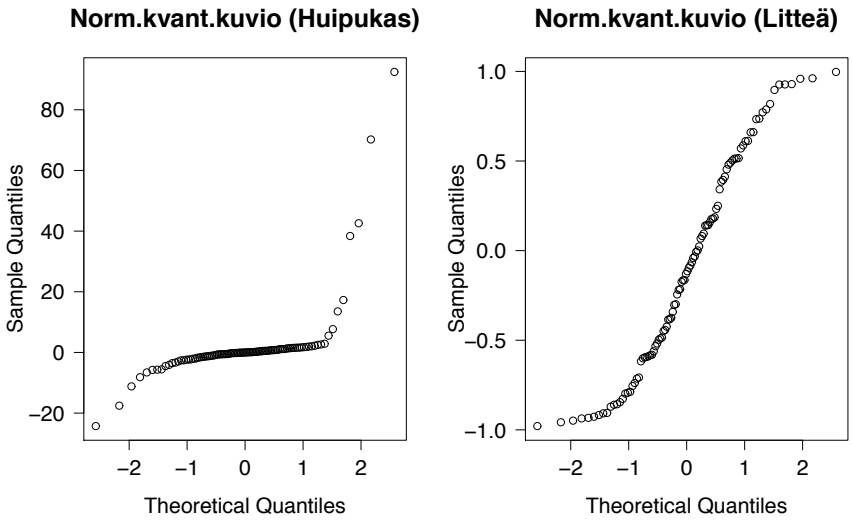
jossa f_i on aineistosta arvioitu kertynyt suhteellinen osuus, joka vastaa f_i-kvantiilia $x_{(i)}$. Jos normaalisuusoletus pätee, niin pisteiden $(q(f_i \mid 0, 1), x_{(i)}), i = 1, \ldots, n$ pitäisi piirrettäessä osua likimääräisesti suoralle. Tätä pistekuviota kutsutaan **normaalikvantiilikuvioksi**. Kuvion pisteiden pitäisi varsinkin kuvaajan keskivaiheilla osua hyvin suoralle, jotta havaintoaineistoa voitaisiin pitää normaalijakautuneena. Kuvaan 2.13 on piirretty normaalikvantiilikuvio havaintoaineistolle ($n = 100$), joka on generoitu jakaumasta $\mathsf{Normal}\left(10, 5^2\right)$. Kuvaan 2.14 on piirretty vasemmalle vinosta sekä oikealle vinosta jakaumasta generoitujen havaintoaineistojen normaalikvantiilikuviot, sekä Kuvaan 2.15 on piirretty huipukkaasta sekä litteästä jakaumasta generoitujen havaintoaineistojen normaalikvantiilikuviot.



Kuva 2.13: Normaalijakautunuteen havaintoaineiston normaalikvantiilikuvio.



Kuva 2.14: Vasemmalle ja oikealle vinojen jakaumien normaalikvantiilikuviot.



Kuva 2.15: Huipukkaan ja litteän jakauman normaalikvantiilikuviot.

Luku 3

Tilastollinen päättely

Tässä kappaleessa tutustutaan tilastolliseen päättelyn perusteisiin. Luonteeltaan tilastollinen päättely on **induktiivista** loogista päättelyä, joka tarkoittaa, että otoksen perusteella yritetään tehdä päätelmiä koko populaatiosta. Induktiiviseen päättelyyn liittyy siis aina jonkinlaista epävarmuutta.

3.1 Yleisiä otossuureita

Kun ajatellaan, että kerättävät tilastollisten muuttujien havainnot x_i ovat saatu satunnaisesti arpomalla populaatiojakaumasta, pidetään havaintoja ennen arvon kiinnittämistä satunnaismuuttujina X_i. Näillä satunnaismuuttujilla on todennäköisyysjakauma, jonka määrittelevät tilastollisten muuttujien arvojen jakauma populaatiossa sekä käytettävä otantamenetelmä. Kutsumme tätä havaintojen jakaumaa populaatiossa **populaatiojakaumaksi**. Tilastollisessa päättely perustuu klassisesti erilaisiin **otossuureisiin**, jotka ovat satunnaismuuttujien $X_1, \ldots, X_n$ funktioita. Yleisesti voidaan merkitä otossuureita funktioina $\Theta(X_1, \ldots, X_n)$. Satunnaismuuttujien funktiot ovat edelleen satunnaismuuttujia, joten otossuureella on **otosjakaumaksi** kutsuttu todennäköisyysjakauma, joka saattaa olla otossuureesta riippuen hyvinkin hankalaa muotoa. Otossuureen keskihajontaa kutsutaan **keskivirheeksi**.

Määritelmä 29 (Keskivirhe)**.** Otossuureen $\Theta(X_1, \ldots, X_n)$ keskihajonta on otossuureen keskivirhe, jota merkitään $\mathsf{SE}(\Theta(X_1, \ldots, X_n))$.

Aiemmin käsiteltyjä havaintoaineiston tunnuslukuja voidaan pitää otossuureina ennen havaintojen arvojen kiinnittämistä. Etenkin otoskvantiilien otosjakaumat ovat hyvin hankalia käsiteltäviä. Tässä kappaleessa on tarkasteltu tilastollisen päättelyn perusotossuureita ja niiden jakaumia. Kaikki otosjakaumat ovat normaalijakaumasta johdettuja jakaumia.

3.1.1 Otoskeskiarvo

Tarkastellaan otoskeskiarvoa, kun havainnot $X_1, \ldots, X_n$ ovat satunnaisia. Merkitään

$$\bar{X} = \Theta(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Oletetaan, että satunnaismuuttujilla X_i on odotusarvo μ sekä varianssi σ^2. Lisäksi oletetaan, että havainnot ovat toisistaan riippumattomia. Tällöin otoskeskiarvon odotusarvolle sekä varianssille pätee

$$\mathsf{E}\left(\bar{X}\right) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left(X_i\right) = \frac{1}{n} n\mu = \mu$$

$$\mathsf{Var}\left(\bar{X}\right) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathsf{Var}\left(X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Otoskeskiarvon keskivirhe on siis $\mathsf{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. Jos havaintojen X_i jakauma on normaalijakauma odotusarvolla μ ja varianssilla σ^2, ja havainnot ovat toisistaan riippumattomat, on otoskeskiarvon

jakauma edelleen normaalijakauma

$$\bar{X} \sim \mathsf{Normal}\left(\mu, \frac{\sigma^2}{n}\right).$$

Vaikka havaintojen jakauma ei olisi normaalijakauma, niin otoskeskiarvolle pätee keskeinen raja-arvolause (Lause 2), jonka mukaan otoskeskiarvon jakauma lähestyy normaalijakaumaa havaintojen määrän kasvaessa.

Lause 2 (Keskeinen raja-arvolause)**.** Olkoon $X_1, \ldots, X_n$ samoinjakautuneita ja toisistaan riippumattomia satunnaismuuttujia siten, että $\mathsf{E}(X_i) = \mu$ ja $\mathsf{Var}(X_i) = \sigma^2 > 0$. Tällöin pätee

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \to \mathsf{Normal}\left(0, \sigma^2\right), \text{ kun } n \to \infty.$$

Keskeisen raja-arvolauseen perusteella voidaan arvioida otoskeskiarvon jakaumaa

$$\bar{X} \dot{\sim} \mathsf{Normal}\left(\mu, \frac{\sigma^2}{n}\right),$$

kun otoskoko on 'suuri'. Merkitään $X \dot{\sim} F_X$, jos X:llä on likimääräisesti jakauma F_X. Suuren otoskoon rajana voidaan pitää esimerkiksi 30 havaintoa, mutta pienikin määrä havaintoja riittää jos otosjakauma on muodoltaa lähellä normaalijakaumaa.

Esimerkki 25. Erään kemikaalipakkauksen painon pitäisi olla 100 grammaa. Kemikaalipakkauksien keskihajonta on 0.2 grammaa. Punnitaan 25 kemikaalipakkausta, ja havaitaan otoskeskiarvo 100.1 grammaa. Otoskeskiarvon keskivirhe $\mathsf{SE}(\bar{X}) = \frac{0.2}{\sqrt{25}}$ grammaa $= 0.04$ grammaa. Todennäköisyys, että otoskeskiarvo poikkeaisi yli 0.1 grammaa jakauman todellisesta odotusarvosta on

$$\mathsf{P}\left(\left|\bar{X} - 100\right| > 0.1\right) = 2\mathsf{P}\left(\frac{\bar{X} - 100}{0.04} > \frac{0.1}{0.04}\right) = 2\mathsf{P}\left(Z > 2.5\right) = 0.0124.$$

Todennäköisyys on hyvin pieni oletuksella, että $\mu = 100$, joten on syytä epäillä olisiko kemikaapakkauksien jakauman odotusarvo oikeasti suurempi kuin 100 grammaa.

Tutkitaankin kahta toisistaan riippumatonta otosta $X_1, \ldots, X_{n_1}$ sekä $Y_1, \ldots, Y_{n_2}$ kahdesta eri populaatiosta, joilla on odotusarvot μ_1 ja μ_2 sekä varianssit σ_1^2 ja σ_2^2. Tarkastellaan kahden havaintoaineiston otoskeskiarvojen erotusta. Erotuksella on odotusarvo sekä varianssi

$$\begin{aligned}
\mathsf{E}\left(\bar{X} - \bar{Y}\right) &= \mathsf{E}\left(\bar{X}\right) - \mathsf{E}\left(\bar{Y}\right) = \mu_1 - \mu_2 \\
\mathsf{Var}\left(\bar{X} - \bar{Y}\right) &= \mathsf{Var}\left(\bar{X}\right) + (-1)^2\mathsf{Var}\left(\bar{Y}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.
\end{aligned}$$

Jos otoskoot ovat suuria, niin pätee likimääräisesti

$$\bar{X} - \bar{Y} \dot{\sim} \mathsf{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Lisäksi, jos populaatiojakaumat ovat normaalijakaumia, niin erotuksen normaalijakautuneisuus pätee tarkasti.

Esimerkki 26. Tarkastellaan kahdelta linjastolta valmistuvien samanlaisten koneenosien pituuksia. Tiedetään, että molempien linjastojen koneen osien pituuksien populaatiovarianssit ovat $\sigma_1^2 = \sigma_2^2 = 0.1$ mm. Molemmilta linjastoilta kerättiin 20 osan otos, ja pituuksiksi havaitaan $\bar{x}_1 = 20.2$ mm sekä $\bar{x}_2 = 20.0$ mm. Lasketaan todennäköisyys, että $\bar{X}_1 - \bar{X}_2 > 0.2$ olettaen $\mu_1 - \mu_2 = 0$,

$$\mathsf{P}\left(\bar{X}_1 - \bar{X}_2 > 0.2\right) = \mathsf{P}\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{0.1/20 + 0.1/20}} > \frac{0.2}{\sqrt{0.1/20 + 0.1/20}}\right) = \mathsf{P}\left(Z > 2\right) = 0.0228.$$

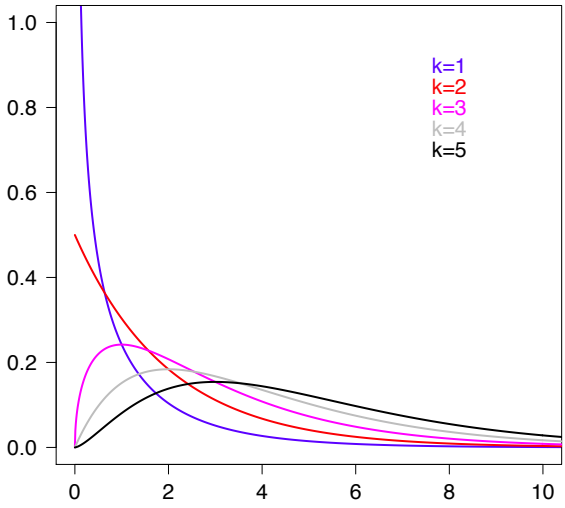
Koska todennäköisyys on kovin pieni oletuksella, että $\mu_1 = \mu_2$, niin luultavasti linjastolta 1 valmistuu pidempiä koneenosia kuin linjastolta 2, $\mu_1 > \mu_2$.

3.1.2 Otosvarianssi

Tarkastellaan tilannetta, jossa tutkitaan normaalijakautunutta populaatiota Normal (μ, σ^2). Populaatiosta kerätään riippumattomat havainnot $X_1, \ldots, X_n$, joista lasketaan otoskeskiarvo $\bar{X}$ ja otosvarianssi S^2. Z-muunnettujen satunnaismuuttujien neliösummalle pätee

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n),$$

jossa $\chi^2(n)$ on parametrinen jakauma, jota kutsutaan **khiin neliö-jakaumaksi vapausastein** n. Khiin neliö-jakaumia eri vapausastein on piirrettynä Kuvaan 3.1.



Kuva 3.1: Khiin neliö-jakaumia eri vapausastein k.

Usein Z-muunnosta ei voi suoraan käyttää otosvarianssin jakauman hakemisessa, sillä yleisesti populaatiojakauman odotusarvo sekä varianssi ovat tuntemattomia. Voidaan kuitenkin näyttää, että

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1),$$

eli otoskeskiarvoa tuntemattoman odotusarvon sijasta käyttämällä ylläoleva neliösumma olisi khiin neliö-jakautunut, mutta vapausastein $n-1$. Otosvarianssi otosuureena on muotoa

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

joten tästä seuraa

$$\frac{(n-1)S^2}{\sigma^2} = \Theta(X_1, \ldots, X_n) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Määritelmä 30 (Khiin neliö-jakauma)**.** Olkoon $Z_1, \ldots, Z_k$ ovat standardinormaalijakautuneita riippumattomia satunnaismuuttujia. Näiden neliösummalla on khiin neliö-jakauma ($\chi^2(k)$-jakauma) vapausastein k

$$Q = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2(k),$$

jossa $\chi^2(k)$-jakauman tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k-2}{2}} \exp\left(-\frac{x}{2}\right), & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Khiin neliö-jakauman tiheysfunktiosta löytyy usein todennäköisyysjakaumissa esiintyvä gamma-funktio

$$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x) \mathrm{d}x,$$

joka on tietyssä mielessä kertoman $n!$ yleistys kokonaislukujen ulkopuolelle ominaisuuksiensa $\Gamma(n+1) = n\Gamma(n)$ ja $\Gamma(1) = 1$ perusteella. Ominaisuuksista seuraa $\Gamma(n+1) = n!$, kun n on kokonaisluku.

Nyt on tärkeää huomata, että otosvarianssille ei päde keskeisen raja-arvolauseen kaltaista tulosta kuten otoskeskiarvolle. Jotta neliösummilla olisi khiin neliö-jakauma, pitää populaation siis olla normaalijakautunut.

Esimerkki 27. Tarkastellaan erään opiskelijapopulaation älykkyysosamäärien jakaumaa. Oletetaan, että älykkyysosamäärien jakauma on normaalijakauma odotusarvolla μ ja varianssilla $\sigma^2 = 10$. Erään testin avulla mitataan 11 opiskelijan älykkyysosamäärä, ja lasketaan otosvarianssiksi $s^2 = 14.6$. Nyt tiedetään, että

$$\frac{(11-1)S^2}{10} = S^2 \sim \chi^2(10).$$

Nyt voidaan laskea todennäköisyys, että otosvarianssiksi saataisiin otoksen perusteella vähintään havaittu $s^2 = 14.6$

$$\mathsf{P}\left(S^2 \geq 14.6\right) = 0.1473.$$

Nyt todennäköisyys on laskettu R:llä. Khiin neliö-jakauman yläkvantiilitaulukosta voitaisiin lukea todennäköisyydelle rajat $0.1 < \mathsf{P}\left(S^2 \geq 14.6\right) < 0.95$.

3.1.3 t-testisuure

Tarkastellaan tilannetta, jossa tutkitaan normaalijakautunutta populaatiota $\mathsf{Normal}\left(\mu, \sigma^2\right)$. Popuulatiosta kerätään havaintoaineisto $X_1, \ldots, X_n$, josta lasketaan otoskeskiarvo $\bar{X}$ ja otosvarianssi S^2, joihin liittyy todennäköisyysjakaumat

$$\bar{X} \sim \mathsf{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Voidaan osoittaa, että otoskeskiarvo $\bar{X}$ ja otosvarianssi S^2 ovat toisistaan riippumattomat satunnaismuuttujat, siitä huolimatta että otossuureet lasketaan samoista havainnoista. Tällöin otossuureella

$$T = \Theta(X_1, \ldots, X_n) = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

on Studentin t-jakauma vapausastein $n-1$. Otossuuretta kutsutaan t-testisuureeksi. Studentin t-jakaumia eri vapausasteilla on piirretty Kuvaan 3.2. Kun Studentin t-jakauman vapausasteet kasvavat, lähestyy t-jakauman muoto normaalijakauman muotoa, mutta etenkin pienillä vapausasteilla t-jakaumalla on paksut hännät verrattuna normaalijakaumaan.

Määritelmä 31 (t-jakauma)**.** Olkoon X_1 ja X_2 toisistaan riippumattomia satunnaismuuttujia

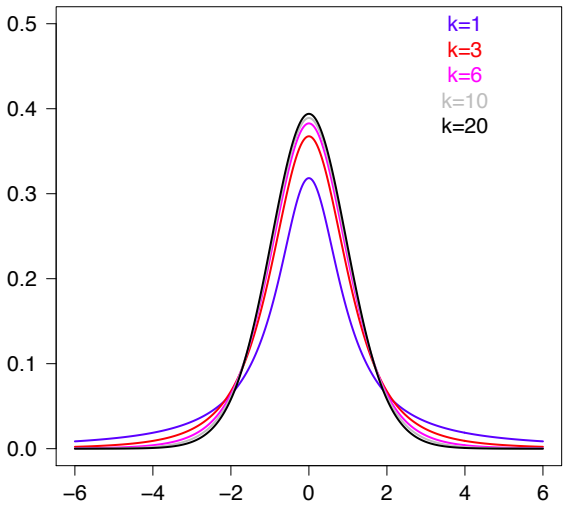
$$X_1 \sim \mathsf{Normal}\,(0,1)$$
$$X_2 \sim \chi^2(k).$$

Tällöin satunnaismuuttujalla T on Studentin t-jakauma vapausastein k, $T \sim \mathrm{t}(k)$ kun

$$T = \frac{X_1}{\sqrt{X_2/k}}.$$

Jakauman $\mathrm{t}(k)$ tiheysfunktio on

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{1}{k}x^2\right)^{-\frac{k+1}{2}}.$$



Kuva 3.2: Studentin t-jakaumia eri vapausastein k.

Nyt on tärkeää huomata, että t-testisuureelle ei päde keskeisen raja-arvolauseen kaltaista tulosta kuten otoskeskiarvolle. Jotta t-testisuure olisi Studentin t-jakautunut, pitää populaation siis olla normaalijakautunut.

Esimerkki 28. Tarkastellaan erään opiskelijapopulaation älykkyysosamäärien jakaumaa. Oletetaan, että älykkyysosamäärien jakauma on normaalijakauma odotusarvolla μ ja varianssilla σ^2. Erään testin avulla mitataan 11 opiskelijan älykkyysosamäärä, ja lasketaan otoskeskiarvoksi 99.7 ja otosvarianssiksi $s^2 = 14.6$. Nyt tiedetään, että

$$T = \frac{99.7 - \mu}{\sqrt{14.6}/\sqrt{11}} \sim \mathsf{t}(10).$$

Jos oletetaan, että opiskelijapopulaatiossa älykkyysosamäärän odotusarvo olisi 102, niin todennäköisyys, että otoskeskiarvo poikkeasi yli 2.3 yksikköä todellisesta odotusarvosta olisi

$$\mathsf{P}\left(\left|\bar{X} - 102\right| > 2.3\right) = 2\mathsf{P}\left(\frac{\bar{X} - 102}{\sqrt{14.6}/\sqrt{11}} > \frac{2.3}{\sqrt{14.6}/\sqrt{11}}\right) = 2\mathsf{P}\left(T > \frac{2.3}{\sqrt{14.6}/\sqrt{11}}\right) = 0.03691471 \approx 0.04.$$

Nyt todennäköisyys on laskettu R:llä. Studentin $\mathsf{t}(k)$-jakauman yläkvantiilitaulukosta todennäköisyydelle voitaisiin lukea rajat $2 \cdot 0.025 < \mathsf{P}\left(\left|\bar{X} - 102\right| > 2.3\right) < 2 \cdot 0.05$, sillä $\frac{2.3}{\sqrt{14.6}/\sqrt{11}} = 1.996401$.

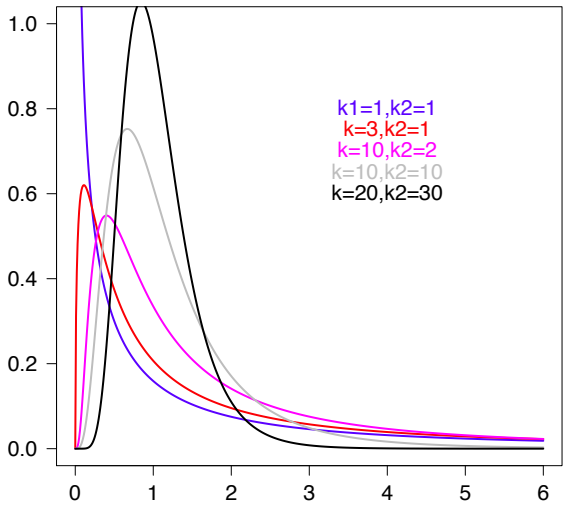
3.1.4 F-testisuure

Tarkastellaan tilannetta, jossa tutkitaan kahta normaalijakautunutta populaatiota, $\mathsf{Normal}\left(\mu_X, \sigma_X^2\right)$ ja $\mathsf{Normal}\left(\mu_Y, \sigma_Y^2\right)$. Populaatioista kerätään havaintoaineistot $X_1, \ldots, X_{n_X}$ ja $Y_1, \ldots, Y_{n_Y}$, joista lasketaan otoskeskiarvot $\bar{X}$, $\bar{Y}$ sekä edelleen otosvarianssit S_X^2, S_Y^2. Otosvariansseihin liittyvät siis jakaumat

$$\frac{(n_X - 1)S_X^2}{\sigma_X^2} \sim \chi^2(n_X - 1)$$
$$\frac{(n_Y - 1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n_Y - 1).$$

Nyt voidaan osoittaa, että otossuureen

$$F = \Theta(X_1, \ldots, X_{n_X}, Y_1, \ldots, Y_{n_Y}) = \frac{\frac{(n_X-1)S_X^2}{\sigma_X^2}/(n_X - 1)}{\frac{(n_Y-1)S_Y^2}{\sigma_Y^2}/(n_Y - 1)} = \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} = \frac{\sigma_Y^2}{\sigma_X^2}\frac{S_X^2}{S_Y^2}$$



Kuva 3.3: F-jakaumia eri vapausastein k_1 ja k_2.

jakauma on parametrista muotoa oleva F-jakauma vapausastein $n_X - 1$ ja $n_Y - 1$. Otossuuretta kutsutaan F-testisuureeksi. F-jakaumia eri vapausastein on piirretty Kuvaan 3.3 Nyt on tärkeää huomata, että F-testisuureelle ei päde keskeisen raja-arvolauseen kaltaista tulosta kuten otoskeskiarvolle. Jotta F-testisuure olisi F-jakautunut, pitää populaation siis olla normaalijakautunut. Jos populaatioiden keskihajonnat ovat samat $\sigma_X^2 = \sigma_Y^2$, niin F-testisuure supistuu muotoon

$$F = \frac{\sigma_Y^2}{\sigma_X^2}\frac{S_X^2}{S_Y^2} = \frac{S_X^2}{S_Y^2}.$$

Määritelmä 32 (F-jakauma)**.** Olkoon X_1 ja X_2 toisistaan riippumattomia satunnaismuuttujia

$$X_1 \sim \chi^2(k_1)$$
$$X_2 \sim \chi^2(k_2).$$

Tällöin satunnaismuuttujalla F on F-jakauma vapausastein k_1 ja k_2, $F \sim \mathsf{F}(k_1, k_2)$ kun

$$F = \frac{X_1/k_1}{X_2/k_2}.$$

Jakauman $\mathsf{F}(k_1, k_2)$ tiheysfunktio on

$$f(x) = \begin{cases} \frac{1}{B\left(\frac{k_1}{2}, \frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1} \left(1 + \frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

F-jakauman määrittelyssä esiintyy beta-funktio

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Esimerkki 29. Tarkastellaan kahdelta linjastolta valmistuvien samanlaisten koneenosien pituuksia. Kerätään molemmilta linjastoilta 10 koneenosan kokoinen otos ja havaitaan otosvarianssit 0.032 ja 0.010. Oletetaan, että varianssit ovat samat. Nyt voidaan laskea todennäköisyys

$$\mathsf{P}\left(\frac{S_X^2}{S_Y^2} \geq \frac{s_X^2}{s_Y^2}\right) = \mathsf{P}\left(F \geq \frac{0.032}{0.010}\right) = \mathsf{P}\left(F \geq 3.2\right),$$

jossa $F \sim \mathsf{F}(9, 9)$. R:llä voidaan laskea $\mathsf{P}\left(F \geq 3.2\right) \approx 0.04908$. $\mathsf{F}(k_1, k_2)$-jakauman yläkvantiilitaulukosta voitaisiin arvioida todennäköisyys $\mathsf{P}\left(F \geq 3.2\right) \approx 0.05$.

3.2 Estimointi

Piste-estimointi

Populaatiojakaumat ovat mallinnettu usein parametristen todennäköisyysjakaumien avulla. Tavallista on, että populaation jakaumaperhe on kiinnitetty (normaalijakauma, binomijakauma, Poisson-jakauma, ...) , mutta jakauman parametreista yksi tai useampi on tuntematon. Piste-estimoinnissa pyritään etsimään populaation tuntemattomille parametreille jossakin mielessä parhaat mahdolliset likimääräiset luvut. Otossuuretta, jonka avulla estimoidaan parametrien arvoa, kutsutaan funktiona **estimaattoriksi**, johon havaintojen arvot sijoittamalla saadaan parametrin **estimaatti**. Estimaattoreiden toivottavia ominasuuksia ovat mm. harhattomuus sekä tarkentuvuus. Parametrin θ estimoimiseksi käytettävän estimaattorin $\Theta(X_1, \ldots, X_n)$ harhattomuus tarkoittaa, että

$$\mathsf{E}\left(\Theta(X_1, \ldots, X_n)\right) = \theta.$$

Vastaavasti estimaattorin tarkentuvuus tarkoittaa, että otoskoon n kasvaessa estimaattorin arvo $\hat{\theta} = \Theta(x_1, \ldots, x_n)$ lähestyy oikeaa parametrin arvoa θ.

Väliestimointi

Väliestimointi on tapa kuvata tilastolliseen päättelyyn liittyvää epävarmuutta estimoimalla otoksesta väli, joka sisältäisi oikean parametrin θ arvon on tietyn suuruisella luottamuksella. Luottamusväliestimaatin pituudesta voidaan vetää päätelmiä piste-estimoinnin tarkuudesta, ja sen perusteella voidaan tarkastella erilaisten estimointia koskevien väitteiden luotettavuutta. Luottamusväli voidaan hakea yksipuolisena (a, Θ_U), (Θ_L, b) tai kaksipuolisena estimaattorina (Θ_L, Θ_U). Kaksipuolisessa väliestimoinnissa estimoidaan alempi päätepiste θ_L sekä ylempi päätepiste θ_U, jotka muodostavat välin (θ_L, θ_U). Usein luottamusvälistä puhuttaessa kuitenkin tarkoitetaan kaksipuolista luottamusväliä. Yksipuolisessa väliestimoinnissa haetaan vain toinen päätepiste, ja toinen on itsestään selvä otossuureen jakauman perusteella. Yksipuolisia välejä ovat esimerkiksi (θ_L, ∞) tai $(-\infty, \theta_U)$.

3.2.1 Populaation odotusarvon estimointi

Otoskeskiarvo sekä otosvarianssi ovat populaation odotusarvon sekä varianssin harhattomia ja tarkentuvia estimaattoreita.

$$\begin{aligned}
\mathsf{E}\left(\bar{X}\right) &= \mu \\
\mathsf{E}\left(S^2\right) &= \mathsf{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu - (\bar{X} - \mu))^2\right) \\
&= \frac{1}{n-1}\mathsf{E}\left(\sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}2(X_i - \mu)(\bar{X} - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2\right) \\
&= \frac{1}{n-1}\mathsf{E}\left(\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)n(\bar{X} - \mu) + n(\bar{X} - \mu)^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathsf{E}\left((X_i - \mu)^2\right) - \frac{n}{n-1}\mathsf{E}\left((\bar{X} - \mu)^2\right) \\
&= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} \\
&= \sigma^2
\end{aligned}$$

Estimaattorina käytettävän otossuureen keskivirhe on luonnollinen estimaattorin hyvyyden mitta. Jos estimaattori on harhaton, niin mitä pienempi varianssi sillä on, niin vähemmän estimaatit vaihtelevat otoksesta toiseensa oikean parametrin arvon ympärillä. Esimerkiksi odotusarvon keskivirhe on siis

$$\mathsf{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

joka riippuu populaatiovarianssista. Kun populaation varianssi ei ole tunnettu, keskivirheeksi arvioidaan

$$\mathsf{SE}(\bar{X}) \approx \frac{s}{\sqrt{n}}.$$

Usein estimoinnissa myös likimääräisesti laskettua keskivirhettä kutsutaan keskivirheeksi. On kuitenkin hyvä olla tietoinen millon käytetään likimääräistä virhe-estimaattia, sillä erilaiset keskivirhettä hyödyntävät otossuureet eivät säilytä tilastollisia ominaisuuksiaan tarkasti, jos likimääräisiä arvoja käytetään. Joskus olisi hyvä käyttää eri merkintää likimääräiselle keskivirheelle, kuten $\widehat{\mathsf{SE}(\bar{X})}$, mutta notaation yksinkertaistuksen vuoksi tässä tekstissä ei noudateta tätä käytäntöä.

Esimerkki 30. Tiedetään, että erään lintulajin munien paino on normaalijakautunut tuntemattomilla parametreille μ ja σ^2. Kerätään havaintoaineisto punnitsemalla 10 munan painot. Saadaan havainnot

$$11 \quad 11 \quad 12 \quad 13 \quad 13 \quad 13 \quad 16 \quad 16 \quad 18 \quad 19$$

Painojen otoskeskiarvo ja otosvarianssi ovat

$$\bar{x} = 14.2$$
$$s^2 = 8.177778 \approx 8.2.$$

Otoskeskiarvon keskivirheeksi arvioidaan

$$\mathsf{SE}(\bar{X}) = \frac{\sqrt{8.177778}}{\sqrt{10}} = 0.9043107 \approx 0.9.$$

Populaation odotusarvon luottamusväli, kun populaation varianssi on tunnettu

Populaation odotusarvon μ harhaton piste-estimaattori on n havainnon otoskeskiarvo $\bar{X}$, jolla on (ainakin likimääräisesti) jakauma

$$\bar{X} \dot{\sim} \mathsf{Normal}\left(\mu, \frac{\sigma^2}{n}\right).$$

Suurilla otoksilla tulos pätee Keskeisen raja-arvolauseen perusteella. Oletetaan jakauman varianssi σ^2 tunnetuksi. Nyt Z-muunnetulla satunnaismuuttujalla

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

on standardinormaalijakauma $\mathsf{Normal}\,(0, 1)$.

Luottamusvälin muodostamiseksi valitaan luottamusvälin peittotodennäköisyys $1 - \alpha$. Sanotaan että estimoidaan yksipuolista tai kaksipuolista $100(1-\alpha)\%$-luottamusväliä, ($100(1-\alpha)\%$-confidence interval, $100(1-\alpha)\%$-CI). Usein luottamusväli ilman muita määreitä tarkoittaa kaksipuolista luottamusväliä.

Tarkastellaan ensin yksipuolista väliestimointia. Haetaan alaraja yksipuoliselle $100(1-\alpha)\%$-luottamusvälille. Olkoon z_α standardinormaalijakauman $(1-\alpha)$-kvantiili, eli $\mathsf{P}\,(Z \geq z_\alpha) = \alpha$ Tällöin pätee

$$1 - \alpha = \mathsf{P}\,(Z < z_\alpha) = \mathsf{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right)$$
$$= \mathsf{P}\left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu\right).$$

Tämän perusteella $100(1-\alpha)\%$-CI estimaattoriksi saadaan (Θ_L, ∞), jossa alaraja on

$$\Theta_L = \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Jos havaittujen arvojen pohjalta otoskeskiarvon arvoksi saadaan $\bar{x}$, niin luottamusväli on siis muotoa (θ_L, ∞), jossa

$$\theta_L = \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Populaatiojakauman odotusarvo on tuntematon vakio, joten μ joko sisältyy tai ei sisälly väliin (θ_L, ∞). Laskettu $100(1-\alpha)\%$-CI tarkoittaa, että sadalla eri havaintoaineistolla samasta populaatiojakaumasta lasketut luottamusvälit sisältävät todellisen odotusarvon noin $100(1-\alpha)$ kertaa. Kun luottamusväli lasketaan yhdellä tietyllä havaintoaineistolla, niin se joko sisältää tai ei sisällä todellista parametrin arvoa. Yksipuolisille luottamusväleille saadaan vastaavasti ylärajaestimaattori

$$\Theta_U = \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Esimerkki 31. Tarkastellaan tietyn maalin kuivumisaikoja $n = 20$ testin avulla. Aiempien testien perusteella tiedetään hyvin tarkasti, että kuivumisaikojen hajonta on 1 h. Otoskeskiarvoksi lasketaan $\bar{x} = 8$ h. Haetaan yksipuoliselle 95%-luottamusvälille yläraja. Tämän perusteella maalin tuoteselosteessa voidaan ilmoittaa kuinka kauan maalin täytyy antaa kuivua vähintään. Nyt 0.95-kvantiili on $z_{0.05} = 1.64$, joten yläraja yksipuoliselle luottamusvälille on

$$\theta_U = 8 + 1.64 \frac{1}{\sqrt{20}} = 8.366715 \approx 8.4.$$

Tarkastellaan kaksipuolista $100(1-\alpha)\%$-luottamusväliä, joka haetaan symmetrisesti siten, että välin ulkopuolelle jää molempiin häntiin $\alpha/2$ verran todennäköisyyttä. Valitaan siis symmetriaan pohjautuen kvantiilit $-z_{\alpha/2}$ ja $z_{\alpha/2}$, joille

$$\mathsf{P}\left(Z > \alpha/2\right) = \alpha/2 = \mathsf{P}\left(Z < -z_{\alpha/2}\right),$$

jossa $Z \sim \mathsf{Normal}\,(0,1)$. Kvantiilit voidaan hakea normaalijakauman kertymäfunktion taulukosta. Tällöin saadaan

$$\begin{aligned} 1-\alpha &= \mathsf{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\ &= \mathsf{P}\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/n} < z_{\alpha/2}\right) \\ &= \mathsf{P}\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \end{aligned}$$

Tässä tapauksessa $100(1-\alpha)\%$-luottamusväli on siis (Θ_L, Θ_U), jossa

$$\begin{aligned} \Theta_L &= \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \Theta_U &= \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Jos realisoitunut otoskeskiarvo on $\bar{x}$, niin luottamusväliksi realisoituu (θ_L, θ_U), jossa siis päätepisteet ovat

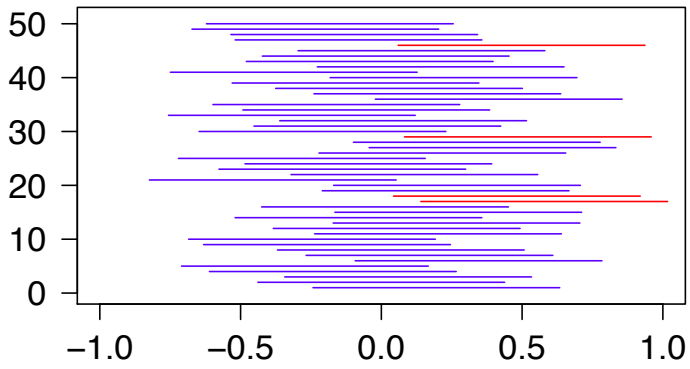
$$\begin{aligned} \theta_L &= \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \theta_U &= \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Samoin kuin edellä, nyt laskettu luottamusväli tarkoittaa, että jos koetta toistetaan 100 kertaa ja lasketaan $100(1-\alpha)\%$-luottamusvälit, niin noin $100(1-\alpha)$ tapauksessa realisoitunut luottamusväli sisältää todellisen odotusarvon μ. Kuvassa 3.4 on piirretty viisikymmentä 95%-luottamusväliä, jotka ovat laskettu standardinormaalipopulaatiosta 20 havainnon perusteella, olettaen populaatiovarianssin tunnetuksi.

Luottamusvälin rajat voidaan kirjoittaa myös

$$100(1-\alpha)\% - \text{CI}: \ \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

joka on muotoa estimaatti $\pm$ $\alpha/2$-yläkvantiili $\times$ estimaattorin keskivirhe.



Kuva 3.4: Viisikymmentä 95%-luottamusväliä, jotka ovat laskettu 20 standardinormaalijakautuneen havainnon avulla. Punaisella piirretyt luottamusvälit eivät sisällä populaation odotusarvoa $\mu = 0$.

Esimerkki 32. Mehuautomaatti on säädetty siten, että jaettava juomamäärä vaihtelee keskihajonnalla $\sigma = 0.2$ cl, ja voidaan olettaa, että juomamäärät noudattavat normaalijakaumaa. Satunnaisesti valitussa 36 näytteen otoksessa keskimääräinen sisälly oli 20.3 cl. Nyt 95%-luottamusvälin ala- ja yläraja juomamäärien odotusarvolle on

$$\theta_L = 20.3 - 1.960\frac{0.2}{\sqrt{36}} = 20.23467 \approx 20.23$$
$$\theta_U = 20.3 + 1.960\frac{0.2}{\sqrt{36}} = 20.36533 \approx 20.37,$$

jossa ollaan käytetty standardinormaalijakauman yläkvantiilia $z_{0.025} = 1.960$.

Populaation odotusarvon luottamusväli, kun populaation varianssi ei ole tunnettu

Oletetaan, että populaatiojakauma on normaalijakauma, mutta populaatiovarianssia σ^2 ei tunneta. Tällöin varianssi estimoidaan n-kokoisesta otoksesta otosvarianssin avulla. Luottamusvälin hakemisessa voidaan käyttää hyödyksi aiemmin esitettyä t-testisuuretta, joka lasketaan otoksen perusteella ja noudattaa Studentin t-jakaumaa vapausastein $n - 1$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t[n-1].$$

Koska Studentin t-jakauma on symmetrinen, valitaan jälleen kvantiilit $-t_{\alpha/2}^{(n-1)}$ ja $t_{\alpha/2}^{(n-1)}$, jossa $t_{\alpha/2}^{(n-1)}$ on yläkvantiili siten, että

$$\mathsf{P}\left(T > t_{\alpha/2}^{(n-1)}\right) = \frac{\alpha}{2} = \mathsf{P}\left(T < -t_{\alpha/2}^{(n-1)}\right),$$

jossa $T \sim \mathsf{t}(n-1)$. Kvantiilit voidaan hakea t-jakauman yläkvantiilitaulukosta. Samoin kuin edellä, voidaan johtaa

$$1 - \alpha = \mathsf{P}\left(\bar{X} - t_{\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}}\right),$$

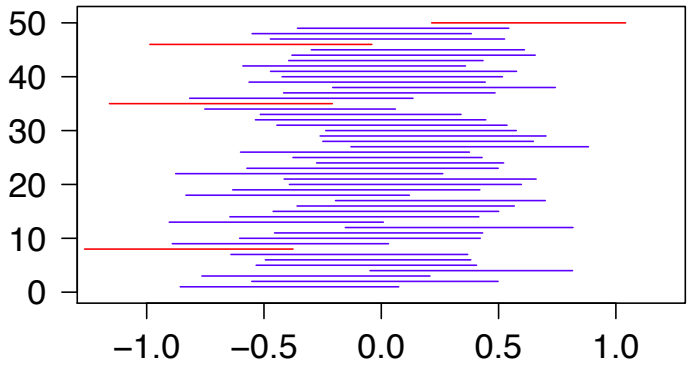
jolloin luottamusväli-estimaattori on (Θ_L, Θ_U), jossa

$$\Theta_L = \bar{X} - t_{\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}}$$
$$\Theta_U = \bar{X} + t_{\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}}.$$

Luottamusvälin rajat voidaan kirjoittaa myös muodossa

$$100(1-\alpha)\% - \text{CI}: \ \bar{x} \pm t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}},$$

joka on taas muotoa estimaatti $\pm$ $\alpha/2$-yläkvantiili $\times$ estimaattorin keskivirhe. Kuvassa 3.5 on piirretty viisikymmentä 95%-luottamusväliä, jotka ovat laskettu standardinormaalipopulaatiosta 20 havainnon perusteella. Huomaa, että toisin kuin aiemmin, kun hajontaa ei tunneta ja käytetään otosvarianssia, niin luottamusvälin pituus riippuu havaintoaineistoista.



Kuva 3.5: Viisikymmentä 95%-luottamusväliä, jotka ovat laskettu 20 standardinormaalijakautuneen havainnon avulla. Punaisella piirretyt luottamusvälit eivät sisällä populaation odotusarvoa $\mu = 0$.

Esimerkki 33. Tutkitaan tupakoivien äitien lasten syntymäpainoja. Oletetaan aiemman kerätyn tiedon perusteella, että syntymäpainot ovat normaalijakautuneita. Kerätään 14 lapsen syntymäpaino x_i kilogrammoissa, josta lasketaan otoskeskiarvo $\bar{x} = 3.37$, sekä otosvarianssi $s^2 = 0.15$. Haetaan 95%-luottamusvälin rajat syntymäpainojen populaation odotusarvolle

$$\theta_L = 3.37 - 2.160\frac{\sqrt{0.15}}{\sqrt{14}} = 3.146419 \approx 3.15$$
$$\theta_U = 3.37 + 2.160\frac{\sqrt{0.15}}{\sqrt{14}} = 3.593581 \approx 3.59,$$

jossa Studentin t-jakauman yläkvantiili on $t_{0.025}^{(13)} = 2.160$.

3.2.2 Parittaisten havaintojen erotuksen estimointi

Tarkastellaan tilannetta, jossa vertaillaan jossakin mielessä parittaisia tilastollisia muuttujia. Jokaiseen havaintoon X_i liittyy siis toinen havainto Y_i. Kerätään havainnot $X_1, \ldots, X_n$ sekä $Y_1, \ldots, Y_n$. Jossa X_i ja Y_i voivat olla tilastoyksikön tilastollisen muuttujan arvo ennen käsittelyä sekä käsittelyn jälkeen. Nyt ei oleta, että havainnot $X_1, \ldots, X_n$ tai $Y_1, \ldots, Y_n$ olisivat normaalijakautuneet tai edes samoin jakautuneet. Sen sijaan mallinnetaan toisistaan riippumattomia erotuksia $D_i = X_i - Y_i$ normaalijakautuneina $D_i \sim \text{Normal}\left(\mu, \sigma^2\right)$. Piste-estimaattori erotusjakauman odotusarvolle on otoskeskiarvo

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i,$$

jonka keskivirheeksi arvioidaan

$$\mathsf{SE}(\bar{D}) = \frac{s}{\sqrt{n}},$$

jossa s on erotusten otoskeskihajonta.

Luottamusväli parittaisille havainnoille, kun erotusta voidaan mallintaa normaalijakaumalla, voidaan estimoida samoin kuten aiemmin populaation odotusarvon luottamusväli.

Esimerkki 34. Tarkastellaan erään dieetin vaikutusta 10 koehenkilön avulla. Jokainen koehenkilö punnitaan sekä ennen että jälkeen dieettijakson. Saadaan havainnot (kilogrammoissa)

Paino ennen:	x_i	94	92	103	100	97	99	87	78	77	96
Paino jälkeen:	y_i	93	90	95	96	90	100	82	80	73	92
Erotus:	d_i	1	2	8	4	7	-1	5	-2	4	4

Erotusten otoskeskiarvo ja otosvarianssi ovat

$$\bar{d} = 3.2$$
$$s^2 = 10.4$$

Keskivirheeksi arvioidaan

$$\mathsf{SE}(\bar{D}) = \sqrt{\frac{10.4}{10}} = 1.019804 \approx 1.02$$

Nyt 95%-luottamusväliksi (θ_L, θ_U) erotusten odotusarvolle voidaan estimoida

$$\theta_L = 3.2 - 2.262\sqrt{\frac{10.4}{10}} \approx 0.893$$
$$\theta_L = 3.2 + 2.262\sqrt{\frac{10.4}{10}} \approx 5.507,$$

jossa $t^{(9)}_{0.025} = 2.262$.

3.2.3 Kahden populaation odotusarvojen erotuksen estimointi

Kun halutaan vertailla kahden populaation keskimääräisiä arvoja, estimoidaan usein niiden odotusarvojen erotusta. Käydään läpi neljä erilaista tilannetta, jotka tulevat esiin odotusarvojen vertailussa. Oletetaan, että on kerätty populaatioista riippumattomat havainnot $X_1, \ldots, X_{n_1}$ sekä $Y_1, \ldots, Y_{n_2}$. Tarkastellaan ensin odotusarvoja μ_1, μ_2 sekä erotusta $\mu_1 - \mu_2$. Estimaattorit ovat $\bar{X}, \bar{Y}$ sekä $\bar{X} - \bar{Y}$, joka on harhaton ja tarkentuva estimaattori erotukselle $\mu_1 - \mu_2$.

Ensin käydään läpi tilanne, jossa populaatioiden varianssit σ_1^2 ja σ_2^2 ovat tunnettuja. Riippumattomuuden nojalla

$$\mathsf{Var}\left(\bar{X} - \bar{Y}\right) = \mathsf{Var}\left(\bar{X}\right) + \mathsf{Var}\left(\bar{Y}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

joten erotuksen $\mu_1 - \mu_2$ estimaattorin keskivirhe on

$$\mathsf{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Jos tiedetään vielä lisäksi, että populaatioiden varianssit ovat yhtäsuuret $\sigma_1^2 = \sigma_2^2 = \sigma^2$, voidaan odotusarvojen erotuksen estimaattorin keskivirhe ilmaista muodossa

$$\mathsf{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Jos populaation varianssit eivät ole tunnettuja, täytyy meidän estimoida ne otoksen perusteella käyttäen otosvariansseja. Tarkastellaan ensin tilannetta, jossa varianssit ovat tuntemattomia, mutta yhtäsuuria $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Merkitään otoksien otosvariansseja S_1^2 sekä S_2^2. Nyt erotuksen varianssin estimaattorina käytetään **yhteisotosvarianssia** (pooled sample variance)

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

jolloin erotuksen estimaattorin keskivirhe on likimääräisesti

$$\mathsf{SE}(\bar{X} - \bar{Y}) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

jossa s_p on realisoitunut yhteisotosvarianssi.

Jos ei voida olettaa, että varianssit olisivat yhtäsuuria, erotuksen keskivirhe on likimääräisesti

$$\mathsf{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Huomaa, että vaikka ei voida olettaa, että varianssit ovat yhtäsuuria, se ei tarkoita etteivät ne voisi olla.

Esimerkki 35. Kymmenen 15-vuotiasta poikaa ja kahdeksan 15-vuotiasta tyttöä osallistui tiettyyn testiin. Testissä havaittiin poikien pistemäärien keskiarvoksi 10.5 ja tyttöjen 9.2. Keskihajonnat olivat vastaavasti 3.1 ja 2.9. Testin standardoinnista johtuen voidaan pistemäärien jakaumia pitää normaaleina. Piste-estimaatti poikien ja tyttöjen populaatioiden odotusarvojen erotukselle on

$$\bar{x} - \bar{y} = 10.5 - 9.2 = 1.3.$$

Oletetaan ensin, että populaatioiden hajontoja voidaan pitää yhtäsuurina. Tällöin yhteisotosvarianssi on

$$s_p^2 = \frac{(10 - 1)3.1^2 + (8 - 1)2.9^2}{10 + 8 - 2} = 9.085,$$

jonka avulla saadaan likimääräinen keskivirhe

$$\mathsf{SE}(\bar{X} - \bar{Y}) = \sqrt{9.085}\sqrt{\frac{1}{10} + \frac{1}{8}} = 1.429729 \approx 1.423.$$

Jos ei voida olettaa, että hajonnat ovat yhtäsuuria, keskivirhettä estimoidaan

$$\mathsf{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{3.1^2}{10} + \frac{2.9^2}{8}} = 1.418538 \approx 1.419.$$

Kahden populaation odotusarvojen erotuksen luottamusväli

Oletetaan, että populaatioista on kerätty riippumattomat havainnot $X_1, \ldots, X_{n_1}$ sekä $Y_1, \ldots, Y_{n_2}$. Tarkastellaan erotusta $\mu_1 - \mu_2$. Estimaattori on $\bar{X} - \bar{Y}$, joka on harhaton ja tarkentuva estimaattori erotukselle $\mu_1 - \mu_2$. Kun varianssit σ_1^2 ja σ_2^2 ovat tunnettuja, $100(1 - \alpha)\%$-luottamusvälien rajoiksi saadaan täsmälleen samoin kuin yllä

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

ja

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

kun $\sigma_1 = \sigma_2 = \sigma$.

Tilanne jossa varianssit ovat tuntemattomia on hankalampi. Oletetaan ensin, että populaatioiden tuntemattomat varianssit ovat yhtäsuuria $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Nyt erotuksen varianssin estimaattorina käytetään yhteisotosvarianssia

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Voidaan näyttää, että otossuureella

$$Q = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} = \frac{(n_1+n_2-2)S_p^2}{\sigma^2}$$

on khiin neliö-jakauma vapausastein n_1+n_2-2. Tästä seuraa, että otossuureella

$$T = \frac{Z}{\sqrt{Q/(n_1+n_2-2)}} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(\bar{X}-\bar{X})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

on Studentin t-jakauma vapausastein n_1+n_2-2. Samoin kuin yllä, saadaan nyt $100(1-\alpha)\%$-luottamusvälin rajoiksi

$$\bar{x}-\bar{y} \pm t_{\alpha/2}^{(n_1+n_2-2)} s_y \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Tarkastellaan viimeiseksi tilanne, kun ei voida olettaa tuntemattomien varianssien yhtäsuuruutta. Tässä tilanteessa ei pääse suoraan käsiksi mihinkään otossuureeseen, jolla olisi parametrista muotoa oleva todennäköisyysjakauma. Sen sijaan käytetään Welch-Satterthwaite-approksimaatiota, jonka mukaan otossuureella

$$T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}},$$

likimääräisesti Studentin t-jakauma vapausastein

$$v = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}.$$

Tämän likimääräisen t-jakauman vapausasteet eivät ole välttämättä kokonaislukuja, mutta t-jakauman vapausasteiden ei tarvitse olla kokonaislukuja. Tällöin $100(1-\alpha)\%$-luottamusvälin rajoiksi saadaan

$$\bar{x}-\bar{y} \pm t_{\alpha/2}^{(v)} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}.$$

Esimerkki 36. Tarkastellaan Esimerkin 35 poikien ja tyttöjen tietyn testin arvosanojen jakaumien odotusarvojen erotusten luottamusväliä. Testissä havaittiin poikien pistemäärien keskiarvoksi 10.5 ja tyttöjen 9.2. Keskihajonnat olivat vastaavasti 3.1 ja 2.9. Piste-estimaatiksi poikien ja tyttöjen populaatioiden odotusarvojen erotukselle laskettiin

$$\bar{x}-\bar{y} = 10.5 - 9.2 = 1.3.$$

Jos populaatioiden hajonnat voidaan olettaa yhtäsuuriksi, niin

$$\mathsf{SE}(\bar{X}-\bar{Y}) = 1.429729,$$

ja odotusarvojen erotuksen 95%-luottamusvälin rajoiksi saadaan

$$\theta_L = 1.3 - 2.120 \cdot 1.429729 = -1.731025 \approx -1.73$$
$$\theta_U = 1.3 + 2.120 \cdot 1.429729 = 4.331025 \approx 4.33,$$

jossa t-jakauman kvantiili $t_{0.025}^{(10+8-2)} = 2.120$.

Jos populaatioiden hajontoja ei voida olettaa yhtäsuuriksi, niin

$$\mathsf{SE}(\bar{X} - \bar{Y}) = 1.418538,$$

ja odotusarvojen erotuksen 95%-luottamusvälin rajoiksi saadaan

$$\theta_L = 1.3 - 2.125 \cdot 1.418538 = -1.714393 \approx -1.71$$
$$\theta_U = 1.3 + 2.125 \cdot 1.418538 = 4.314393 \approx 4.31,$$

jossa t-jakauman kvantiili $t_{0.025}^{15.54444} = 2.125$. T-jakauman vapausaste 15.54444 lasketaan kuten Welch-Satterthwaite-approksimaatiossa.

3.2.4 Suhteellisen osuuden estimointi

Oletetaan, että populaatio on Bernoulli-jakautunut tuntemattomalla onnistumistodennäköisyydellä p. Tarkastellaan siis tilannetta, jossa otoksen $X_1, \ldots, X_n$ perusteella yritetään estimoida parametria p. Yhtäpitävästi voidaan tutkia frekvenssiä $Y = X_1 + \ldots + X_n$, jossa $Y \sim \mathsf{Binomial}\,(n, p)$. Piste-estimaattoriksi parametrille p otetaan suhteellinen frekvenssi

$$\hat{P} = \Theta(X_1, \ldots, X_n) = \bar{X} = \frac{Y}{n},$$

joka on Bernoulli-jakautuneiden muuttujien otoskeskiarvo. Suhteellisen frekvenssin odotusarvo ja varianssi ovat

$$\mathsf{E}\left(\hat{P}\right) = \mathsf{E}\left(\frac{Y}{n}\right) = \frac{1}{n}\mathsf{E}\,(Y) = \frac{1}{n}np = p$$
$$\mathsf{Var}\left(\hat{P}\right) = \mathsf{Var}\left(\frac{Y}{n}\right) = \left(\frac{1}{n}\right)^2 \mathsf{Var}\,(Y) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Suhteellinen frekvenssi on p:n harhaton ja tarkentuva estimaattori. Keskeiseen raja-arvolauseeseen vedoten suhteellisen frekvenssin jakaumaa voidaan approksimoida normaalijakaumalla, kun n on suuri. Suhteellisen frekvenssin keskivirheeksi voidaan likimääräisesti ottaa

$$\mathsf{SE}(\hat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

jossa $\hat{p}$ on suhteellisen frekvenssin realisoitunut arvo.

Esimerkki 37. Tarkastellaan erästä edullisten mikroprosessoreiden tuotantolinjastoa. Linjastolta kerätään 1000 mikroprosessorin otos, josta 920 täyttää etukäteen asetetut laatuvaatimukset. Olkoon p laatuvaatimukset täyttävien mikroprosessoireiden suhteellinen osuus tuotanlinjastolta valmistettavista prosessoreista. Suhteellisen osuuden piste-estimaatti on suhteellinen frekvenssi

$$\hat{p} = \frac{920}{1000} = 0.92.$$

Suhteellisen frekvenssin keskivirhe on

$$\mathsf{SE}(\hat{P}) = \sqrt{\frac{p(1-p)}{1000}},$$

jota arvioidaan

$$\mathsf{SE}(\hat{P}) = \sqrt{\frac{0.92 \cdot 0.08}{1000}} = 0.008579044 \approx 0.009.$$

Suhteellisen osuuden luottamusväli

Keskeiseen raja-arvolauseeseen vedoten suhteellisen frekvenssin jakaumaa voidaan approksimoida normaalijakaumalla, kun n on suuri. Tällöin voidaan samoin kuin hakea suhteellisen osuuden $100(1-\alpha)\%$-luottamusväli ratkaisemalla epäyhtälö

$$1 - \alpha = \mathsf{P}\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right)$$

todellisen suhteellisen osuuden p suhteen. Havaintojen realisoitumisen jälkeen päätepisteet voidaan hakea ratkaisemalla toiseen asteen yhtälö

$$(\hat{p} - p)^2 = \frac{z_{\alpha/2}^2}{n} p(1-p).$$

Likimääräinen luottamusväli saadaan kuitenkin sitä arvoimalla

$$\hat{p} \pm z_{\alpha/2} \mathsf{SE}(\hat{P}),$$

jossa keskivirhe on

$$\mathsf{SE}(\hat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Tämä approksimaatio ei kuitenkin ole kovin hyvä, ja varsinkin tietokoneella tehtävissä laskuissa käytetään parempia estimaatteja.

Esimerkki 38. Tarkastellaan Esimerkin 37 mikroprosessorien tuotantolinjastoaineistoa. Havaintoainestosta ($n = 1000$) laskettu suhteellinen frekvenssi on

$$\hat{p} = \frac{920}{1000} = 0.92,$$

sekä likimääräinen suhteellisen frekvenssin keskivirhe on

$$\mathsf{SE}(\hat{P}) = \sqrt{\frac{0.92 \cdot 0.08}{1000}} = 0.008579044 \approx 0.009.$$

Nyt laadukkaiden prosessoreiden suhteelliselle osuudelle saadaan $100(1-\alpha)\%$-luottamusvälin rajat

$$\theta_L = 0.92 - 1.960 \cdot 0.008579044 = 0.9031851 \approx 0.903$$
$$\theta_U = 0.92 + 1.960 \cdot 0.008579044 = 0.9368149 \approx 0.937,$$

jossa standardinormaalijakauman 0.025-yläkvantiili on $z_{0.025} = 1.960$.

3.2.5 Kahden suhteellisen osuuden erotuksen estimointi

Tarkastellaan kahta Bernoulli-jakautunutta populaatiota, joilla on tuntemattomat parametrit p_1 ja p_2. Kun halutaan vertailla parametrien p_1 ja p_2 arvoja, estimoidaan parametrien erotusta $p_1 - p_2$, populaatioista kerättävien riippumattomien otosten $V_1, \ldots, V_{n_1}$, sekä $W_1, \ldots, W_{n_2}$ perusteella. Perustetaan estimointi jälleen binomijakautuneisiin frekvensseihin

$$Y_1 = \sum_{i=1}^{n_1} V_i \sim \mathsf{Binomial}\,(n_1, p_1)$$
$$Y_2 = \sum_{i=1}^{n_2} W_i \sim \mathsf{Binomial}\,(n_2, p_2)\,.$$

Suhteelliset frekvenssit $\hat{P}_1$ ja $\hat{P}_2$ ovat parametrien p_1 ja p_2 harhattomia ja tarkentuvia estimaattoreita. Erotuksen estimaattorina käytetään erotusta $\hat{P}_1 - \hat{P}_2$, jolla on odotusarvo ja varianssi

$$\mathsf{E}\left(\hat{P}_1 - \hat{P}_2\right) = p_1 - p_2$$
$$\mathsf{Var}\left(\hat{P}\right) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Suurilla otoksilla erotus noudattaa likimääräisesti normaalijakaumaa kyseisillä parametrien arvoilla. Suhteellisten frekvenssien erotuksen keskivirheeksi voidaan arvioida

$$\mathsf{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

jossa $\hat{p}_1$ ja $\hat{p}_2$ ovat suhteellisen frekvenssin realisoituneita arvoja.

Esimerkki 39. Vertaillaan 10- ja 12-vuotiaiden koululaisten sukupuolijakaumia. Ensimmäisessä populaatiossa, eli 740 10-vuotiaan koululaisen joukossa on 350 poikaa, ja vastaavasti toisessa populaatiossa, eli 690 12-vuotiaan koululaisen joukossa on 340 poikaa. Suhteellisten osuuksien eroa estimoidaan suhteellisten frekvenssien erolla

$$\hat{p}_1 - \hat{p}_2 = \frac{350}{740} - \frac{340}{690} = -0.01978065 \approx -0.0198.$$

Keskivirhe on likimääräisesti

$$\mathsf{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\frac{350}{740}\frac{390}{740}}{740} + \frac{\frac{340}{690}\frac{350}{690}}{690}} = 0.02644038 \approx 0.0264.$$

Kahden suhteellisen osuuden erotuksen luottamusväli

Suurilla otoksilla kahden suhteellisen osuuden erotus noudattaa likimääräisesti normaalijakaumaa odotusarvolla ja varianssilla $\mathsf{E}\left(\hat{P}_1 - \hat{P}_2\right)$ ja $\mathsf{Var}\left(\hat{P}\right)$, vastaavasti. Erotuksen tarkan luottamusvälin etsiminen on hankalaa, mutta likimääräinen luottamusväli saadaan kuitenkin sitä arvioimalla

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\mathsf{SE}(\hat{P}_1 - \hat{P}_2),$$

jossa likimääräinen keskivirhe on

$$\mathsf{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Esimerkki 40. Tarkastellaan Esimerkin 39 koululaisten sukupuolijakauma-aineistoa. Ensimmäisessä populaatiossa, eli 740 10-vuotiaan koululaisen joukossa on 350 poikaa, ja vastaavasti toisessa populaatiossa, eli 690 12-vuotiaan koululaisen joukossa on 340 poikaa. Suhteellisten osuuksien eroa estimoidaan suhteellisten frekvenssien erolla

$$\hat{p}_1 - \hat{p}_2 = \frac{350}{740} - \frac{340}{690} = -0.01978065 \approx -0.0198.$$

Suhteellisten frekvenssien keskivirhe on likimääräisesti

$$\mathsf{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\frac{350}{740}\frac{390}{740}}{740} + \frac{\frac{340}{690}\frac{350}{690}}{690}} = 0.02644038 \approx 0.0264.$$

Nyt populaatioiden suhteellisten osuuksien erotukselle 95%-luottamuvälin rajat ovat

$$\theta_L = -0.0198 - 1.960 \cdot 0.02644038 = -0.07162314 \approx -0.072$$
$$\theta_U = -0.0198 + 1.960 \cdot 0.02644038 = 0.03202314 \approx 0.032.$$

Luku 4

Hypoteesien testaus

Estimoinnin ohella toinen tilastollisen päättelyn tärkeä osa-alue on hypoteesien testaus. Piste-estimointi, väliestimointi ja hypoteesien testaus ovat toisiaan tukevia tilastollisen päättelyn muotoja ja niitä käytetään antamaan toisiaan täydentävää informaatiota populaatiosta. Tilastolliset hypoteesit koskevat jotain yhden tai useamman populaation ominaisuutta. Usein hypoteesi on väittämä populaatiojakaumien parametrien arvoista. Todellisuudessa populaatiojakaumalla joko on tai ei kyseinen ominaisuus, mutta koska päättely perustuu satunnaisotokseen , voidaan puhua korkeintaan hypoteesiin jollakin tavalla liittyvistä todennäköisyyksistä. Testauksessa pyritään tietenkin siihen, että virheellisen hypoteesia koskevan päätelmän tekeminen olisi epätodennäköistä.

Perinteinen hypoteesin testaus etenee siten, että parametreille asetetaan kaksi hypoteesia, joista ensimmäinen on **nollahypoteesi** ja toista kutsutaan **vaihtoehtoiseksi hypoteesiksi**. Nollahypoteesia merkitään tässä luentomateriaalissa symbolilla H_0, ja vaihtoehtoista hypoteesia (vastahypoteesia) symbolilla H_v. Itse testaus tapahtuu olettamalla, että **nollahypoteesi on tosi**. Tämän jälkeen arvioidaan voisiko havaintoaineisto olla nollahypoteesin mukaisesta mallista. Nyt tutkitaan siis, että voisiko havaittu ero havaintoaineiston ja nollahypoteesin välillä johtua sattumasta, vai voisiko ero olla todellista. Jos aineisto ei tue nollahypoteesia, niin analyysiä jatketaan vaihtoehtoisen hypoteesin kanssa. Jos aineiston perusteella nollahypoteesia **ei voida pitää epätodennäköisenä**, niin tilastollista analyysiä jatketaan nollahypoteesiin pohjaten. Hypoteesin testauksen jälkeen ei vieläkään siis tiedetä varmuudella kumpi hypoteeseista on oikea, mutta mahdollisesti analyysiä jatketaan ainakin malleista paremman kanssa.

Hypoteesin testauksen perusteella pyritään siis tekemään päätös, että voidaanko nollahypoteesi hylätä vai ei. Testauksen satunnaisuudesta johtuen eri päätöksiin liittyy siis jotkin todennäköisyydet. Nollahypoteesiin liittyvät päätökset sekä näihin liittyvät virheet ovat

	H_0 on oikea	H_0 on väärä
H_0:aa ei hylätä	Oikea päätös	Tyypin II virhe
H_0 hylätään	Tyypin I virhe	Oikea päätös.

Jos nollahypoteesi H_0 hylätään vaikka se olisi tosi, niin tehdään ns. tyypin I virheen. Vastaavasti jos nollahypoteesia H_0 ei hylätä vaikka se ei olisi tosi, niin tehtäisiin ns. tyypin II virheen. Muissa tapauksissa päätös on oikea, eli virhettä ei tehdä. Tyypin I virheen todennäköisyyttä α kutsutaan testin merkitsevyydeksi, riskiksi, tai riskitasoksi. Hypoteesin testauksen mahdollinen lähtökohta on valita suurin sallittu riski. Tyypin II virheen todennäköisyyttä, jota kutsutaan testin voimakkuudeksi, ei usein voida laskea, koska testauksen kohteena on vain nollahypoteesin sopivuutta malliin. Tällöin vastahypoteesin perusteella valitaan ainoastaan mitkä poikkeamat nollahypoteesista ovat merkityksellisiä, eli poikkevat tietyssä mielessä vastahypoteesin suuntaan.

Tässä kappaleessa tarkastellaan jakaumaperheiden parametriin θ liittyvää hypoteesin testausta. Nollahypoteesit ovat muotoa $\theta = \theta_0$, jonka mukaisesti parametrin arvo olisi jokin tietty θ_0. Tarkastellaan kolmen tyyppisiä hypoteesipareja, jotka ovat muotoa

$\mathsf{H}_0: \quad \theta = \theta_0$	$\mathsf{H}_0: \quad \theta = \theta_0$	$\mathsf{H}_0: \quad \theta = \theta_0$
$\mathsf{H}_v: \quad \theta < \theta_0$	$\mathsf{H}_v: \quad \theta > \theta_0$	$\mathsf{H}_v: \quad \theta \neq \theta_0.$

Kaksi vasemmanpuolista testiä ovat **toispuolisia testejä**, ja oikeanpuoleisin hypoteesipari vastaa **kaksipuolista testiä**.

Käytännössä tilastollisen testin tekeminen tapahtuu siten, että riippuen populaatiojakauma-oletuksesta ja testattavasta populaatiojakauman parametrista, valitaan ensiksi jokin otossuure, jota kutsutaan tilastollisessa testauksessa **testisuureeksi**. Yleisimmille testisuureille on olemassa tunnettua muotoa oleva todennäköisyysjakama, kun nollahypoteesi oletetaan todeksi. Kun testisuure on valittu, valitaan testin merkitsevyystaso α. Hyvin usein merkitsevyystasoksi valitaan (kovin mielivaltaisesti) esimerkiksi 0.05 tai 0.01. Testin merkitsevyystason valitsemisen jälkeen voidaan edetä kahdella vaihtoehtoisella tavalla.

Ensinnäkin merkitsevyystason α perusteella voitaisiin hakea ns. **kriittisen alueen** testisuureen mahdollisille arvoille vastahypoteesin perusteella. Kriittinen alue valitaan siten, että testisuureen realisoituminen kyseiselle alueelle antaisi evidenssiä vaihtoehtoisen hypoteesin puolesta nollahypoteesia vastaan. Kriittisen alueelle realisoitumisen todennäköisyys tulisi olla α. Seuraavaksi lasketaan **havaittu testisuureen arvo**, eli havaintojen arvot sijoitetaan testisuureen kaavaan. Jos nyt havaittu testisuureen arvo osuu kriittiselle alueelle, nollahypoteesin voidaan hylätä riskitasolla α. Jos arvo ei osu alueelle, niin nollahypoteesiä ei hylätä riskitasolla α.

Vaihtoehtoinen tapa on laskea havaittu testisuureen arvo, jonka perusteella lasketaan merkitsevyystestin **p-arvo**. Merkitsevyystestin p-arvo (havaittu merkitsevyystaso) on todennäköisyys, että esiintyy vähintään havaitun suuruinen poikkeama vastahypoteesin suuntaan, kun pidetään nollahypoteesiä on totena. Toisin sanottuna testin p-arvo on pienin α, jolla nollahypoteesi voidaan hylätä. **Nyt jos p-arvo on pienempi kuin valittu merkitsevyystaso α, niin nollahypoteesi hylätään riskitasolla α. Vastaavasti jos p-arvo on suurempi kuin valittu α, niin nollahypoteesia ei hylätä riskitasolla α.**

Tapana usein on, että käyttää kumpaa lähestymistapaa tahansa, niin testiin liittyvä p-arvo ilmoitetaan. Tämä siksi, että satunnaisuuden luonteen vuoksi usein ei ole kovinkaan paljon merkitystä onko laskettu p-arvo 0.04999 vai 0.05001. Kuitenkin toinen lasketuista p-arvoista johtaisi nollahypoteesin hylkäämiseen riskillä 0.05 ja toisella ei hylättäisi nollahypoteesia riskillä 0.05. P-arvon ilmoittamisella tutkimustuloksien luotettavuudesta voidaan vetää parempia johtopäätöksiä.

Tässä luentomonisteessa käsiteltävät testit

- Normaalijakautuneen populaation odotusarvon testaus yhden otoksen z- tai t-testillä
- Kahden normaalijakautuneen populaation odotusarvojen erotuksen testaus kahden riippumattoman otoksen t-testillä
- Populaation suhteellisen osuuden testaus z-testillä
- Kahden populaation suhteellisten osuuksien erotuksen testaus z-testillä
- Kahden normaalijakautuneen populaation varianssien yhtäsuuruuden testaus F-testillä
- Populaation mediaaniin testaus ilman normaalijakauma-oletusta Wilcoxonin merkityn järjestyksen testillä
- Kahden populaation sijaintien eron testaus ilman normaalijakauma-oletusta Mann-Whitneyn U-testillä
- Kahden diskreetin tai kategorisen muuttujan riippumattomuuttomuuden testaus χ^2-testillä
- Usean diskreetin tai kategorisen muuttujan jakaumien homogeenisuuden testaus χ^2-testillä

4.1 Z-testi

Odotusarvon testaus kun varianssi on tunnettu

Tarkastellaan tilannetta, jossa hypoteesit koskevat populaatiojakauman odotusarvoa. Populaatiojakauman odotusarvo ja keskihajonta ovat μ ja σ, jossa keskihajonta on tunnettu. Tässä tilanteessa odotusarvo testiä voidaan kutsua **z-testiksi**. Olkoon populaatiojakauma on normaalijakauma, tai otoskoko tarpeeksi iso, jotta otoskeskiarvon jakauma on hyvin tarkasti normaalijakauma $\bar{X} \sim \mathsf{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$. Odotusarvoa koskevat hypoteesiparit ovat muotoa

$\mathsf{H}_0: \quad \mu = \mu_0$	$\mathsf{H}_0: \quad \mu = \mu_0$	$\mathsf{H}_0: \quad \mu = \mu_0$
$\mathsf{H}_v: \quad \mu < \mu_0$	$\mathsf{H}_v: \quad \mu > \mu_0$	$\mathsf{H}_v: \quad \mu \neq \mu_0.$

Oletetaan, että nollahypoteesi on tosi. Tällöin $\mu = \mu_0$, ja otoskeskiarvolle voidaan tehdä Z-muunnos, jonka lopputulosta kutsutaan odotusarvon testauksessa **z-testisuureeksi**

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathsf{Normal}\,(0, 1)\,.$$

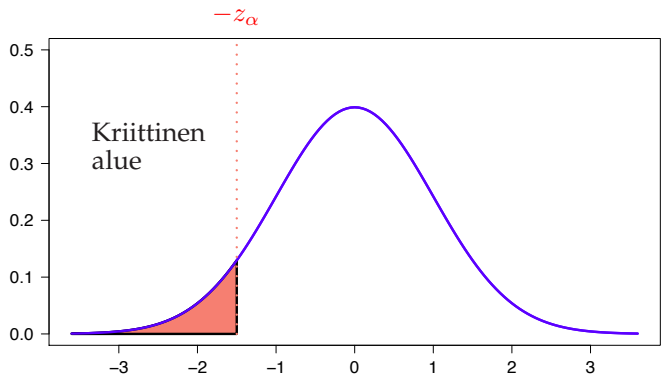
Valitaan testin merkitsevyystasoksi α.

Tarkastellaan hypoteesiparia $\mathsf{H}_0 : \mu = \mu_0, \mathsf{H}_v : \mu < \mu_0$. Kriittinen alue haetaan siten, että testisuureen realisoituminen kyseiselle alueelle antaisi evidenssiä nollahypoteesiä vastaan vastahypoteesin hyväksi. Koska nyt vastahypoteesi on muotoa $\mu < \mu_0$, niin mahdollisimman pienen testisuureen arvon havaitseminen antaisi evidenssiä vastahypoteesin suuntaan ja nollahypoteesiä vastaan. Tarkastellaan tilannetta nollahypoteesin voimassa ollessa testisuureen jakauman kautta. Kriittinen alue, joka sisältäisi α-verran todennäköisyyttä haetaan nyt standardinormaalijakauman yläkvantiilin z_α avulla. Koska $\mathsf{P}\,(Z \geq z_\alpha) = \alpha$, niin symmetrian vuoksi $\mathsf{P}\,(Z \leq -z_\alpha) = \alpha$. Kuvaan 4.1 on piirretty testisuureen jakauma, sekä kriittinen alue, johon kuuluu α-verran todennäköisyyttä. Seuraavaksi lasketaan havaitun testisuureen arvo z_{hav}. Jos z_{hav} osuu kriittiselle alueelle, niin hylätään nollahypoteesin riskillä α.

Vastaavasti voidaan havaitun testisuureen z_{hav} laskemisen jälkeen laskea todennäköisyyden, että havaittaisiin vähintään z_{hav} suuruinen poikkeama vastahypoteesin suuntaan. Tämä on merkitsevyystestin p-arvo, joka nyt käsitellyssä testissä on

$$\mathsf{p-arvo} = \Phi(z_{\text{hav}}),$$

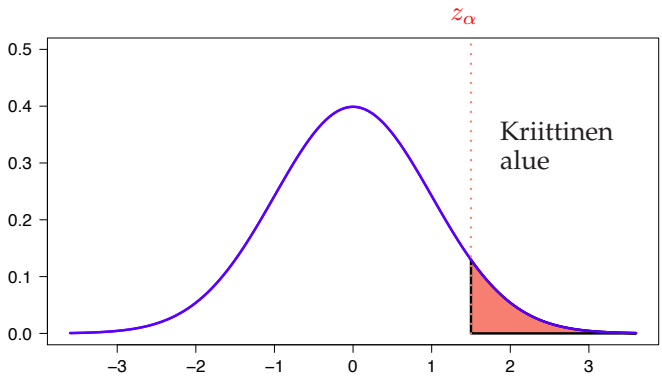
jossa $\Phi(\cdot)$ on standardinormaalijakauman kertymäfunktio.



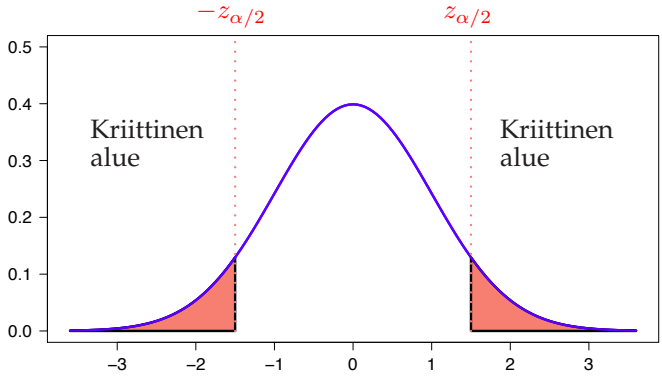
Kuva 4.1: z-testisuureen jakauma ehdolla, että nollahypoteesi $\mu = \mu_0$ on voimassa. Vastahypoteesina on $\mu < \mu_0$. Jos havaittu testisuureen arvo osuu kriittiselle alueelle, joka vastaa vasemmanpuolista häntätodennäköisyyttä, niin nollahypoteesi hylätään riskillä α.

Hypoteesiparin $\mathsf{H}_0 : \mu = \mu_0, \mathsf{H}_v : \mu > \mu_0$ testaus tapahtuu täysin vastaavasti. Erona on, että suuret havaitun testisuureen arvot antavat evidenssiä nollahypoteesiä vastaan vastahypoteesin suntaan. Tällöin kriittinen alue ja sen todennäköisyys vastaavat testisuuren jakauman oikeanpuolista häntää. Kriittinen alue on piirretty Kuvaan 4.2.

Tarkastellaan viimeiseksi hypoteesiparia $\mathsf{H}_0 : \mu = \mu_0, \mathsf{H}_v : \mu \neq \mu_0$. Tällöin poikkeamia vastahypoteesin suuntaan ovat jakauman molempia häntiä vastaavat alueet, jotka valitaan symmetrisiksi. Kriittisen alueen rajoittamat kvantiilit haetaan siten, että molemmille hännille annetaan $\alpha/2$ verran todennäköisyyttä, jolloin häntien yhteenlaskettu todennäköisyys vastaa testin merkitsevyystasoa α. Seuraavaksi lasketaan havaitun testisuureen arvo z_{hav}. Jos z_{hav} osuu kriittiselle alueelle, niin nollahypoteesi hylätään riskillä α.



Kuva 4.2: z-testisuureen jakauma ehdolla, että nollahypoteesi $\mu = \mu_0$ on voimassa. Vastahypoteesina on $\mu > \mu_0$. Jos havaittu testisuureen arvo osuu kriittiselle alueelle, joka vastaa oikeanpuolista häntätodennäköisyyttä, niin nollahypoteesi hylätään riskillä α



Kuva 4.3: z-testisuureen jakauma ehdolla, että nollahypoteesi $\mu = \mu_0$ on voimassa. Vastahypoteesina on $\mu \neq \mu_0$. Jos havaittu testisuureen arvo osuu kriittiselle alueelle, joka vastaa vasemman- sekä oikeanpuolista häntätodennäköisyyttä, niin nollahypoteesi hylätään riskillä α

Eri tilanteet ovat esitetty tiivistetysti seuraavassa taulukossa, kun halutaan testata nollahypoteesia $\mathsf{H}_0 : \mu = \mu_0$ merkitsevyystasolla α.

H_1	Kriittinen alue	P-arvo
$\mu < \mu_0$	$(-\infty, -z_\alpha)$	$\Phi(z_{\text{hav}})$
$\mu > \mu_0$	(z_α, ∞)	$1 - \Phi(z_{\text{hav}})$
$\mu \neq \mu_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$	$2 \cdot (1 - \Phi(\lvert z_{\text{hav}} \rvert))$

Taulukon kriittisistä alueista huomataan, että merkitsevyystason α testi voidaan itse asiassa tehdä hakemalla yksi tai kaksipuolinen $100(1-\alpha)\%$-luottamusväli, jolloin nollahypoteesia ei voida hylätä, jos luottamusväli sisältää nollahypoteesin mukaisen odotusarvon. Esimerkiksi tapaukselle $\mu > \mu_0$ on yhtäpitävää, että havaittu testisuureen arvo osuu kriittiselle alueelle kuin, että μ_0 ei kuulu yläpuoliseen $100(1-\alpha)\%$-luottamusväliin.

$$z_{\text{hav}} \geq z_\alpha \quad \Leftrightarrow \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_\alpha \quad \Leftrightarrow \quad \mu_0 \leq \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Esimerkki 41. Älykkyystestin pisteiden jakaumaa tietyn suuruisen oppivelvollisuuskoulun oppilaille tiedetään voitavan kuvata normaalijakaumalla, jonka keskihajonta on 12. Koulun oppilaista valittiin

sadan kokoinen otos, ja pisteiden otoskeskiarvoksi saatiin 102.6. Halutaan testata tasolla $\alpha = 0.01$, että onko kyseisen koulun opiskelijoiden älykkyystestin tulos parempi kuin valtakunnallisella tasolla, jossa pisteiden jakauman odotusarvo on 100. Hypoteesipariksi valitaan

$$\begin{aligned} \mathsf{H}_0 &: \mu = 100 \\ \mathsf{H}_v &: \mu > 100 \end{aligned}$$

Haetaan kriittisen alueen rajoittava yläkvantiili $z_{0.01} = 2.33$. Havaittu testisuureen arvo on

$$z_{\text{hav}} = \frac{102.6 - 100}{12/\sqrt{100}} \approx 2.17.$$

Koska havaittu testisuureen arvo ei osu kriittiselle alueelle $[2.33, \infty)$, niin nollahypoteesia ei hylätä tasolla $\alpha = 0.01$. Testin p-arvo on

$$\mathsf{p-arvo} = 1 - \Phi(z_{\text{hav}}) = 1 - \Phi(2.17) = 1 - 0.9850 = 0.0150.$$

Testin p-arvosta nähdään, että nollahypoteesi oltaisiin voitu hylätä esimerkiksi riskillä $\alpha = 0.05$.

Esimerkki 42. Tutkittaessa suomalaisten perheiden TV:n katselua havaittiin, että TV:tä pidettiin avoinna keskimäärin 30 h viikossa. Tutkimus perustui 60 perheen otokseen. Aiempien tutkimusten perusteella voidaan olettaa, että TV:n avoinna pitämisajan keskihajonta on 10.5 h. Läntisten teollisuusmaiden perheissä TV:n viikottainen aukioloaika on keskimäärin 34 h. Halutaan näiden havaintojen perusteella selvittää merkitsevyystasolla 0.05, että katsotaanko Suomessa yhtälailla TV:tä kuin muissa läntisissä teollisuumaissa.

Hypoteesipari on nyt muotoa $\mathsf{H}_0 : \mu = 34, \quad \mathsf{H}_v : \mu \neq 34$. Testi tehdään tasolla $\alpha = 0.05$, joten kriittinen alue on

$$|z| \geq z_{0.025} = 1.96.$$

Z-testisuureen havaittu arvo on

$$z_{\text{hav}} = \frac{30 - 34}{10.5/\sqrt{60}} \approx -2.95.$$

Koska z_{hav} kuuluu kriittiselle alueelle $|-2.95| \geq 1.96$, niin nollahypoteesi hylätään riskitasolla 0.05. Testin p-arvo on

$$\mathsf{p-arvo} = 2(1 - \Phi(|-2.95|)) = 2 \cdot (1 - 0.9984) = 0.0032.$$

Joten pienin riskitaso, jolla nollahypoteesi voitaisiin hylätä on 0.0032.

Suhteellisen osuuden testaus

Tarkastellaan tilannetta, jossa hypoteesi koskee populaatiojakauman suhteellista osuutta. Jonkin kaksiarvoisen tilastollisen muuttujan toisen arvon suhteellinen osuus populaatiossa on p. Piste-estimoinnista muistetaan, että suhteellisella frekvenssillä $\hat{P}$ on odotusarvo sekä varianssi

$$\begin{aligned} \mathsf{E}\left(\hat{P}\right) &= p \\ \mathsf{Var}\left(\hat{P}\right) &= \frac{p(1-p)}{n}. \end{aligned}$$

Lisäksi, jos otoskoko n on suuri, niin pätee likimääräisesti

$$\hat{P} \sim \mathsf{Normal}\left(p, \frac{p(1-p)}{n}\right).$$

Suhteellista osuutta koskevat hypoteesiparit ovat muotoa

$\mathsf{H}_0 : \quad p = p_0$	$\mathsf{H}_0 : \quad p = p_0$	$\mathsf{H}_0 : \quad p = p_0$
$\mathsf{H}_v : \quad p < p_0$	$\mathsf{H}_v : \quad p > p_0$	$\mathsf{H}_v : \quad p \neq p_0$

Oletetaan, että nollahypoteesi on tosi. Tällöin $p = p_0$, ja suureen otoskokoon n vedottuna pätee

$$\hat{P} \sim \mathsf{Normal}\left(p_0, \frac{p_0(1-p_0)}{n}\right),$$

jossa odotusarvo ja varianssi ovat tunnettuja. Nyt suhteellisen osuuden testaus voidaan tehdä z-testillä, jossa testisuure on muotoa

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathsf{Normal}\,(0,1)\,.$$

Esimerkki 43. Erään puolueen kannatus valtiollisissa vaaleissa oli 18.7%. Uusien vaalien kynnyksellä tehdyssä mielipidetiedustelussa, joka suoritettiin otantatutkimuksen vaatimukset täyttäen, haastateltiin 1500 äänioikeutettua. Haastatelluista 253 ilmoitti äänestävänsä kyseistä puoluetta. Halutaan selvittää, että onko puolueen kannatus laskenut valtiollisista vaaleista. Hypoteesipari on nyt

$$\mathsf{H}_0 : p = 0.187 \quad \mathsf{H}_v : p < 0.187.$$

Tehdään testi tasolla $\alpha = 0.05$, jolloin kriittinen alue on muotoa

$$z \leq -z_{0.05} = -1.64.$$

Lasketaan ensin suhteellinen frekvenssi

$$\hat{p} = \frac{253}{1500} \approx 0.1687.$$

Havaittu testisuureen arvo on

$$z_{\text{hav}} = \frac{0.1687 - 0.187}{\sqrt{\frac{0.187(1-0.187)}{1500}}} \approx -1.82,$$

joka kuuluu kriittiselle alueelle. Nollahypoteesi siis hylätään riskillä 0.05. Testin p-arvo on

$$\mathsf{p-arvo} = \Phi(-1.82) \approx 0.03,$$

joten pienin riski, jolla nollahypoteesi voidaan hylätä on n. 0.03.

Kahden suhteellisen osuuden testaus

Tarkastellaan tilannetta, jossa hypoteesi koskee kahden populaatiojakauman suhteellisten osuuksien yhtäsuuruutta. Kahden kaksiarvoisen tilastollisen muuttujan toisen arvon suhteelliset osuudet kahdessa populaatiossa ovat p_1 ja p_2. Piste-estimoinnista muistetaan, että suhteellisten frekvenssien erotuksella $\hat{P}_1 - \hat{P}_2$ on odotusarvo sekä varianssi

$$\begin{aligned}\mathsf{E}\left(\hat{P}_1 - \hat{P}_2\right) &= p_1 - p_2 \\ \mathsf{Var}\left(\hat{P}_1 - \hat{P}_2\right) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.\end{aligned}$$

Lisäksi, jos otoskokoot n_1 sekä n_2 on suuria, niin pätee likimääräisesti

$$\hat{P}_1 - \hat{P}_2 \sim \mathsf{Normal}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Suhteellisia osuuksia koskevat hypoteesiparit ovat muotoa

$\mathsf{H}_0 : \quad p_1 - p_2 = 0$	$\mathsf{H}_0 : \quad p_1 - p_2 = 0$	$\mathsf{H}_0 : \quad p_1 - p_2 = 0$
$\mathsf{H}_v : \quad p_1 - p_2 < 0$	$\mathsf{H}_v : \quad p_1 - p_2 > 0$	$\mathsf{H}_v : \quad p_1 - p_2 \neq 0$

Oletetaan, että nollahypoteesi on tosi. Tällöin $p_1 = p_2 = p$, ja suuriin otoskokoihin n_1 ja n_2 vedottuna pätee likimääräisesti

$$\hat{P}_1 - \hat{P}_2 \sim \mathsf{Normal}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right) = \mathsf{Normal}\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

jossa varianssi riippuu nyt tuntemattomasta parametrista p. Estimoidaan keskivirhettä likimääräisesti

$$\mathsf{SE}(\hat{P}_1 - \hat{P}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

jossa

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}.$$

On tärkeää huomata, että koska nollahypoteesi-oletuksen mukaisesti pätee $p_1 = p_2 = p$, niin tuntematonta suhteellista osuutta voidaan estimoida p ottamalla molemmista populaatiosta tilastoyksiköt, joilla on haluttu ominaisuus ($y_1 + y_2$ kpl) ja suhteuttamalla tämän frekvenssin molempien populaatioiden yhteenlaskettuun kokoon ($n_1 + n_2$ kpl).

Nyt suhteellisten osuuksien yhtäsuuruuden testaus voidaan tehdä z-testillä, jossa testisuure on muotoa

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathsf{Normal}\,(0,1)\,,$$

jossa

$$\hat{P}_1 = \frac{Y_1}{n_1}, \quad \hat{P}_2 = \frac{Y_2}{n_2}, \quad \hat{P} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Esimerkki 44. Ruuveja tuottavien koneiden A ja B tuotannosta poimittiin otokset. Koneen A 200 kpl kokoisessa otoksessa oli 22 viallista ruuvia ja koneen B 100 kpl kokoisessa otoksessa oli 4 viallista ruuvia. Halutaan tarkastella onko koneiden tuottamien viallisten ruuvien suhteellinen osuuksissa (A: p_1 ja B: p_2) eroa. Tehdään testi tasolla 0.01. Hypoteesipari on

$$\mathsf{H}_0 : p_1 - p_2 = 0, \quad \mathsf{H}_v : p_1 - p_2 \neq 0$$

Kriittinen alue on tällöin

$$|z| \geq z_{0.005} = 2.57.$$

Lasketaan ensin välituloksia

$$\begin{aligned}
\hat{p}_1 &= \frac{22}{200} = \frac{11}{100} \\
\hat{p}_2 &= \frac{4}{100} \\
\hat{p}_1 - \hat{p}_2 &= \frac{11}{100} - \frac{4}{100} = \frac{7}{100} = 0.07 \\
\hat{p} &= \frac{4+11}{200+100} = \frac{13}{150} \approx 0.0867 \\
\mathsf{SE}(\hat{P}_1 - \hat{P}_2) &= \sqrt{\frac{13}{150}\left(1 - \frac{13}{150}\right)\left(\frac{1}{200} + \frac{1}{100}\right)} \approx 0.03446.
\end{aligned}$$

Nyt havaittu testisuureen arvo on

$$z_{\text{hav}} = \frac{0.07}{0.03446} \approx 2.03.$$

Koska havaittu testisuureen arvo ei osu kriittiselle alueelle, niin nollahypoteesia ei voida hylätä tasolla 0.01. Testin p-arvo on

$$\mathsf{p-arvo} = 2(1 - \Phi(|2.03|)) = 2 \cdot 0.0212 = 0.0424.$$

Joten pienin riski, jolla nollahypoteesi voitaisiin hylätä, on 0.0424.

4.2 t-testi

Odotusarvon testaus kun varianssi ei ole tunnettu

Tarkastellaan tilannetta, jossa populaatiojakauma on normaalijakauma Normal (μ, σ^2), jonka varianssia ei tunneta. Ollaan kiinnostuttu populaatiojakauman odotusarvosta μ, jota koskevat hypoteesiparit ovat

$\mathsf{H}_0: \quad \mu = \mu_0$	$\mathsf{H}_0: \quad \mu = \mu_0$	$\mathsf{H}_0: \quad \mu = \mu_0$
$\mathsf{H}_v: \quad \mu < \mu_0$	$\mathsf{H}_v: \quad \mu > \mu_0$	$\mathsf{H}_v: \quad \mu \neq \mu_0.$

Oletetaan, että nollahypoteesi on tosi. Tällöin $\mu = \mu_0$, ja valitaan odotusarvon testaukseen liittyväksi otossuureeksi **t-testisuureen**, jonka jakaumaksi tunnetaan Studentin t-jakauma vapausastein $n-1$, jossa n on populaatiosta kerättävien havaintojen määrä

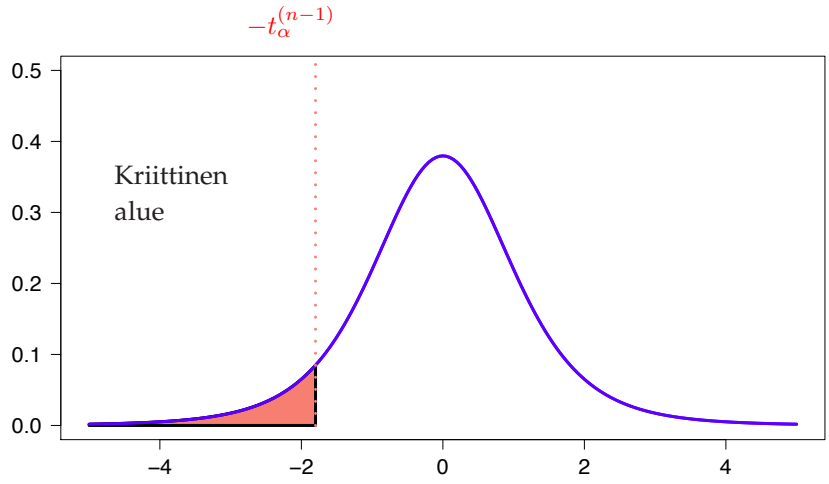
$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim \mathsf{t}(n-1),$$

jossa $\bar{X}$ ja S ovat otoskeskiarvo sekä otoshajonta, ja n on havaintoaineiston koko.

Testaaminen tapahtuu samalla tavalla kuin z-testissä. Valitaan testin merkitsevyystasoksi α. Haetaan α-kokoiset kriittinen alueet hypoteesiparin perusteella ja katsotaan osuuko havaittu testisuureen arvo

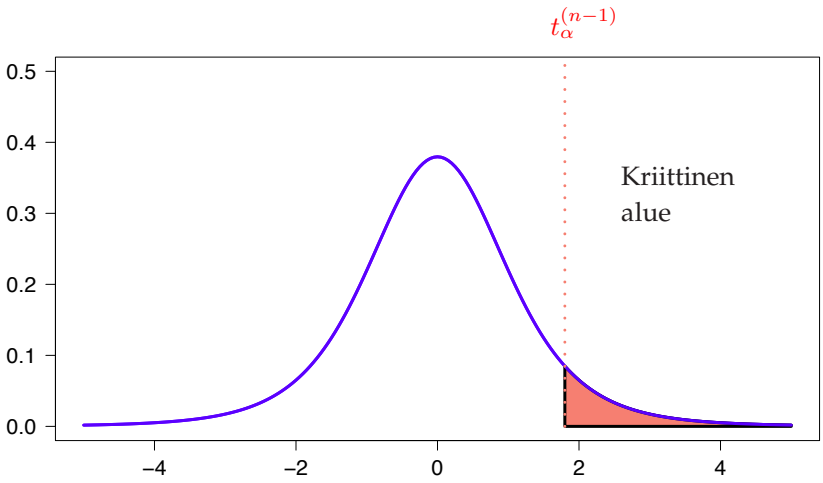
$$t_{\text{hav}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

kriittiselle alueelle vai ei. Eri hypoteesiparien kriittiset alueet ovat havainnollistettuina Kuvissa 4.4–4.6 Jos havaittu testisuure osuu kriittiselle alueelle, niin nollahypoteesi voidaan hylätä riskillä α, muuten jatekaan nollahypoteesin kanssa. Lisäksi lasketaan testille p-arvo, eli todennäköisyys, että havaitaan vähintään havaitun suuruisen poikkeama vastahypoteesin suuntaan.

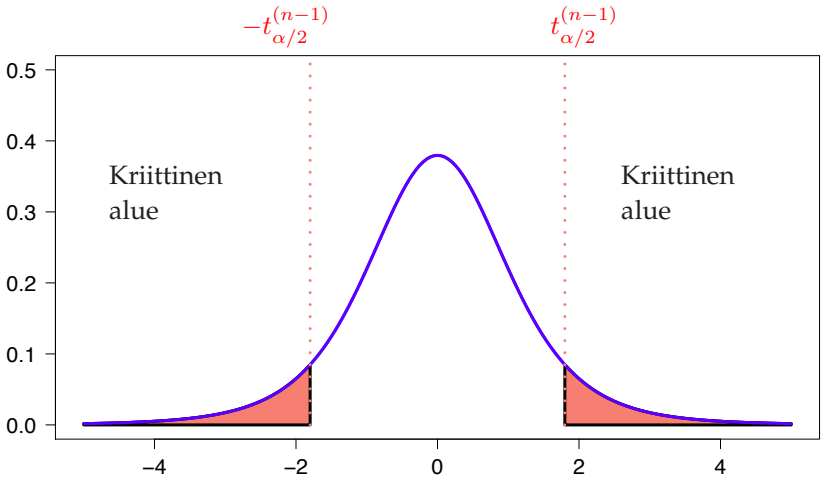


Kuva 4.4: t-testisuureen jakauma ehdolla, että nollahypoteesi $\mu = \mu_0$ on voimassa, ja havaintoja on kerätty n kpl. Vastahypoteesina on $\mu < \mu_0$. Jos havaittu testisuureen arvo osuu kriittiselle alueelle, joka vastaa vasemmanpuolista häntätodennäköisyyttä, niin nollahypoteesi voidaan hylätä riskillä α

Merkitään nyt t$(n-1)$-jakauman kertymäfunktiota symbolilla $F_{T_{n-1}}$, tällöin hypoteesin testauksen tärkeät arvot voidaan tiivistää eri hypoteesipareille seuraavaan taulukkoon



Kuva 4.5: t-testisuureen jakauma ehdolla, että nollahypoteesi $\mu = \mu_0$ on voimassa, ja havaintoja on kerätty n kpl. Vastahypoteesina on $\mu > \mu_0$. Jos havaittu testisuureen arvo osuu kriittiselle alueelle, joka vastaa oikeanpuolista häntätodennäköisyyttä, niin nollahypoteesin voidaan hylätä riskillä α



Kuva 4.6: t-testisuureen jakauma ehdolla, että nollahypoteesi $\mu = \mu_0$ on voimassa, ja havaintoja on kerätty n kpl. Vastahypoteesina on $\mu \neq \mu_0$. Jos havaittu testisuureen arvo osuu kriittiselle alueelle, joka vastaa vasemman- sekä oikeanpuolista häntätodennäköisyyttä, niin nollahypoteesi voidaan hylätä riskillä α

H_1	Kriittinen alue	P-arvo
$\mu > \mu_0$	$t \geq t_\alpha^{(n-1)}$	$1 - F_{T_{n-1}}(t_{\text{hav}})$
$\mu < \mu_0$	$t \leq -t_\alpha^{(n-1)}$	$F_{T_{n-1}}(t_{\text{hav}})$
$\mu \neq \mu_0$	$\lvert t\rvert \geq t_{\alpha/2}^{(n-1)}$	$2(1 - F_{T_{n-1}}(\lvert t_{\text{hav}}\rvert))$.

Käytännössä testattaessa hypoteeseja käsin voidaan kriittisten alueiden rajat lukea t-jakauman yläkvantiilitaulukosta. Testin p-arvoja hakiessa taulukon yläkvantiilien avulla, täytyy meidän tyytyä ala- ja ylärajoihin p-arvolle, sillä taulukon α-yläkvantiilien määrä on hyvin rajallinen. Rajojen hakeminen tapahtuu seuraavasti. Olkoon t_{hav} havaittu testisuureen arvo, ja olkoon testisuureen T vapausteiden määrä k. Etsitään yläkvantiilitaulukon vapausastetta k vastaava rivi. Tarkastellessa vastahypoteesia

$\mu < \mu_0$ haetaan kaksi vierekkäistä saraketta, joiden arvojen väliin $-t_{\text{hav}}$ osuu. Vastaavasti vastahypoteesin $\mu > \mu_0$ tapauksessa haetaan kaksi vierekkäistä saraketta, joiden arvojen väliin t_{hav} osuu, ja vastahypoteesin $\mu \neq \mu_0$ tapauksessa haetaan kaksi vierekkäistä saraketta, joiden arvojen väliin $|t_{\text{hav}}|$ osuu. Olkoon nämä yläkvantiilit $t_{\alpha_1}^{(k)}$ sekä $t_{\alpha_2}^{(k)}$. Nyt p-arvolle saadaan haettua α_1:n ja α_2:n avulla rajat seuraavan taulukon mukaisesti.

H_1	Yläkvantiiliväli	P-arvo:n rajat
$\mu > \mu_0$	$t_{\alpha_2}^{(k)} < t_{\text{hav}} < t_{\alpha_1}^{(k)}$	$\alpha_1 < \mathsf{p-arvo} < \alpha_2$
$\mu < \mu_0$	$t_{\alpha_2}^{(k)} < -t_{\text{hav}} < t_{\alpha_1}^{(k)}$	$\alpha_1 < \mathsf{p-arvo} < \alpha_2$
$\mu \neq \mu_0$	$t_{\alpha_2}^{(k)} < \lvert t_{\text{hav}}\rvert < t_{\alpha_1}^{(k)}$	$2\alpha_1 < \mathsf{p-arvo} < 2\alpha_2.$

Esimerkki 45. Tutkitaan tupakoivien äitien lasten syntymäpainoja, ja ollaan kiinnostuttu olisiko tupakoivien äitien lapsien syntymäpainot keskimäärin pienempiä kuin tupakoimattomien äitien lasten keskimääräinen syntymäpaino, joka on 3.56 kg. Oletetaan aiemman kerätyn tiedon perusteella, että syntymäpainot ovat normaalijakautuneita. Kerätään 14 lapsen syntymäpaino x_i kilogrammoissa, josta lasketaan otoskeskiarvo $\bar{x} = 3.37$, sekä otosvarianssi $s^2 = 0.15$.

Testin hypoteesipari on $\mathsf{H}_0 : \mu = 3.56, \quad \mathsf{H}_v : \mu < 3.56$, ja testi tehdään tasolla 0.05. Nollahypoteesin voimassa ollessa t-testisuure on muotoa

$$T = \frac{\bar{X} - 3.37}{\sqrt{0.15}/\sqrt{14}} \sim \mathsf{t}(13).$$

Kriittinen alue on nyt muotoa $t \leq -t_{0.05}^{(13)} = -1.771$. Havaittu testisuureen arvo on nyt

$$t_{\text{hav}} = \frac{3.37 - 3.56}{\sqrt{0.15}/\sqrt{14}} \approx -1.84,$$

joka osuu kriittiselle alueelle. Nyt siis nollahypoteesi voidaan hylätä riskillä 0.05. Rajat p-arvolle voidaan hakea t-jakauman yläkvantiilitaulukosta hakemalla vapausasteiden 13 kohdalta yläkvantiilit joiden väliin $-t_{\text{hav}}$ osuu. Taulukosta löydetään yläkvantiilit $t_{0.05}^{(13)} = 1.771$ sekä $t_{0.025}^{(13)} = 2.160$. Tästä seuraa, että $0.025 < \mathsf{p-arvo} < 0.05$.

Parittaisten havaintojen erotuksen testaus

Tarkastellaan tilannetta jossa ollaan kiinnostuttu kahden populaation parittaisten havaintojen odotusarvojen eroista. Kun ensimmäisestä populaatiosta on kerätty riippumattomat havainnot $X_1, \ldots, X_n$ ja toisesta $Y_1, \ldots, Y_n$, niin lasketaan erotukset $D_i = X_i - Y_i$, joiden oletetaan olevan normaalijakautuneita

$$D_i \sim \mathsf{Normal}\left(\mu, \sigma^2\right).$$

Mielenkiinto kohdistuu nyt parittaisten havaintojen erotuksen jakauman odotusarvoon μ, jota koskevat hypoteesiparit ovat

$\mathsf{H}_0 : \quad \mu = \mu_0$	$\mathsf{H}_0 : \quad \mu = \mu_0$	$\mathsf{H}_0 : \quad \mu = \mu_0$
$\mathsf{H}_v : \quad \mu < \mu_0$	$\mathsf{H}_v : \quad \mu > \mu_0$	$\mathsf{H}_v : \quad \mu \neq \mu_0.$

Hyvin usein $\mu_0 = 0$, jolloin testataan onko parittaisissa havainnoissa ero johonkin suuntaan. Testaus tapahtuu käytännössä t-testillä kuten edellä. Toki jos erotusten jakauman hajontakin olisi tunnettu, niin z-testin käyttö olisi mahdollista.

Esimerkki 46. Tarkastellaan erään dieetin vaikutusta 10 koehenkilön avulla. Jokainen koehenkilö punnitaan sekä ennen että jälkeen dieettijakson. Saadaan havainnot (kilogrammoissa)

Paino ennen:	x_i	94	92	103	100	97	99	87	78	77	96
Paino jälkeen:	y_i	93	90	95	96	90	100	82	80	73	92
Erotus:	d_i	1	2	8	4	7	-1	5	-2	4	4

Erotusten otoskeskiarvo ja otosvarianssi ovat

$$\bar{d} = 3.2$$
$$s^2 = 10.4$$

Keskivirheeksi arvioidaan

$$\mathsf{SE}(\bar{D}) = \sqrt{\frac{10.4}{10}} = 1.019804 \approx 1.02$$

Oletetaan, että erotukset ovat normaalijakautuneita $D_i \sim \mathsf{Normal}\left(\mu, \sigma^2\right)$. Testataan tasolla 0.05 onko dieettijaksolla alentavaa vaikutusta painoon. Hypoteesipari on nyt (huomaa, että erotukset ovat paino ennen - paino jälkeen)

$$\mathsf{H}_0 : \mu = 0$$
$$\mathsf{H}_v : \mu > 0.$$

Kriittinen alue on nyt $t \geq t^{(9)}_{0.05} = 1.833$. Havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{3.2}{1.019804} = 3.138,$$

joka kuuluu kriittiselle alueelle, joten nollahypoteesi hylätään tasolla 0.05.

Lisäksi koska havaittu testisuureen arvo osuu yläkvantiilien $t^{(9)}_{0.01} = 2.821$ ja $t^{(9)}_{0.005} = 3.250$ väliin, niin $0.005 < \mathsf{p-arvo} < 0.01$.

Kahden populaatiojakauman odotusarvojen erotuksen testaus

Tarkastellaan tilannetta, jossa halutaan tutkia kahden populaation keskimääräisiä arvoja. Tutkitaan siis jakaumien odotusarvojen eroa havaintojen riippumattomien havaintojen $X_1, \ldots, X_{n_1}$ ja $Y_1, \ldots, Y_{n_2}$ avulla. Oletetaan, että molemmat populaatiot ovat normaalijakautuneita

$$X_i \sim \mathsf{Normal}\left(\mu_1, \sigma_1^2\right)$$
$$Y_i \sim \mathsf{Normal}\left(\mu_2, \sigma_2^2\right).$$

Populaatioiden keskimääräisten arvojen erotuksia testatessa mielenkiinto kohdistuu odotusarvojen erotukseen $\mu_1 - \mu_2 = \delta$. Hypoteesiparit ovat muotoa

$\mathsf{H}_0 : \quad \mu_1 - \mu_2 = \delta_0$	$\mathsf{H}_0 : \quad \mu_1 - \mu_2 = \delta_0$	$\mathsf{H}_0 : \quad \mu_1 - \mu_2 = \delta_0$
$\mathsf{H}_v : \quad \mu_1 - \mu_2 < \delta_0$	$\mathsf{H}_v : \quad \mu_1 - \mu_2 > \delta_0$	$\mathsf{H}_v : \quad \mu_1 - \mu_2 \neq \delta_0.$

Tarkastellaan kahta tilannetta, joissa populaatioiden varianssit σ_1^2 ja σ_2^2 ovat yhtäsuuria tai ne eivät ole yhtäsuuria.

Populaatioiden varianssit ovat yhtäsuuria

Oletetaan, että populaatioiden variansseille pätee $\sigma_1^2 = \sigma_2^2$. Tällöin voidaan yhteisotosvarianssilla S_p^2 estimoida otoskeskiarvojen erotuksen keskivirhettä $\mathsf{SE}(\bar{X} - \bar{Y})$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
$$\mathsf{SE}(\bar{X} - \bar{Y}) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Luottamusvälejä käsitellessä todettiin, että testisuureelle T pätee

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\mathsf{SE}(\bar{X} - \bar{Y})} \sim \mathsf{t}(n_1 + n_2 - 2).$$

Oletetaan nyt, että nollahypoteesi $\mu_1 - \mu_2 = \delta_0$ on tosi. Valitaan testin merkitsevyystasoksi α. Testaaminen tapahtuu taas täsmälleen samoin tavoin kuin odotusarvon testaaminen t-testillä. Nyt havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{\bar{x} - \bar{y} - \delta_0}{s_y\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

ja t-jakaumalla on vapausasteita $n_1 + n_2 - 2$.

Esimerkki 47. Fyysikot A ja B mittasivat erään aineen tiettyä ominaisuutta ja saivat tuloksiksi eri kerroilla seuraavat arvot:

Mittaaja	Arvot						
A	87.3	91.8	91.3	93.8	89.4	90.7	92.2
B	89.8	87.2	88.5	89.1	92.1	84.9	90.6

Oletetaan, että mittaukset ovat normaalijakautuneita odotusarvoilla μ_A, μ_B ja mittausten keskihajonnat ovat samat $\sigma_A = \sigma_B = \sigma$. Testataan tasolla 0.05 poikkevatko fyysikoiden tulokset toisistaan tilastollisesti merkitsevästi. Aineistosta voidaan laskea

$$n_A = 7, \quad \bar{x}_A = 90.929, \quad s_A^2 = 4.386$$
$$n_B = 7, \quad \bar{x}_B = 88.886, \quad s_B^2 = 5.505,$$

joiden avulla lasketaan yhteisotosvarianssi ja keskivirhe estimaattorille $\bar{X}_A - \bar{X}_B$

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = 4.9455$$
$$\mathsf{SE}(\bar{X}_A - \bar{X}_B) = \sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} = \sqrt{4.9455 \cdot \left(\frac{1}{7} + \frac{1}{7}\right)} \approx 1.1887.1.$$

Hypoteesipari on

$$\mathsf{H}_0 : \mu_A - \mu_B = 0$$
$$\mathsf{H}_v : \mu_A - \mu_B \neq 0.$$

Kriittinen alue on nyt $|t| \geq t_{(0.025)}^{(7+7-2)} = 2.179$. Havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{\bar{x}_A - \bar{x}_B}{\mathsf{SE}(\bar{X}_A - \bar{X}_B)} = \frac{90.929 - 88.886}{1.1887} = 1.719,$$

joka ei osu kriittiselle alueelle. Nollahypoteesi jää siis voimaan tasolla 0.05. Koska havaitun testisuureen itsearvo osuu yläkvantiilien $t_{0.10}^{(12)} = 1.356$ ja $t_{0.05}^{(12)} = 1.782$ väliin ja vastahypoteesi on kaksipuolinen, niin $2 \cdot 0.05 < \mathsf{p-arvo} < 2 \cdot 0.10$.

Populaatioiden varianssit eivät ole yhtäsuuria

Tarkastellaan tilannetta, jossa ei ole perusteltua olettaa, että $\sigma_1^2 = \sigma_2^2$. Luottamusvälien tarkastelussa todettiin, ettei ole mahdollista päästä käsiksi parametrista muotoa olevaan todennäköisyysjakaumaan, jolla voitaisiin tarkastella odotusarvoa $\mu_1 - \mu_2$. Sen sijaan käytetään Welch-Satterthwaite-approksimaatiota, jonka mukaan otossuureella

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}},$$

likimääräisesti Studentin t-jakauma vapausastein

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Oletetaan nyt, että nollahypoteesi $\mu_1 - \mu_2 = \delta_0$ on tosi. Valitaan testin merkitsevyystasoksi α. Testaaminen tapahtuu taas täsmälleen samoin tavoin kuin odotusarvon testaaminen t-testillä. Nyt havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}},$$

ja t-jakaumalla vapausasteet ovat Welch-Satterthwaite approksimaation mukaiset v.

Esimerkki 48. Kymmenen 15-vuotiasta poikaa ja kahdeksan 15-vuotiasta tyttöä osallistui tiettyyn testiin. Testissä havaittiin poikien pistemäärien keskiarvoksi 10.5 ja tyttöjen 9.2. Keskihajonnat olivat vastaavasti 3.1 ja 2.9. Piste-estimaatiksi poikien ja tyttöjen populaatioiden odotusarvojen erotukselle laskettiin

$$\bar{x} - \bar{y} = 10.5 - 9.2 = 1.3.$$

Jos populaatioiden hajontoja ei voida olettaa yhtäsuuriksi, niin

$$\mathsf{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = \sqrt{\frac{3.1^2}{10} + \frac{2.9^2}{8}} = 1.418538,$$

Testataan tasolla 0.05 poikkeavatko poikien ja tyttöjen keskimääräiset arvosanat toisistaan. Hypoteesipari on

$$\mathsf{H}_0 : \mu_x - \mu_y = 0$$
$$\mathsf{H}_v : \mu_x - \mu_y \neq 0$$

Nollahypoteesin ollessa voimassa testisuureella on likimääräisesti t-jakauma

$$T = \frac{\bar{X} - \bar{Y}}{\mathsf{SE}(\bar{X} - \bar{Y})} \sim \mathsf{t}(15.54444)$$

T-jakauman vapausaste 15.54444 lasketaan kuten Welch-Satterthwaite-approksimaatiossa.

Kriittiseksi alueeksi voidaan likimääräisesti arvioida taulukoitujen arvojen perusteella $|t| \geq \frac{1}{2}(t^{(15)}_{0.025} + t^{(16)}_{0.025}) = \frac{1}{2}(2.131 + 2.120) = 2.1255$. Havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{1.3}{1.418538} = 0.9164365,$$

joka ei kuulu kriittiselle alueelle. Nollahypoteesi jää siis voimaan tasolla 0.05. Sekä vapausasteille 15 että 16 havaittu testisuureen arvo osuu $0.25-$ ja $0.10-$yläkvantiilien väliin, joten $2 \cdot 0.10 < \mathsf{p} - \mathsf{arvo} < 2 \cdot 0.25$.

4.3 F-testi

Tarkastellaan tilannetta jossa halutaan selvittää onko kahdella normaalijakautuneella populaatiolla yhtäsuuri varianssi. Testaamiseen voidaan käyttää aiemmin esiteltyä F-testisuuretta. Tarkastellaan populaatioista kerättäviä riippumattomia havaintoja $X_1, \ldots, X_n$ ja $Y_1, \ldots, Y_m$, joista lasketaan otosvarianssit S_X^2 ja S_Y^2. Nyt $X_i \sim \mathsf{Normal}\left(\mu_X, \sigma_X^2\right)$, ja $Y_i \sim \mathsf{Normal}\left(\mu_Y, \sigma_Y^2\right)$. F-testisuureelle F pätee

$$F = \frac{\sigma_Y^2}{\sigma_X^2}\frac{S_X^2}{S_Y^2} \sim \mathsf{F}(n-1, m-1),$$

jossa $\mathsf{F}(n-1, m-1)$ on F-jakauma vapausastein $n-1$ ja $m-1$. Varianssien yhtäsuuruutta testattaessa mielenkiinto kohdistuu populaatiovarianssien suhteeseen. Hypoteesiparit ovat muotoa

$\mathsf{H}_0: \ \frac{\sigma_Y^2}{\sigma_X^2} = 1$	$\mathsf{H}_0: \ \frac{\sigma_Y^2}{\sigma_X^2} = 1$	$\mathsf{H}_0: \ \frac{\sigma_Y^2}{\sigma_X^2} = 1$
$\mathsf{H}_v: \ \frac{\sigma_Y^2}{\sigma_X^2} < 1$	$\mathsf{H}_v: \ \frac{\sigma_Y^2}{\sigma_X^2} > 1$	$\mathsf{H}_v: \ \frac{\sigma_Y^2}{\sigma_X^2} \neq 1.$

Nollahypoteesin $\mathsf{H}_0 : \sigma_X^2 = \sigma_Y^2$ ollessa voimassa

$$F = \frac{S_X^2}{S_Y^2} \sim \mathsf{F}(n-1, m-1).$$

Kun käytetään taukoituja F-jakauman 0.05-yläkvantiileja hakemaan kriittisiä alueita varianssien yhtäsuuruuden F-testille, niin kriittiset alueet jotka on mahdollista hakea, vastaavat 0.05 tasoisia toispuolisia testejä ja 0.10 tasoista kaksipuolista testiä. Merkitseviä poikkeamia vastaavat kriittiset alueet vastahypoteesin suuntaan saadaan eri tapauksissa seuraavan taulukon mukaisesti

H_1	Kriittinen alue
$\frac{\sigma_Y^2}{\sigma_X^2} < 1$	$f \geq f_{0.05}^{(n-1,m-1)}$
$\frac{\sigma_Y^2}{\sigma_X^2} > 1$	$f \leq \frac{1}{f_{0.05}^{(m-1,n-1)}}$
$\frac{\sigma_Y^2}{\sigma_X^2} \neq 1$	$f \leq \frac{1}{f_{0.05}^{(m-1,n-1)}}$ tai $f \geq f_{0.05}^{(n-1,m-1)}$.

Kriittisen alueen hakemisen jälkeen taskistetaan osuuko havaittu testisuureen arvo

$$f_{\text{hav}} = \frac{s_X^2}{s_Y^2}$$

kriittiselle alueelle vai ei. Jos f_{hav} osuu kriittiselle alueelle, niin toispuolisissa testeissä hylätään nollahypoteesi riskillä 0.05 ja jos ei osu, niin nollahypoteesia ei voida hylätä riskillä 0.05. Kaksipuolisissa testeissä taso on vastaavasti 0.10. F-testi on hyvin herkkä poikkeamille normaalijakautuneisuudesta. Sen vuoksi usein käytetään robustimpia menetelmiä, kuten Levenen testiä varianssien yhtäsuuruuden testaamiseksi.

Esimerkki 49. Fyysikot A ja B mittasivat erään aineen tiettyä ominaisuutta ja saivat tuloksiksi eri kerroilla seuraavat arvot:

Mittaaja	Arvot						
A	87.3	91.8	91.3	93.8	89.4	90.7	92.2
B	89.8	87.2	88.5	89.1	92.1	84.9	90.6

Käytetään F-testiä, jolla testataan tasolla 0.10 poikkevatko fyysikoiden tuloksien varianssit toisistaan merkitsevästi. Hypoteesipari on $\mathsf{H}_0 : \sigma_A^2/\sigma_B^2 = 1, \quad \mathsf{H}_v : \sigma_A^2/\sigma_B^2 \neq 1$. Aineistosta voidaan laskea

$$n_A = 7, \quad s_A^2 = 4.386$$
$$n_B = 7, \quad s_B^2 = 5.505.$$

Kriittinen alue on muotoa $f \leq 1/f_{0.05}^{(6,6)} = 1/4.28 = 0.2336$ ja $f \geq f_{0.05}^{(6,6)} = 4.28$, ja havaittu testisuureen arvo on

$$f_{\text{hav}} = \frac{s_A^2}{s_B^2} = \frac{4.386}{5.505} = 0.797.$$

Koska havaittu testisuureen arvo ei kuulu kriittiselle alueelle, jää nollahypoteesi voimaan tasolla 0.10.

4.4 Normaalisuusoletuksen tarkastelu

Monet tarkastellut menetelmät perustuvat oletukseen normaalijakautuneisuudesta. Normaalisuutta voidaan tarkastella erilaisilla tilastollisilla testeillä, mutta tässä luentomonisteessa rajoitutaan tarkastelemaan normaalisuutta kuvailevan tilastotieteen menetelmillä. Normaalikvantiilikuviolla saadaan hyvä kuva aineistosta, jonka lisäksi voidaan laskea vinous, sekä huipukkuuskertoimet aineistolle.

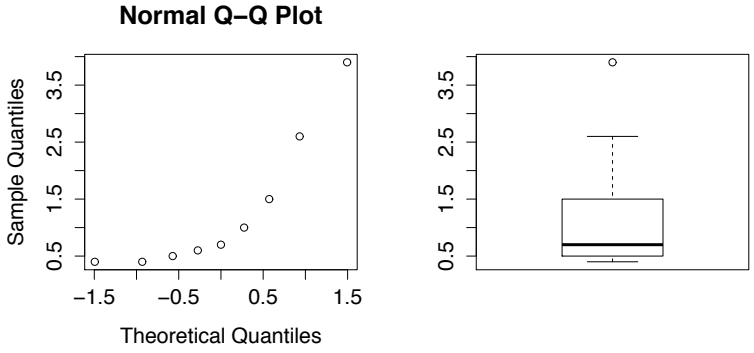
4.4.1 Muunnokset

Kun havaintoaineisto on selkeästi vino ja tilastollinen päättely kohdistyy yhden tai useamman jakauman sijaintiin, kannattaa usein tarkistaa voidaanko havaintoaineiston muotoon vaikuttaa erilaisilla muunnoksilla, jolloin aineisto saataisiin symmetriseksimmäksi tai lähemmäksi normaalijakaumaa. Havaintoaineisto voidaan muuntaa laskemalla jokaisen aineiston arvo jollakin tietyllä funktiolla. Yleisiä käytettyjä muunnostyyppejä ovat erilaiset potenssimuunnokset x^p tai logaritmimuunnos $\ln(x)$, jotka ovat käyttökelpoisia kun havainnot ovat positiivisia. Kun tässä muunnoksessa valitaan $p > 1$, niin saadaan vähennettyä vasemmalle vinoutta, ja jos valitaan $p < 1$, niin saadaan oikealle vinoutta vähennettyä. Usein käytettyjä potenssimuunnoksia on $x^{-1}, x^{1/2}$ ja x^2.

Esimerkki 50. Tarkastellaan havaintoaineistoa

$$0.4 \quad 0.4 \quad 0.5 \quad 0.6 \quad 0.7 \quad 1.0 \quad 1.5 \quad 2.6 \quad 3.9,$$

jota ollaan havainnollistettu normaalikvantiilikuviolla Kuvassa 4.7. Halutaan testata hypoteesiparia $\mathsf{H}_0 : \mu = 2, \mathsf{H}_v : \mu \neq 2$ tasolla 0.05.



Kuva 4.7: Esimerkin 50 aineistojen normaalikvantiilikuviot.

Aineisto vaikuttaa vinolta, ja lasketut vinous- ja huipukkuuskertoimet tukevat vaikutelmaa

$$g_1 = \frac{m_3}{m_2^{1.5}} \approx 1.32$$
$$g_2 = \frac{m_4}{m_2^2} - 3 \approx 0.39.$$

Kokeillaan aineistoon muunnoksia

Muunnos	Vinous	Huipukkuus
$x^{1/2}$	0.99	-0.38
x^{-1}	0.06	-1.40
$\ln(x)$	0.62	-0.98

joista potenssimuunnos x^{-1} vähentää vinoutta huomattavasti. Suoritetaan testi x^{-1}-muunnetulle aineistolle, joka on muotoa

$$1/0.4 \quad 1/0.4 \quad 1/0.5 \quad 1/0.6 \quad 1/0.7 \quad 1.0 \quad 1/1.5 \quad 1/2.6 \quad 1/3.9,$$

ja hypoteesipari on muunnettuna $: \mu = 1/2, \mathsf{H}_v : \mu \neq 1/2$. Kriittinen alue on muotoa $|t| \geq t^{(8)}_{0.025} = 2.306$. Otoskeskiarvo, otosvarianssi ja havaittu testisuureen arvo muunnetulle aineistolle on

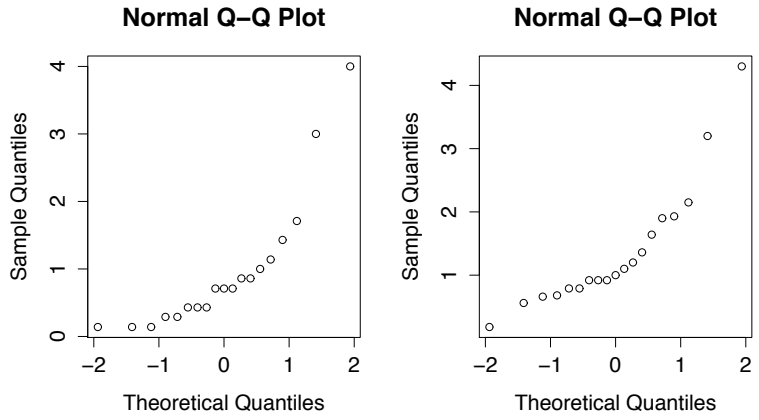
$$\bar{x} = 1.378103$$
$$s^2 = 0.7355242$$
$$t_{\text{hav}} = \frac{\bar{x} - 1/2}{s/\sqrt{9}} = 3.071627,$$

joka osuu kriittiselle alueelle. Voidaan siis tasolla 0.05 hylätä nollahypoteesi, että populaatiolle $\mu = 2$. Jos oltaisiin testattu hypoteesiparia ilman muunnosta aluperäisellä vinolla aineistolla, niin tällöin nollahypoteesi olisi jäänyt voimaan tasolla 0.05.

Esimerkki 51. Halutaan vertailla kahden populaation odotusarvoja. Populaatiosta on kerätty havaintoaineistot

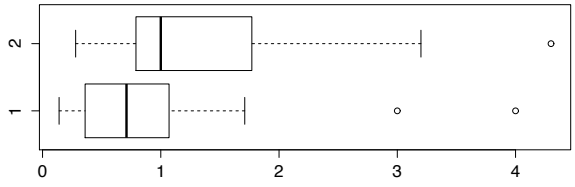
Populaatio 1	0.14	0.14	0.14	0.29	0.29	0.43	0.43	0.43	0.71	0.71
	0.71	0.86	0.86	1.00	1.14	1.43	1.71	3.00	4.00	
Populaatio 2	0.28	0.56	0.66	0.68	0.79	0.79	0.92	0.92	0.92	1.00
	1.10	1.20	1.36	1.90	1.64	1.93	2.15	3.20	4.30.	

Kahden riippumattoman otoksen t-testi perustuu populaatioiden normaalijakautuneisuuteen. Piirretään aineistoista normaalikvantiilikuvaajat (Kuva 4.8), josta nähdään, että normaalikvantiilikuvion pisteet eivät sijoitu hyvin suoralle. Normaalijakautuneisuusoletus ei siis luultavasti ole kovin hyvä.



Kuva 4.8: Esimerkin 51 aineistojen normaalikvantiilikuviot.

Piirretään myös Tukeyn laatikko-janakuviot (Kuva 4.9). Lasketaan vinous- ja huipukkuuskertoimet



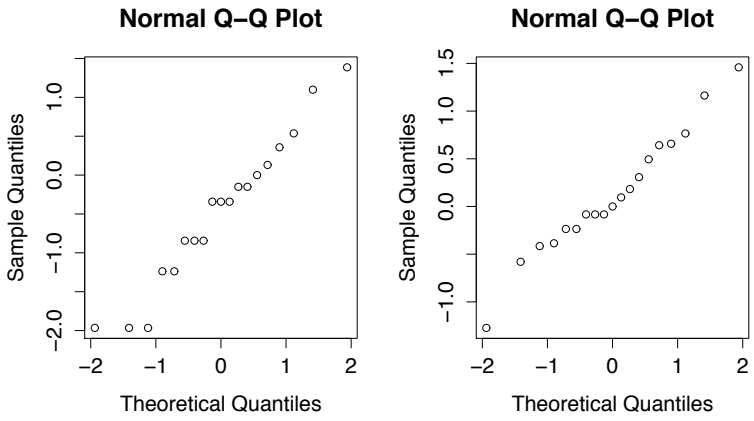
Kuva 4.9: Esimerkin 51 aineistojen normaalikvantiilikuviot.

$$\begin{aligned}
\text{Populaatio 1:} \quad & g_1 = \frac{m_3}{m_2^{1.5}} \approx 1.90 \\
& g_2 = \frac{m_4}{m_2^2} - 3 \approx 2.96 \\
\text{Populaatio 1:} \quad & g_1 \approx 1.67 \\
& g_2 \approx 2.34.
\end{aligned}$$

Molemmat aineistot ovat oikealle vinoja. Testataan aineistoon erilaisia potenssimuunnoksia ja logaritmimuunnosta. Muunnetuille aineistoille saadaan kertoimet

	Populaatio 1		Populaatio 2	
Muunnos	Vinous	Huipukkuus	Vinous	Huipukkuus
$x^{1/2}$	1.06	0.68	0.99	0.66
x^{-1}	1.31	0.36	2.05	4.99
$\ln(x)$	0.04	−0.60	0.13	0.09

joista logaritmimuunnettu aineisto on lähimpänä normaalijakautunutta aineistoa. Muunnettujen aineistojen normaalikvantiilikuviot löytyvät Kuvasta 4.10. Testataan muunnettujen jakaumien odotusar-



Kuva 4.10: Esimerkin 51 logaritmimuunnettujen aineistojen normaalikvantiilikuviot.

vojen yhtäsuuruutta t-testillä tasolla 0.05. Hypoteesipari on $\mathsf{H}_0 : \mu_1 - \mu_2 = 0, \mathsf{H}_v : \mu_1 - \mu_2 \neq 0$. Testataan F-testillä varianssien yhtäsuurutta, jolloin saadaan F-testille tietokoneella laskettuna $\mathsf{p-arvo} = 0.09$ (havaintojen suurestä määrästä johtuen varianssien yhtäsuuruustestiä ei voida tehdä taulukoitujen arvojen perusteella). Edetään siis oletuksella, että jakaumien varianssit ovat samat. Odotusarvot, varianssit, yhteisotosvarianssi sekä likimääräinen keskivirhe ovat

$$\begin{aligned}
\bar{x}_1 &= -0.4592111, \quad s_1^2 = 0.9305988 \\
\bar{x}_2 &= 0.1258352, \quad s_2^2 = 0.4111931 \\
s_p^2 &= 0.670896 \\
\mathsf{SE}(\bar{X}_1 - \bar{X}_2) &= 0.2657454.
\end{aligned}$$

Kriittinen alue testille on $|t| \geq t_{0.025}^{(36)} = 2.028$. Koska havaittu testisuureen arvo on

$$t_{\text{hav}} = \frac{-0.4592111 - 0.1258352}{0.2657454} = -2.201529,$$

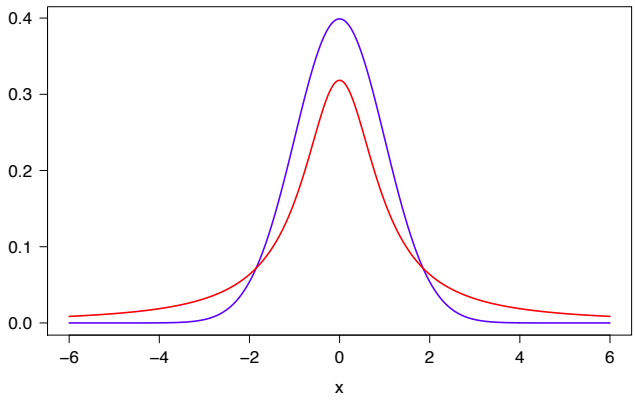
joka kuuluu kriittiselle alueelle, hylätään testin nollahypoteesi riskillä 0.05. Jos oltaisiin testattu odotusarvojen yhtäsuuruutta muuntamattomalla aineistolla, olisi nollahypoteesi jäänyt voimaan jopa tasolla 0.10.

4.4.2 Poikkeavien havaintojen tarkastelu

Poikkeva havainto (vieras, ulkolainen, outlier) aineistossa on sellainen äärimmäinen havainto, jonka arvo on huomattavasti suurempi tai pienempi kuin aineiston arvot yleisesti. Poikkeavien havaintojen esiintyminen usein tarkoittaa joko virheellistä havaintoa tai **paksuhäntäistä** todellista populaatiojakaumaa. Virheellinen havainto voi olla esimerkiksi seurausta virheestä aineiston kirjaamisessa tai se voi johtua virheestä kokeen toteutuksessa. Esimerkiksi kokeellisessa tutkimuksessa jollekulle subjektille on annettu väärä annos testattavaa lääkeainetta. Jakaumaa sanotaan usein paksuhäntäiseksi jos sen ääriarvojen todennäköisyydet eivät pienene tarpeeksi nopeasti tarkasteltaessa yhä äärimmäisempiä arvoja. Paksuhäntäiseksi saatetaan sanoa esimerkiksi jakaumaa, jonka tiheysfunktio vähenee hitaampaa kuin normaalijakauma kuten Kuvassa 4.11. Kuvan 4.11 populaatiojakaumista on kerätty 10 havaintoa, ja ollaan saatu kahden desimaalin tarkkuudella ilmoitetut arvot

Popuulatio 1	0.55	−0.28	1.78	0.19	1.14	0.42	1.23	0.24	−0.37	1.11
Popuulatio 2	−7.53	−3.36	−0.11	0.13	0.69	1.44	0.66	−40.86	−1.24	−0.60,

josta huomataan, että verrattuna normaalipopulaatioon, paksuhäntäisesti jakautuneella populaatiosta kerätyissä havainnoissa on suhteettoman suuria arvoja ($-7.53, -40.86$).



Kuva 4.11: Sinisellä piirrettyyn normaalijakaumaan verrattuna punaisella piirretyllä jakaumalla on paksut hännät.

Poikkeavat havainnot aiheuttavat usein hankaluuksia aineiston analysoinnissa, etenkin jos havaintoaineisto ei ole suuri kooltaan. Jos poikkeava havainto on virheellisen mittauksen seurausta, voidaan havainto jättää pois aineistosta, mutta jos aineistoa muokataan jättämällä pois havaintoja, tulee tämä aina dokumentoida hyvin. Havaintoaineiston muokkaamisesta helpommin tulkittavaksi ja analysoitavaksi pitää aina ilmoittaa. Näin pitää toimia, sillä ongelmana on, ettei usein pystytä varmasti sanomaan mistä poikkeava havainto johtuu. Jos poikkeava havainto kuuluu populaatioon, eikä havaittu arvo ole virhe, saattaa tämä havainto olla tieteellisesti hyvinkin tärkeä. Analysointimenetelmiä jotka eivät ole herkkiä poikkeaville havainnoille sanotaan **robusteiksi menetelmiksi**.

Kun havaintoaineistossa on poikkeavia havaintoja, on näiden vaikutus normaalijakautuneisuusoletukseen perustuviin testeihin hyvin voimakas. Ennen aineiston analysointia on hyvä kuvailevan tilastotieteen keinoin, esimerkiksi normaalikvantiilikuviolla sekä Tukeyn laatikko-janakuviolla tarkastella aineistoa, jos aineisto noudattaisi läheisesti normaalijakaumaa lukuunottamatta muutamia poikkeavia havaintoja. Jos mielenkiinto kohdistuu jakauman sijantiin, eikä muunnoksilla saada aineistoa normaalijakautuneeksi, voidaan analyysi suorittaa poikkeavien havaintojen poistamisen jälkeen. Mahdollisesti poistetut havainnot kuitenkin aina ilmoitetaan analyysin tuloksissa, esimerkiksi mainitsemalla, että tulokset pätevät siivotulle aineistolle, josta on poistettu poikkeavat havainnot x_i.

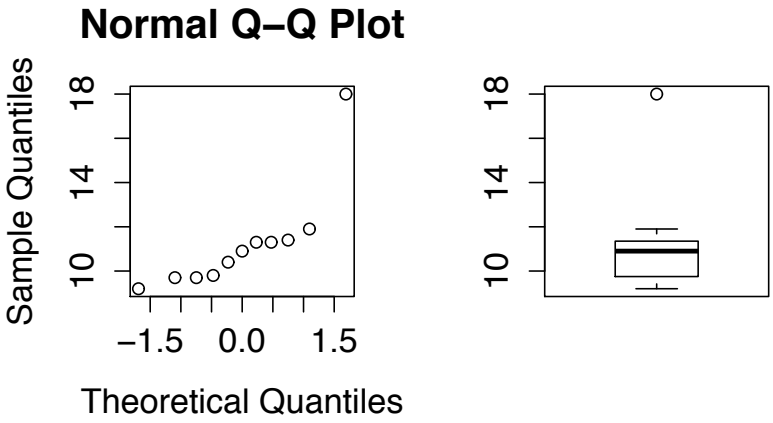
Esimerkki 52. Tarkastellaan havaintoaineistoa

$$\begin{array}{cccccc} 9.2 & 9.7 & 9.7 & 9.8 & 10.4 & 10.9 \\ 11.3 & 11.3 & 11.4 & 11.9 & 18.0, & \end{array}$$

ja halutaan testata hypoteesiparia $\mathsf{H}_0 : \mu = 12.5, \mathsf{H}_v : \mu \neq 12.5$. Aineiston normaalikvantiilikuvio sekä Tukeyn laatikko-janakuviot ovat piirrettynä Kuvaan 4.12. Tehdään testi tasolla 0.05 koko aineistolle ilman poikkeavien havaintojen siivoamista. Nyt kriittinen alue on $|t| \geq t^{(10)}_{0.025} = 2.228$, ja koska havaittu testisuureen arvo on nyt

$$\begin{aligned} \bar{x} &= 11.23636 \\ s^2 &= 5.796545 \\ t_{\text{hav}} &= \frac{\bar{x} - 12.5}{s/\sqrt{11}} \approx -1.74, \end{aligned}$$

niin nollahypoteesi jää voimaan tasolla $\alpha = 0.05$. P-arvolle voidaan hakea taulukosta rajat $2 \cdot 0.05 < \mathsf{p-arvo} < 2 \cdot 0.10$.



Kuva 4.12: Esimerkin 52 aineiston normaalikvantiilikuvio sekä Tukeyn laatikko-janakuvio.

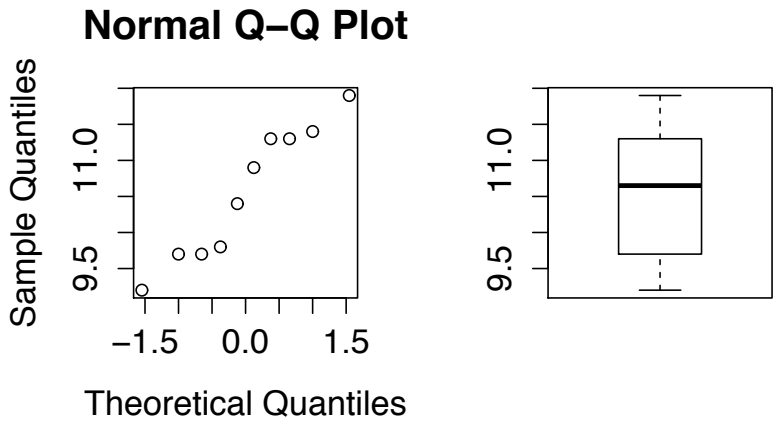
Aineistosta lasketut vinous ja huippukkuuskertoimet, joista nähdään ettei etenkään vinouskerrointa saada näillä muunnoksilla pieniksi.

Muunnos	Vinous	Huipukkuus
x	2.18	3.96
$x^{1/2}$	2.01	3.44
x^{-1}	-1.37	1.71
$\ln(x)$	1.81	2.88

Poistetaan aineistosta poikkeavaksi havainnoksi epäilty $x_{11} = 18$, ja tarkastellaan aineistoa uudelleen. Siivotun aineiston normaalikvantiilikuvio sekä Tukeyn laatikko-janakuvio on piirretty Kuvaan 4.13. Siivotulle aineistolle vinous- ja huipukkuuskertoimet ovat

$$g_1 \approx -0.06, \quad g_2 \approx -1.43.$$

Tehdään nyt testi jälleen tasolla $\alpha = 0.05$. Kriittinen alue on nyt muotoa (vapausasteita on siivotun



Kuva 4.13: Esimerkin 52 aineiston normaalikvantiilikuvio sekä Tukeyn laatikko-janakuvio havainnon $x_{11} = 18$ poistamisen jälkeen.

aineiston jäljiltä yksi vähemmän) $|t| \geq t^{(9)}_{0.025} = 2.262$. Havaittu testisuureen arvo voidaan nyt laskea

$$\begin{aligned} \bar{x} &= 10.56 \\ s^2 &= 0.8493333 \\ t_{\text{hav}} &= \frac{\bar{x} - 12.5}{s/\sqrt{10}} \approx -6.66, \end{aligned}$$

jonka perusteella nollahypoteesi hylätään tasolla 0.05, kun aineistosta on poistettu poikkeavat havainnot. P-arvolle voidaan hakea taulukosta raja $\mathsf{p-arvo} < 2 \cdot 0.0005$.

Esimerkiksi mediaani ja leikattu otoskeskiarvo ovat sijainnin robusteja tunnuslukuja, kun taas kvartiilipoikkeama ja pseudokeskihajonta ovat hajonnan robusteja tunnuslukuja.

Luku 5

Epäparametriset testit

Kun otoskoko on pieni, eikä voida tehdä oletusta parametrisista populaatiojakaumista, voidaan populaatioiden ominaisuuksia testata epäparametristen (parametrittomien) testien avulla. Myös epäparametrisissa menetelmissä oletetaan toki jotain jakaumista, mutta olettamukset eivät koske jakaumien parametreja.

5.1 Wilcoxonin merkityn järjestyksen testi

Oletetaan, että tarkasteltava jakauma on symmetrinen jonkin pisteen suhteen ja jatkuva. Kyseinen piste on jakauman mediaani m_d. Tarkastellaan hypoteesia, joka koskee jakauman tuntematonta mediaania, eli $\mathsf{H}_0 : m_d = m_0$. Kiinnostavat hypoteesiparit muotoillaan

$\mathsf{H}_0: \quad m_d = m_0$	$\mathsf{H}_0: \quad m_d = m_0$	$\mathsf{H}_0: \quad m_d = m_0$
$\mathsf{H}_v: \quad m_d < m_0$	$\mathsf{H}_v: \quad m_d > m_0$	$\mathsf{H}_v: \quad m_d \neq m_0.$

Kerätään havaintoaineisto $x_1, \ldots, x_n$. Lasketaan luvut

$$d_i = x_i - m_0, \quad i = 1, \ldots, n,$$

ja jätetään pois kaikki erotukset joille $d_i = 0$. Järjestetään jäljelle jääneet d_i:t (n_r kpl) suuruusjärjestykseen **itseisarvonsa** mukaisesti. Merkitään erotuksen itseisarvon $|d_i|$ suuruusjärjestysnumeroa r_i.

Erotus	d_1	d_2	d_3	$\cdots$	d_{n_r}
Itseisarvojärjestys	r_1	r_2	r_3	$\cdots$	r_{n_r}
Etumerkki	$+/-$	$+/-$	$+/-$	$\cdots$	$+/-,$

jossa $+/-$ valitaan d_i:n etumerkin mukaan. Jos lukujen d_i joukossa on itseisarvoltaan yhtäsuuria lukuja, joita kutsutaan **sidoksiksi**, niin jokaisen yhtäsuuren luvun järjestysnumeroksi asetetaan vastaavien järjestysnumeroiden keskiarvo, esimerkiksi jos $4.$ ja $5.$ suurimmat luvut ovat samoja, saavat molemmat luvut järjestysluvuksi $(4+5)/2 = 4.5$. Sidosten mukanaolo vaikuttaa Wilcoxonin testin jakaumaoletuksiin, mutta yksinkertaisuuden vuoksi jatketaan päättelyä sidoksista välittämättä. Mahdollisiksi testisuureiksi lasketaan positiivisiksi merkittyjen ja negatiivisiksi merkittyjen järjestyslukujen summa

$$V^+ = \text{ Positiivisiksi merkittyjen järjestyslukujen summa}$$
$$V^- = \text{ Negatiivisiksi merkittyjen järjestyslukujen summa}$$

Nollahypoteesin voimassa ollessa molemmille mahdollisille testisuureille pätee.

$$\mathsf{E}(V) = \mathsf{E}\left(V^+\right) = \mathsf{E}\left(V^-\right) = \frac{n_r(n_r+1)}{4}$$
$$\mathsf{Var}(V) = \mathsf{Var}\left(V^+\right) = \mathsf{Var}\left(V^-\right) = \frac{n_r(n_r+1)(2n_r+1)}{24}$$

Testisuure valitaan vastahypoteesin perusteella siten, että testisuureen havaitun arvon pienuus tukee vastahypoteesia.

- Vastahypoteesi $\mathsf{H}_v : m_d < m_0$, eli erotusmuuttujien D_i symmetrisen jakauman mediaani on nollaa pienempi. Tällöin positiiviseksi merkittyjen lukujen itseisarvojärjestysten summan pitäisi olla pienempi. Testisuureeksi valitaan siis V^+.
- Vastahypoteesi $\mathsf{H}_v : m_d > m_0$, eli erotusmuuttujien D_i symmetrisen jakauman mediaani on nollaa suurempi. Tällöin negatiiviseksi merkittyjen lukujen itseisarvojärjestysten summan pitäisi olla pienempi. Testisuureeksi valitaan siis V^-.
- Vastahypoteesi $\mathsf{H}_v : m_d \neq m_0$, eli erotusmuuttujien D_i symmetrisen jakauman mediaani poikkeaa nollasta. Tällöin joko negatiiviseksi tai positiiviseksi merkittyjen lukujen itseisarvojärjestysten summan pitäisi olla pieni. Testisuureeksi valitaan se kumman havaittu arvo on pienempi.

Testin p-arvo voidaan lukea taulukoiduista kertymäfunktion arvoista. Jos aineistossa on sidoksia, voidaan tämän kurssin puitteissa pyöristää havaittu testisuureen arvo lähimpään kokonaislukuun. Toispuolisille testeille pätee

$$\mathsf{p-arvo} = \mathsf{P}\left(V \leq v_{\text{hav}}\right),$$

ja kaksipuoliselle testille

$$\mathsf{p-arvo} = 2\mathsf{P}\left(V \leq v_{\text{hav}}\right).$$

Haettua p-arvoa verrataan testin tasoon. Jos $\mathsf{p-arvo}$ on pienempi kuin testin valittu taso, voidaan nollahypoteesi hylätä kyseisellä tasolla. Satunnaismuuttujan V kertymäfunktion arvoja on taulukoitu todennäköisyyteen 0.5 asti kun havaintojen määrä $n \leq 12$. Jos havaintoja on niin monta ettei taulukoiduista arvoista voida lukea p-arvoa testille, voidaan testi tehdä likimääräisesti z-testinä käyttäen testisuureena

$$Z = \frac{V - \mathsf{E}\left(V\right)}{\sqrt{\mathsf{Var}\left(V\right)}} \sim \mathsf{Normal}\left(0, 1\right).$$

Esimerkki 53. Erään akkukäyttöisen rikkaimurin akun latautumisaikaa tutkittiin. Latautumisaikoja kerättiin 10 rikkuimurin osalta, kun täysin tyhjä akku ladattiin täyteen varaukseen. Saatiin seuraavat havainnot (tunneissa)

$$1.5 \quad 2.2 \quad 0.9 \quad 1.3 \quad 2.0 \quad 1.6 \quad 1.5 \quad 2.0 \quad 1.2 \quad 1.7$$

Jakaumasta tiedetään vain, että se on jatkuva ja symmetrinen mediaanin suhteen. Testaan tasolla 0.05 pitäisikö paikkaansa valmistajan ilmoittama keskimääräinen latautumisaika 1.8 tuntia. Hypoteesipari on siis $\mathsf{H}_0 : m = 1.8, \mathsf{H}_v : m \neq 1.8$.

Lasketaan erotukset d_i

x_i	1.5	2.2	0.9	1.3	2.0	1.6	1.5	2.0	1.2	1.7
$x_i - 1.8$	-0.3	0.4	-0.9	-0.5	0.2	-0.2	-0.3	0.2	-0.6	-0.1
Itsearvojärjestys:	5.5	7	10	8	3	3	5.5	3	9	1
Etumerkki:	$-$	$+$	$-$	$-$	$+$	$-$	$-$	$+$	$-$	$-$

Nyt mahdollisia testisuureita ovat joko V^+ tai V^-, mutta koska positiiviseksi merkittyjen järjestyslukujen summa on piemempi, valitaan testisuureeksi V^+, jonka havaittu arvo on $v_{\text{hav}} = 7+3+3 = 13$. Koska havaintoaineistossa on sidoksia, voidaan hakea taulukoiduista arvoista ainoastaan likimääräisen p-arvon

$$\mathsf{p-arvo} = 2\mathsf{P}\left(V \leq v_{\text{hav}}\right) = 2 \cdot 0.08 = 0.16.$$

Useimmiten Wilcoxonin merkityn järjestyksen testiä käytetään parittaisten havaintojen erotuksen tarkasteluun, kun mielenkiinto kohdistuu kysymykseen, että onko muuttujien arvoissa systemaattista eroa. Siirrytään parittaisista havainnoista $x_1, y_1, x_2, y_2, \ldots, x_n, y_n$ erotusmuttujiin

$$d_i = x_i - y_i, \quad i = 1, \ldots, n,$$

jossa oletetaan, että erotukset d_i noudattavat symmetristä jakaumaa. Taas jätetään pois erotukset $d_i = 0$. Testi koskee erotusten jakauman mediaania ja testaaminen tapahtuu samalla tavalla kuin yllä.

Esimerkki 54. Seitsemän miehen painot (kg) ennen tupakoinnin lopettamista ja lopettamisen jälkeen olivat

Ennen:	67.5	80	71.5	69.5	52.5	88	93
Jälkeen:	70.5	81.5	71.5	68.5	55	88.5	95

Testataan havaintoaineiston perusteella tasolla 0.05 hypoteesia, että tupakoinnin lopettaminen aiheuttaisi miehillä yleensä painonnousua. Hypoteesipari koskee symmetriseksi oletetun jakauman mediaania m, $\mathsf{H}_0 : m = 0, \mathsf{H}_v : m > 0$.

Käytetään Wilcoxonin merkittyjen järjestyslukujen testiä. Lasketaan erotukset, itseisarvojärjestykset sekä merkitään lukujen etumerkit

Jälkeen - ennen (d_i) :	3	1.5	0	-1	2.5	0.5	2
Itseisarvojärjestys (r_i)	6	3	–	2	5	1	4
Etumerkki:	$+$	$+$		$-$	$+$	$+$	$+$

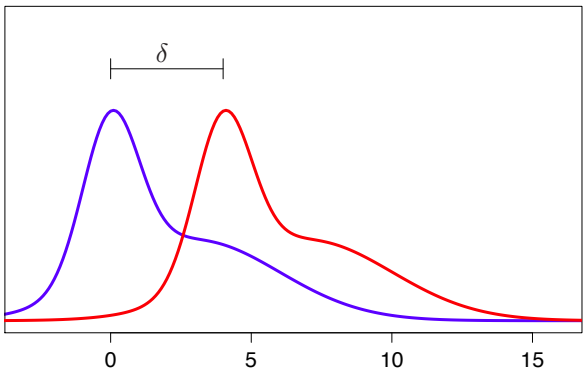
Koska ollaan kiinnnostuttu siitä onko jakauman mediaani nollaa suurempi (eli onko tupakoimisen lopettamisen jälkeen miehillä suurempi paino), niin kyseessä on toispuolinen testi ja negatiiviseksi merkittyjen järjestyslukujen summan pienuus antaa tukea vastahypoteesin suuntaan. Testisuureen havaittu arvo on $v_{\text{hav}} = v^- = 2$, sillä -2 on ainoa negatiiviseksi merkitty järjestysluku. Nyt voidaan hakea taulukoiduista p-arvoista kohdasta $v = 2, n = 6$ arvon $\mathsf{p-arvo} = \mathsf{P}(V \leq v_{\text{hav}}) = 0.047$. Tasolla 0.05 voitaisiin siis hylätä nollahypoteesin.

5.2 Mann-Whitneyn U-testi

Kun vertaillaan kahden populaation sijaintien eroa populaatioista kerättyjen riippumattomien otosten avulla ilman oletusta parametrisista jakaumista, voidaan käyttää Mann-Whitneyn U-testiä (tunnetaan myös nimellä Wilcoxonin järjestyssummatesti). Oletetaan, että populaatioiden jakaumat F_1 ja F_2 ovat samanmuotoisia. Muutoin samanmuotoisten jakaumien sijaintiero voidaan muotoilla

$$F_1(x) = F_2(x + \delta),$$

jota ollaan havainnollistettu Kuvassa 5.1. Kiinnostavat hypoteesiparit muotoillaan



Kuva 5.1: Kaksi samanmuotoista jakaumaa, joilla on δ suuruinen ero sijainnissa.

$\mathsf{H}_0 : \ \delta = 0$	$\mathsf{H}_0 : \ \delta = 0$	$\mathsf{H}_0 : \ \delta = 0$
$\mathsf{H}_v : \ \delta < 0$	$\mathsf{H}_v : \ \delta > 0$	$\mathsf{H}_v : \ \delta \neq 0.$

Kun $\delta < 0$, on jakauman F_2 sijainti jakauman F_1 vasemmalla puolella. Vastaavasti kun $\delta > 0$, niin jakauma F_2 sijaitsee jakauman F_1 oikealla puolella.

Tarkastellaan nyt tilannetta, jossa populaatiojakaumasta F_1 ollaan saatu havainnot $x_{1,1}, \ldots, x_{1,n_1}$ ja jakaumasta F_2 ollaan saatu havainnot $x_{2,1}, \ldots, x_{2,n_2}$. Haetaan yhdistetyille havaintojen arvoille $x_{1,1}, \ldots, x_{1,n_1}, x_{2,1}, \ldots, x_{2,n_2}$ järjestysluvut suuruuden mukaan $r_{1,1}, \ldots, r_{1,n_1}, r_{2,1}, \ldots, r_{2,n_2}$. Jos aineistossa on sidoksia, eli yhtäsuuria lukuja, annetaan näille kaikille vastaavien järjestyslukujen keskiarvo järjestysluvuksi, samoin kuin Wilcoxonin merkittyjen järjestyslukujen testissä.

Havainnot	$x_{1,1}$	$x_{1,2}$	$\cdots$	x_{1,n_1}		
Järjestysnumero	$r_{1,1}$	$r_{1,2}$	$\cdots$	r_{1,n_1}		
Havainnot	$x_{2,1}$	$x_{2,2}$	$\cdots$	x_{2,n_1}	$\cdots$	x_{2,n_2}
Järjestysnumero	$r_{2,1}$	$r_{2,2}$	$\cdots$	r_{2,n_1}	$\cdots$	r_{2,n_2}

Jossa ollaan oletettu, että $n_1 < n_2$.

Järjestysluvuista lasketaan arvot

$$w_1 = r_{1,1} + r_{1,2} + \ldots + r_{1,n_1}$$
$$w_2 = r_{2,1} + r_{2,2} + \ldots + r_{2,n_2},$$

jotka voidaan ennen havaintojen arvojen kiinnittämistä ajatella vastaavina satunnaismuuttujina W_1 ja W_2. Kätevyyssyistä siirrytään satunnaisluvuista W_1 ja W_2 satunnaislukuihin

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$
$$U_2 = W_2 - \frac{n_2(n_2+1)}{2},$$

joilla molemmilla on odotusarvo ja varianssi

$$\mathsf{E}(U_1) = \mathsf{E}(U_2) = \frac{n_1 n_2}{2}$$
$$\mathsf{Var}(U_1) = \mathsf{Var}(U_2) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Testisuureen U:n jakaumista on taulukoita kertymäfunktion alahäntää. Taulukoita käytettäessä testisuureeksi U valitaan testisuureista U_1 ja U_2 se, jonka pieni merkitsee poikkeamaa vastahypoteesin suuntaan.

1. Vastahypoteesi $F_1(x) = F_2(x+\delta), \delta > 0$, eli jakauma F_1 on sijoittunut jakauman F_2 vasemmalle puolelle. Tällöin otossuure U_1 pyrkii olemaan pieni ja U_2 pyrkii olemaan suuri, sillä U_1 on summa pienten arvojen järjestysluvuista ja U_2 on summa suurten arvojen järjestysluvuista. Valitaan testisuureeksi $U = U_1$.

2. Vastahypoteesi $F_1(x) = F_2(x+\delta), \delta < 0$, eli jakauma F_2 on sijoittunut jakauman F_1 vasemmalle puolelle. Tällöin otossuure U_1 pyrkii olemaan suuri ja U_2 pyrkii olemaan pieni, sillä U_1 on summa suurten arvojen järjestysluvuista ja U_2 on summa pienten arvojen järjestysluvuista. Valitaan testisuureeksi $U = U_2$.

3. Vastahypoteesi $F_1(x) = F_2(x+\delta), \delta \neq 0$, eli kumpi tahansa jakaumista F_1 tai F_2 on sijoittunut toisen jakauman vasemmalle puolelle. Tällöin jompikumpi otossuureista U_1 ja U_2 pyrkii olemaan pieni ja toinen pyrkii olemaan suuri, sillä toinen otossuureista U_1 tai U_2 on summa pienten arvojen järjestysluvuista ja toinen on summa suurten arvojen järjestysluvuista. Valitaan testisuureksi U testisuureista U_1 ja U_2 sen jonka havaittu testisuureen arvo on pienempi.

Testissä merkitsevät poikkeamat löytyvät siis aina testisuureen jakauman vasemmanpuolisesta hännästä. Toispuolisen testin p-arvo on

$$\mathsf{p-arvo} = \mathsf{P}(U \leq u_{\mathrm{hav}}),$$

ja kaksipuolisen testin p-arvo on vastaavasti

$$\mathsf{p-arvo} = 2\mathsf{P}(U \leq u_{\mathrm{hav}}).$$

Haettua p-arvoa verrataan testin tasoon. Jos $\mathsf{p}-\mathsf{arvo}$ on pienempi kuin testin valittu taso, voidaan nollahypoteesi hylätä kyseisellä tasolla. Jos havaintoja on niin monta ettei taulukoiduista arvoista voida lukea p-arvoa testille, voidaan testi tehdä z-testinä testisuureena

$$Z = \frac{U - \mathsf{E}(U)}{\sqrt{\mathsf{Var}(U)}} \sim \mathsf{Normal}(0,1).$$

Jos aineistossa on sidoksia, vaikuttavat nämä Mann Whitneyn U-testiin liittyviin todennäköisyysjakaumiin. Yksinkertaisuuden vuoksi kuitenkin suoritetaan päättely sidoksista huolimatta yllä kuvaillulla tavalla. Havaitut testisuureen arvot voidaan pyöristää lähimpään tasalukuun taulukoituja kertymäfunktion arvoja varten.

Esimerkki 55. Uuden opetusmenetelmän testaamiseksi valittiin koululuokalta kaksi ryhmää. Toiseen ryhmään sovellettiin uutta menetelmää (koeryhmä), toiseen vanhaa (kontrolliryhmä). Kokeen jälkeen pidetyn kuulustelun jälkeen tulokset olivat seuraavat (korkea pistemäärä merkitsee hyvää tulosta)

Ryhmä	Arvot						
Koe (populaatio 1)	32	38	41	43	49	54	55
Kontrolli (populaatio 2)	30	33	34	35	40		

Testataan tasolla 0.05 onko menetelmällä edullista vaikutusta. Hypoteesipari koskee jakaumien sijaintia $F_1(x) = F_2(x+\delta)$ ja on muotoa $\mathsf{H}_0 : \delta = 0, \mathsf{H}_v : \delta < 0$. Eli kontrolliryhmän arvosanajakauma sijaitsisi koeryhmän arvosanajakauman vasemmalla puolella (eli koeryhmä saisi useammin parempian arvosanoja). Muodostetaan järjestyslukutaulukko

Havainnot $(x_{1,i})$	32	38	41	43	49	54	55
Järjestysluku $(r_{1,i})$	2	6	8	9	10	11	12
Havainnot $(x_{2,i})$	30	33	34	35	40		
Järjestysluku $(r_{2,i})$	1	3	4	5	7		

Valitaan testisuureeksi U_2, sillä sen pieni arvo tukee vastahypoteesia. Lasketaan $w_2 = 1+3+4+5+7 = 20$, ja $u_2 = w_2 - \frac{1}{2}n_2(n_2+1) = 5$. Nyt $n_2 = 5 < n_1 = 7$, ja kertymäfunktiotaulukosta luetaan $\mathsf{p}-\mathsf{arvo} = 0.024$. Koska $\mathsf{p}-\mathsf{arvo} < 0.05$, niin tasolla 0.05 nollahypoteesi voidaan hylätä.

Luku 6

χ^2-testit

6.1 Riippumattomuustesti

Tarkastellaan tilannetta, jossa tutkittaviin tilastoyksiköihin liittyy kaksi tilastollista muuttujaa X ja Y, jotka ovat kategorisia tai diskreettejä numeerisia muuttujia. Molemmilla tilastollisilla muuttujilla on äärellinen määrä mahdollisia arvoja, ja molemmat muuttujista ovat vasteita. Nyt yhteisfrekvenssijakauma ja reunafrekvenssijakaumat voidaan esittää kontingenssitaulukon, eli ristiintaulukon muodossa. Olkoon X:n mahdolliset arvot $\{x_1, \ldots, x_r\}$ ja Y:n mahdolliset arvot $\{y_1, \ldots, y_s\}$. Nyt olkoon satunnaismuuttujien pistetodennäköisyysfunktiot

$$\begin{aligned} \mathsf{P}(X = x_i) &= p_i \\ \mathsf{P}(Y = y_j) &= q_j. \end{aligned}$$

Satunnaismuuttujat X ja Y ovat toisistaan riippumattomat, jos ja vain jos

$$\mathsf{P}(X = x_i, Y = y_j) = p_{i,j} = \mathsf{P}(X = x_i)\,\mathsf{P}(Y = y_j) = p_i q_j.$$

Halutaan testata ovatko tilastolliset muuttujat X ja Y toisistaan riippumattomat käyttäen havaintoaineistoa, joka voidaan ilmaista yhteisfrekvenssijakauman avulla. Yhteisfrekvenssijakauma on kontingenssitaulukon avulla ilmaistuna

		Y				
	y_1	$\ldots$	y_j	$\ldots$	y_s	
x_1	$f_{1,1}$	$\ldots$	$f_{1,j}$	$\ldots$	$f_{1,s}$	$f_{1,\cdot}$
	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X x_i	$f_{i,1}$	$\ldots$	$f_{i,j}$	$\ldots$	$f_{i,s}$	$f_{i,\cdot}$
	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	$f_{r,1}$	$\ldots$	$f_{r,j}$	$\ldots$	$f_{r,s}$	$f_{r,\cdot}$
	$f_{\cdot,1}$	$\ldots$	$f_{\cdot,j}$	$\ldots$	$f_{\cdot,s}$	n

Kontingenssitaulukossa $f_{i,j}$ on niiden havaintojen lukumäärä, jossa tilastolliset muuttujat saavat arvot x_i ja y_j. Tilastollisten muuttujien X ja Y (reuna)frekvenssijakaumat muodostuvat frekvensseistä $f_{i,\cdot}$ ja $f_{\cdot,j}$, ja ilmoittavat montako kertaa x_i esiintyy aineistossa ($f_{i,\cdot}$) ja y_j esiintyy aineistossa ($f_{\cdot,j}$). Havaintojen kokonaismäärä on n.

Tarkastellaan hypoteesiparia

$$\begin{aligned} \mathsf{H}_0 &: p_{i,j} = p_i q_j, \quad \forall i, j \\ \mathsf{H}_v &: \exists\, i, j \; p_{i,j} \neq p_i q_j. \end{aligned}$$

Nyt odotettu frekvenssi $\mathsf{E}(F_{i,j})$ tapahtumalle $(X = x_i, Y = y_j)$ on binomijakauman (multinomijakauman) mukaisesti $\mathsf{E}(F_{i,j}) = np_{i,j}$, joka taas voidaan nollahypoteesin ollessa voimassa ilmaista $np_i q_j$.

Todennäköisyyksiä p_i ja q_j estimoidaan suhteellisten frekvenssien avulla

$$\hat{p}_i = \frac{f_{i,\cdot}}{n}$$
$$\hat{q}_j = \frac{f_{\cdot,j}}{n},$$

jolloin voidaan estimoida frekvenssin odotusarvoa

$$e_{i,j} = n\hat{p}_i\hat{q}_j = \frac{f_{i,\cdot}f_{\cdot,j}}{n}.$$

Standardoidut jäännökset määritellään

$$d_{i,j} = \frac{f_{i,j} - e_{i,j}}{\sqrt{e_{i,j}}}.$$

Riippumattomuuden testaamiseen voidaan käyttään standardoitujen jäännösten neliösummaa

$$h = \sum_{i=1}^{r}\sum_{j=1}^{s} d_{i,j}^2.$$

Vastaavalla satunnaismuuttujalla

$$H = \sum_{i=1}^{r}\sum_{j=1}^{s} D_{i,j}^2$$

on nollahypoteesin ollessa voimassa Pearsonin approksimaation mukaisesti $\chi^2(v)$, jossa vapausasteet $v = (r-1)(s-1)$. Kriittinen alue testille löytyy jakauman $\chi^2(v)$ oikeasta hännästä, sillä poikkeamat riippumattomuusoletuksesta kasvattavat ylläolevaa neliösummaa. Joskus kirjallisuudessa mainitaan, että kaikkien odotettujen frekvenssien $f_{i,\cdot}f_{\cdot,j}/n$ pitäisi olla vähintään 5, jotta χ^2-approksimaatio olisi hyvä.

Esimerkki 56. Tarkastellaan tilannetta, jossa kolmelta eri linjalta valmistuu tuotteita (L_i) ja tuotteissa on kahden tyyppisiä virheitä (V_i). Halutaan tarkastella onko virhetyyppi ja linja toisistaan riippumattomia. Kerätään havaintoaineisto linjastoilla tapahtuneista erilaisten virheiden lukumäärästä

Virhetyyppi \ Linja	L_1	L_2	L_3	
V_1	15	21	45	81
V_2	26	31	34	91
	41	52	79	172

Testataan hypoteesiparia

$$\mathsf{H}_0 : p_{i,j} = p_i q_j, \quad \forall i,j$$
$$\mathsf{H}_v : \exists\, i,j\ p_{i,j} \neq p_i q_j.$$

tasolla 0.05. Testisuureella

$$H = \sum_{i=1}^{r}\sum_{j=1}^{s} D_{i,j}^2$$

on nollahypoteesin ollessa voimassa Pearsonin approksimaation mukaisesti $\chi^2(2)$, joten kriittinen arvo testille tasolla 0.05 on $q^{(2)}_{0.05} = 5.991$. Havaittu testisuureen arvo on

$$\begin{aligned} h_{\text{hav}} &= \frac{\left(15 - \frac{81\cdot 41}{172}\right)^2}{\frac{81\cdot 41}{172}} + \frac{\left(21 - \frac{81\cdot 52}{172}\right)^2}{\frac{81\cdot 52}{172}} + \frac{\left(45 - \frac{81\cdot 79}{172}\right)^2}{\frac{81\cdot 79}{172}} \\ &\quad + \frac{\left(26 - \frac{91\cdot 41}{172}\right)^2}{\frac{91\cdot 41}{172}} + \frac{\left(31 - \frac{91\cdot 52}{172}\right)^2}{\frac{91\cdot 52}{172}} + \frac{\left(34 - \frac{91\cdot 79}{172}\right)^2}{\frac{91\cdot 79}{172}} \\ &= 5.844, \end{aligned}$$

joka ei osu kriittiselle alueelle (kuitenkin hyvin lähelle kriittistä arvoa). Nyt tasolla 0.05 nollahypoteesi jää voimaan, eli virhetyyppi sekä linjasto ovat toisistaan riippumattomia.

6.2 Homogeenisuustesti

Tarkastellaan nyt tilannetta, jossa Y on edelleen vaste, mutta X on tekijä. Tekijä-vaste tilannetta tarkasteltaessa ei tarkastella riippuvuutta vaan muuttujien assiosiaatiota. Olkoon X:n mahdolliset arvot $\{x_1, \ldots, x_r\}$ ja koska x on tekijä, niin x:n jakauma on ennalta määrätty. Merkitään $f_{i,\cdot} = n_i$, jossa n_i on x_i:n frekvenssi. Nyt Y:n mahdolliset arvot ovat $\{y_1, \ldots, y_s\}$ ja voidaan ajatella, että havaintoaineisto koostuu r kappaleesta Y:n frekvenssijakaumia. Ollaan nyt kiinnostuneita, että ovatko nämä r kpl Y:n jakaumia samanlaisia, eli ovatko jakaumat **homogeenisia**.

		Y					
		y_1	$\cdots$	y_j	$\cdots$	y_s	
X	x_1	$f_{1,1}$	$\cdots$	$f_{1,j}$	$\cdots$	$f_{1,s}$	n_1
		$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	$f_{i,1}$	$\cdots$	$f_{i,j}$	$\cdots$	$f_{i,s}$	n_i
		$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	$f_{r,1}$	$\cdots$	$f_{r,j}$	$\cdots$	$f_{r,s}$	n_r
		$f_{\cdot,1}$	$\cdots$	$f_{\cdot,j}$	$\cdots$	$f_{\cdot,s}$	n

Homogeenisuuden testaaminen tapahtuu täsmälleen samoin kuin riippumattomuuden. Hypoteesipari on nyt (X on tässä muotoilussa tekijänä)

$$\mathsf{H}_0 : p_{1,j} = p_{2,j} = \cdots = p_{r,j}, j = 1, \ldots, s$$
$$\mathsf{H}_v : \exists\, i, j, k \text{ s.e. } p_{i,k} \neq p_{j,k}.$$

Oletetaan, että nollahypoteesi on voimassa. Merkitään nyt nollahypoteesin mukaisia todennäköisyyksiä

$$p_j = p_{1,j} = p_{2,j} = \cdots = p_{r,j}.$$

Nyt frekvenssin $F_{i,j}$ odotusarvo on

$$\mathsf{E}\,(F_{i,j}) = n_i p_j.$$

Tuntematonta todennäköisyyttä estimoidaan suhteellisten frekvenssien avulla

$$\hat{p}_j = \frac{f_{\cdot,j}}{n},$$

jolloin saadaan frekvenssin odotusarvoa $\mathsf{E}\,(F_{i,j})$ estimoivat odotetut frekvenssit

$$e_{i,j} = n_i \hat{p}_j = \frac{n_i f_{\cdot,j}}{n}.$$

Huomaa, että muoto on odotetuille jäännöksille täsmälleen sama kuin riippumattomuuden testauksessa. Nollahypoteesin ollessa voimassa testisuureella

$$H = \sum_{i=1}^{r}\sum_{j=1}^{s} D_{i,j}^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(\frac{F_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}\right)^2$$

on likimääräisesti $\chi^2(v)$ jakauma, jossa $v = (r-1)(s-1)$.

Esimerkki 57. Pesuaineiden markkinatutkimuksessa todettiin satunnaisesti valittujen kotitalouksien edustajien mielipiteen parhaasta pesuaineesta jakautuvan kolmella paikkakunnalla seuraavasti:

	Pesuaine			
	A	B	C	Otoskoko
Turku	232	108	60	400
Tampere	260	139	101	500
Lahti	197	106	97	400
	689	353	258	1300

Testataan onko mieltymykset pesuaineista samanlaisia eri paikkakunnilla. Testataan tasolla 0.05 hypoteesiparia

$$\mathsf{H}_0 : p_{1,j} = p_{2,j} = p_{3,j}, j = 1, 2, 3$$
$$\mathsf{H}_v : \exists\, i, j, k \text{ s.e. } p_{i,k} \neq p_{j,k}.$$

Nyt testisuureella on jakauma

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} D_{i,j}^2 \sim \chi^2(4),$$

ja kriittinen arvo testille on $q_{0.05}^{(4)} = 9.488$. Lasketaan odotetut frekvenssit ja taulukoidaan ne

	Pesuaine		
	A	B	C
Turku	$\frac{400\cdot 689}{1300}$	$\frac{400\cdot 353}{1300}$	$\frac{400\cdot 258}{1300}$
Tampere	$\frac{500\cdot 689}{1300}$	$\frac{500\cdot 353}{1300}$	$\frac{500\cdot 258}{1300}$
Lahti	$\frac{400\cdot 689}{1300}$	$\frac{400\cdot 353}{1300}$	$\frac{400\cdot 258}{1300}$

Havaittu testisuureen arvo on

$$\begin{aligned} h_{\text{hav}} = {} & \frac{\left(232 - \frac{400\cdot 689}{1300}\right)^2}{\frac{400\cdot 689}{1300}} + \frac{\left(108 - \frac{400\cdot 353}{1300}\right)^2}{\frac{400\cdot 353}{1300}} + \frac{\left(60 - \frac{400\cdot 258}{1300}\right)^2}{\frac{400\cdot 258}{1300}} \\ & + \frac{\left(260 - \frac{500\cdot 689}{1300}\right)^2}{\frac{500\cdot 689}{1300}} + \frac{\left(139 - \frac{500\cdot 353}{1300}\right)^2}{\frac{500\cdot 353}{1300}} + \frac{\left(101 - \frac{500\cdot 258}{1300}\right)^2}{\frac{500\cdot 258}{1300}} \\ & + \frac{\left(197 - \frac{400\cdot 689}{1300}\right)^2}{\frac{400\cdot 689}{1300}} + \frac{\left(106 - \frac{400\cdot 353}{1300}\right)^2}{\frac{400\cdot 353}{1300}} + \frac{\left(97 - \frac{400\cdot 258}{1300}\right)^2}{\frac{400\cdot 258}{1300}} \\ = {} & 11.85963, \end{aligned}$$

joka kuuluu kriittiselle alueelle. Nollahypoteesi voidaan siis hylätä tasolla 0.05. Havaintoaineiston perusteella siis eri kaupungeissa preferoidaan eri suhteessa pesuaineita A, B ja C.

Lähteet

[1] T. W. Anderson, *An Introduction to the Statistical Analysis of Data*, Houghton Mifflin Company, 1978

[2] R. B. Ash, *Basic Probability Theory*, Dover Publications, Inc., Mineola, New York

[3] M. Grönroos, *Johdatus tilastotieteeseen: Kuvailu, mallit ja päättely*, Oy Finn Lectura Ab, 2003, Helsinki.

[4] G. Casella, R. L. Berger, *Statistical Inference*, 2ed.

[5] K. Ruohonen, Tilastomatematiikka, Luentomoniste, TTY

[6] G. J. Kerns, *Introduction to Probability and Statistics Using R*, 2010.

[7] I. Mellin, *Todennäköisyyslaskenta: Todennäköisyys ja sen laskusäännöt*, Luentomoniste, TKK.

Liite A

TAULUKOITA

Standardoidun normaalijakauman kertymäfunktion Φ arvoja.

$$\mathsf{P}(Z \leq z) = \Phi(z), \quad \Phi(-z) = 1 - \Phi(z), \quad \Phi^{-1}(p) = -\Phi^{-1}(1-p)$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Jakaumien t(v) p-yläkvantiileja $t_p^{(v)}$ eri vapausasteilla v. $P(T \geq t_p^{(v)}) = p$, kun $T \sim \mathrm{t}[v]$

	p								
v	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
36	0.681	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
38	0.681	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
45	0.680	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
55	0.679	1.297	1.673	2.004	2.396	2.668	2.925	3.245	3.476
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
70	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
150	0.676	1.287	1.655	1.976	2.351	2.609	2.849	3.145	3.357
200	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340
300	0.675	1.284	1.650	1.968	2.339	2.592	2.828	3.118	3.323
500	0.675	1.283	1.648	1.965	2.334	2.586	2.820	3.107	3.310
∞	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Khin-neliö(v)-jakaumien p-yläkvantiileja $q_p^{(v)}$ eri vapausasteilla v.

$$P(Q \geq q_p^{(v)}) = p \ , \ \text{kun} \ \ Q \sim \chi^2(v)$$

	p								
v	0.99	0.975	0.95	0.1	0.05	0.025	0.01	0.005	0.001
1	0.000	0.001	0.004	2.706	3.841	5.024	6.635	7.879	10.828
2	0.020	0.051	0.103	4.605	5.991	7.378	9.210	10.597	13.816
3	0.115	0.216	0.352	6.251	7.815	9.348	11.345	12.838	16.266
4	0.297	0.484	0.711	7.779	9.488	11.143	13.277	14.860	18.467
5	0.554	0.831	1.145	9.236	11.070	12.833	15.086	16.750	20.515
6	0.872	1.237	1.635	10.645	12.592	14.449	16.812	18.548	22.458
7	1.239	1.690	2.167	12.017	14.067	16.013	18.475	20.278	24.322
8	1.646	2.180	2.733	13.362	15.507	17.535	20.090	21.955	26.124
9	2.088	2.700	3.325	14.684	16.919	19.023	21.666	23.589	27.877
10	2.558	3.247	3.940	15.987	18.307	20.483	23.209	25.188	29.588
11	3.053	3.816	4.575	17.275	19.675	21.920	24.725	26.757	31.264
12	3.571	4.404	5.226	18.549	21.026	23.337	26.217	28.300	32.909
13	4.107	5.009	5.892	19.812	22.362	24.736	27.688	29.819	34.528
14	4.660	5.629	6.571	21.064	23.685	26.119	29.141	31.319	36.123
15	5.229	6.262	7.261	22.307	24.996	27.488	30.578	32.801	37.697
16	5.812	6.908	7.962	23.542	26.296	28.845	32.000	34.267	39.252
17	6.408	7.564	8.672	24.769	27.587	30.191	33.409	35.718	40.790
18	7.015	8.231	9.390	25.989	28.869	31.526	34.805	37.156	42.312
19	7.633	8.907	10.117	27.204	30.144	32.852	36.191	38.582	43.820
20	8.260	9.591	10.851	28.412	31.410	34.170	37.566	39.997	45.315
21	8.897	10.283	11.591	29.615	32.671	35.479	38.932	41.401	46.797
22	9.542	10.982	12.338	30.813	33.924	36.781	40.289	42.796	48.268
23	10.196	11.689	13.091	32.007	35.172	38.076	41.638	44.181	49.728
24	10.856	12.401	13.848	33.196	36.415	39.364	42.980	45.559	51.179
25	11.524	13.120	14.611	34.382	37.652	40.646	44.314	46.928	52.620
26	12.198	13.844	15.379	35.563	38.885	41.923	45.642	48.290	54.052
27	12.879	14.573	16.151	36.741	40.113	43.195	46.963	49.645	55.476
28	13.565	15.308	16.928	37.916	41.337	44.461	48.278	50.993	56.892
29	14.256	16.047	17.708	39.087	42.557	45.722	49.588	52.336	58.301
30	14.953	16.791	18.493	40.256	43.773	46.979	50.892	53.672	59.703
31	15.655	17.539	19.281	41.422	44.985	48.232	52.191	55.003	61.098
32	16.362	18.291	20.072	42.585	46.194	49.480	53.486	56.328	62.487
33	17.074	19.047	20.867	43.745	47.400	50.725	54.776	57.648	63.870
34	17.789	19.806	21.664	44.903	48.602	51.966	56.061	58.964	65.247
35	18.509	20.569	22.465	46.059	49.802	53.203	57.342	60.275	66.619
36	19.233	21.336	23.269	47.212	50.998	54.437	58.619	61.581	67.985
37	19.960	22.106	24.075	48.363	52.192	55.668	59.893	62.883	69.346
38	20.691	22.878	24.884	49.513	53.384	56.896	61.162	64.181	70.703
39	21.426	23.654	25.695	50.660	54.572	58.120	62.428	65.476	72.055
40	22.164	24.433	26.509	51.805	55.758	59.342	63.691	66.766	73.402

Jakaumien F(v_1, v_2) 0.05-yläkvantiileja eri vapausasteilla v_1 ja v_2 .

	v_1											
v_2	1	2	3	4	5	6	7	8	9	10	11	12
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.11	2.08
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	2.09	2.06
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10	2.06	2.02
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08	2.04	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00

Wilcoxonin merkittyjen järjestyslukujen testisuureen kertymäfunktion arvoja $P(V \leq v)$ joillakin havaintojen lukumäärillä, kun havaintoaineistossa ei esiinny sidoksia.

	Havaintojen (arvoparien) lukumäärä n								
v	4	5	6	7	8	9	10	11	12
0	0.063	0.031	0.016	0.008	0.004	0.002	0.001	0.000	0.000
1	0.125	0.063	0.031	0.016	0.008	0.004	0.002	0.001	0.000
2	0.188	0.094	0.047	0.023	0.012	0.006	0.003	0.001	0.001
3	0.313	0.156	0.078	0.039	0.020	0.010	0.005	0.002	0.001
4	0.438	0.219	0.109	0.055	0.027	0.014	0.007	0.003	0.002
5	0.563	0.313	0.156	0.078	0.039	0.020	0.010	0.005	0.002
6		0.406	0.219	0.109	0.055	0.027	0.014	0.007	0.003
7		0.500	0.281	0.148	0.074	0.037	0.019	0.009	0.005
8			0.344	0.188	0.098	0.049	0.024	0.012	0.006
9			0.422	0.234	0.125	0.064	0.032	0.016	0.008
10			0.500	0.289	0.156	0.082	0.042	0.021	0.010
11				0.344	0.191	0.102	0.053	0.027	0.013
12				0.406	0.230	0.125	0.065	0.034	0.017
13				0.469	0.273	0.150	0.080	0.042	0.021
14				0.531	0.320	0.180	0.097	0.051	0.026
15					0.371	0.213	0.116	0.062	0.032
16					0.422	0.248	0.138	0.074	0.039
17					0.473	0.285	0.161	0.087	0.046
18					0.527	0.326	0.188	0.103	0.055
19						0.367	0.216	0.120	0.065
20						0.410	0.246	0.139	0.076
21						0.455	0.278	0.160	0.088
22						0.500	0.313	0.183	0.102
23							0.348	0.207	0.117
24							0.385	0.232	0.133
25							0.423	0.260	0.151
26							0.461	0.289	0.170
27							0.500	0.319	0.190
28								0.350	0.212
29								0.382	0.235
30								0.416	0.259
31								0.449	0.285
32								0.483	0.311
33								0.517	0.339
34									0.367
35									0.396
36									0.425
37									0.455
38									0.485
39									0.515

Mann-Whitneyn U-testisuureen kertymäfunktion arvoja P $(U \leq u)$ joillakin havaintojen lukumäärillä k_1 ja $k_2 \leq k_1$. Eli $k_1 = \max\{n_1, n_2\}, \quad k_2 = \min\{n_1, n_2\}$. Havaintoaineistossa ei esiinny sidoksia.

		k_1								
k_2	u	3	4	5	6	7	8	9	10	u
2	0	0.050	0.067	0.048	0.036	0.028	0.022	0.018	0.015	0
	1			0.095	0.071	0.056	0.044	0.036	0.030	1
	2						0.089	0.073	0.061	2
3	0	0.050	0.029	0.018	0.012	0.008	0.006	0.005	0.003	0
	1		0.057	0.036	0.024	0.017	0.012	0.009	0.007	1
	2			0.071	0.048	0.033	0.024	0.018	0.014	2
	3				0.083	0.058	0.042	0.032	0.024	3
	4						0.067	0.050	0.038	4
	5								0.056	5
4	0		0.014	0.008	0.005	0.003	0.002	0.001	0.001	0
	1		0.029	0.016	0.010	0.006	0.004	0.003	0.002	1
	2		0.057	0.032	0.019	0.012	0.008	0.006	0.004	2
	3			0.056	0.033	0.021	0.014	0.010	0.007	3
	4				0.057	0.036	0.024	0.017	0.012	4
	5					0.055	0.036	0.025	0.018	5
	6						0.055	0.038	0.027	6
	7						0.077	0.053	0.038	7
	8								0.053	8
5	0			0.004	0.002	0.001	0.001	0.000	0.000	0
	1			0.008	0.004	0.003	0.002	0.001	0.001	1
	2			0.016	0.009	0.005	0.003	0.002	0.001	2
	3			0.028	0.015	0.009	0.005	0.003	0.002	3
	4			0.048	0.026	0.015	0.009	0.006	0.004	4
	5			0.075	0.041	0.024	0.015	0.009	0.006	5
	6				0.063	0.037	0.023	0.014	0.010	6
	7					0.053	0.033	0.021	0.014	7
	8						0.047	0.030	0.020	8
	9						0.064	0.041	0.028	9
	10							0.056	0.038	10
	11								0.050	11
6	0				0.001	0.001	0.000	0.000	0.000	0
	1				0.002	0.001	0.001	0.000	0.000	1
	2				0.004	0.002	0.001	0.001	0.000	2
	3				0.008	0.004	0.002	0.001	0.001	3
	4				0.013	0.007	0.004	0.002	0.001	4
	5				0.021	0.011	0.006	0.004	0.002	5
	6				0.032	0.017	0.010	0.006	0.004	6
	7				0.047	0.026	0.015	0.009	0.005	7
	8				0.066	0.037	0.021	0.013	0.008	8
	9					0.051	0.030	0.018	0.011	9
	10						0.041	0.025	0.016	10
	11						0.054	0.033	0.021	11
	12							0.044	0.028	12
	13							0.057	0.036	13
	14								0.047	14
	15								0.059	15

(jatkoa) Mann-Whitneyn U-testisuureen kertymäfunktion arvoja joillakin havaintojen lukumäärillä k_1 ja $k_2 \leq k_1$. Eli

$$k_1 = \max\{n_1, n_2\}, \quad k_2 = \min\{n_1, n_2\}$$

	k_2										
	7				8			9		10	
	k_1				k_1			k_1		k_1	
u	7	8	9	10	8	9	10	9	10	10	u
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0
1	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1
2	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2
3	0.002	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	3
4	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	4
5	0.006	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	5
6	0.009	0.005	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000	6
7	0.013	0.007	0.004	0.002	0.003	0.002	0.001	0.001	0.000	0.000	7
8	0.019	0.010	0.006	0.003	0.005	0.003	0.002	0.001	0.001	0.000	8
9	0.027	0.014	0.008	0.005	0.007	0.004	0.002	0.002	0.001	0.001	9
10	0.036	0.020	0.011	0.007	0.010	0.006	0.003	0.003	0.001	0.001	10
11	0.049	0.027	0.016	0.009	0.014	0.008	0.004	0.004	0.002	0.001	11
12	0.064	0.036	0.021	0.012	0.019	0.010	0.006	0.005	0.003	0.001	12
13		0.047	0.027	0.017	0.025	0.014	0.008	0.007	0.004	0.002	13
14		0.060	0.036	0.022	0.032	0.018	0.010	0.009	0.005	0.003	14
15			0.045	0.028	0.041	0.023	0.013	0.012	0.007	0.003	15
16			0.057	0.035	0.052	0.030	0.017	0.016	0.009	0.004	16
17				0.044		0.037	0.022	0.020	0.011	0.006	17
18				0.054		0.046	0.027	0.025	0.014	0.007	18
19						0.057	0.034	0.031	0.017	0.009	19
20							0.042	0.039	0.022	0.012	20
21							0.051	0.047	0.027	0.014	21
22								0.057	0.033	0.018	22
23									0.039	0.022	23
24									0.047	0.026	24
25									0.056	0.032	25
26										0.038	26
27										0.045	27
28										0.053	28