Luku 1
Tilastollisia perusasioita

Seuraavassa kerrataan lyhyesti joitakin tilastollisia peruskäsitteitä - jotka ovat tietenkin periaatteessa kaikille tuttuja, mutta jatkon kannalta myös välttämättömiä. Enimmäkseen puhutaan tavallisista satunnaismuutujista – jokainen muuttujan arvo on ennalta arvaamaton, mutta lukujen jakauma pysyy ajan funktiona vakiona. Satunnais*prosessit* ovat sikäli hankalampia, että niissä myös muuttujien todennäköisyysjakaumat muuttuvat ajan funktiona.

Tähtitieteellisistä kohteista tehtävät johtopäätökset ovat useimmiten luonteeltaan tilastollisia ja perustuvat suuriin havaintoaineistoihin. Havaintojen analyysi edellyttää tilastollisten käsitteiden ja menetelmien tuntemusta. Erityisesti pitää tietää, milloin eri menetelmiä kannattaa tai ylipäätään voi soveltaa. Kaiken perusta on erilaisten todennäköisyysjakaumien tuntemus. Erityisempiin ongelmiin, esim. erilaiset valintaefektit ja niiden aiheuttamat virheet, ei varsinaisesti puututa.

1.1 Tilastolliset tunnusluvut

Seuraavassa merkitään $p(x)$ todennäköisyyttä, että satunnaismuuttujan arvo on välillä $[x, x + d\,x]$. Funktiota $p(x)$ kutsutaan myös kyseisen jakauman todennäköisyystiheysfunktioksi. Se on aina ei-negatiivinen ja sen integraali yli kaikkien mahdollisten parametriarvojen on yksi. Yleisimmin tarvittaviin todennäköisyysjakaumiin tutustutaan tarkemmin luvussa 1.2.

Satunnaismuuttujan x odotusarvo $E(x)$ määritellään integraalina

$$E(x) = \int x\, p(x) d\, x. \tag{1.1}$$

Tämän avulla voidaan määritellä n:s momentti (*ordinary moment*) μ'_n

$$\mu'_n = E(x^n) = \int x^n p(x) d\, x. \tag{1.2}$$

Keskiarvo keskiarvo on siis samalla satunnaismuuttujan ensimmäinen momentti. Korkeampien momenttien yhteydessä puhutaan tavallisimmin keskistetyistä momenteista μ_n, jotka määritellään

$$\mu_n = E([x - E\,(x)]^n). \tag{1.3}$$

Jakauman keskihajonnan, σ, neliö on varianssi. Se määritellään lausekkeena

$$\sigma^2 = E([x - E(x)]^2) = E(x^2) - (E(x))^2. \tag{1.4}$$

Keskihajonta kuvaa todennäköisyysjakauman leveyttä sen odotusarvon ympärillä. Seuraava eli kolmas momentti kuvaa jakauman vinoutta (*skewness*)

$$vinous = \frac{\mu_3}{\sigma^3} = \frac{E([x - E(x)]^3}{\sigma^3}. \tag{1.5}$$

Määritelmässä kolmas momentti on normitettu tekijällä σ^3. Symmetrisille jakaumille (esim. normaalijakauma) vinous on nolla. Neljäs momentti tekijällä σ^4 jaettuna on kurtosis,

$$kurtosis = \frac{\mu_4}{\sigma^4} = \frac{E([x - Ex]^4)}{\sigma^4}. \tag{1.6}$$

Kurtosis kuvaa sitä, kuinka voimakkaat hännät todennäköisyysjakaumassa on. Normaalijakauman kurtosis on edellisestä kaavasta laskettuna 3. Joskus kurtosis määritellään niin, että edellisen kaavan antamasta arvosta vähennetään 3. Näin normaalijakauman kurtosikseksi saadaan nolla.

Edellä puhuttiin jatkuvista todennäköisyysjakaumista. Tavallisesti joudutaan kuitenkin käsittelemään äärellistä joukkoa havaintoja, joita halutaan luonnehtia samoilla käsitteillä. Oletetaan siis, että on tehty N havaintoa satunnaimuuttujasta x. Muuttujan havaitut arvot ovat $x_1, x_2, ..., x_N$. Havaintoaineiston keskiarvo (odotusarvo) $\overline{x}$ on

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{1.7}$$

Tämä on paras estimaatti satunnaismuuttujan x todelliselle keskiarvolle, jos kaikkien mittauspisteiden paino on sama. Havaintoaineiston **moodi** on se piste, jossa funktio $p(x)$ saavuttaa maksiminsa. Moodi ei ole välttämättä lähellä odotusarvoa. **Mediaani** on arvo, jota pienempiä ja suurempia arvoja on havaittu yhtä paljon. Jos havaintoja on pariton määrä, mediaani on näistä suuruusjärjestyksessä keskimmäinen. Jos pisteitä on parillinen määrä, voidaan mediaaniksi määritellä kahden keskimmäisen pisteen keskiarvo.

Keskiarvo kiinnittää satunnaismuuttujan jakauman paikan. Jakauman leveyttä kuvataan keskihajonnalla, σ_x. Havainnoista saatavaa esimaattia merkitään s_x. Tämän neliö, eli havaintojen varianssi on

$$\begin{aligned} s_x^2 &= \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2 \\ &= \frac{1}{N-1} \left(\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2 / N \right). \end{aligned}$$

Kaavan jälkimmäinen muoto on sikäli hyödyllinen, että se sallii keskihajonnan laskemisen ilman, että havaintopisteitä tarvitsee käydä läpi kahta kertaa. Huomaa myös, että tässä ei ole käytetty keskihajonnalle merkintää σ: kreikkalaiset kirjaimet varataan tavallisesti kuvaamaan jakauman todellisia tunnuslukuja, ja s_x on ainoastaan havaintoaineistosta laskettu todellisen keskihajonnan estimaatti.

Havaintoaineiston keskistämisellä tarkoitetaan keskiarvon vähentämistä mitatuista arvoista. Studentisointi (standardointi) merkitsee yleisesti paikkatunnusluvun vähentämistä ja hajontatunnusluvulla jakamista - käytännössä useimmiten siis

$$x_{\mathrm{stud}} = \frac{x-\mu}{\sigma}\,. \tag{1.8}$$

Keskihajontaa käytetään tavallisimmin normaalijakauman yhteydessä (kts. alla). Se ei kerro vielä paljon itse jakauman muodosta. Tämän vuoksi (erityisesti jos jakauma poikkeaa selvästi normaalijakaumasta) kannattaa jakauman karakterisointiin käyttää kvartiileja (kvartiilipiste, 'neljännespiste'). Havaituista pisteistä 25% on pienempiä kuin ensimmäinen kvartiilipiste (Q_1) ja suurimmat 25% kuuluvat ylimpään kvartiiliin ($x \geqslant Q_3$). Toisen ja kolmannen kvartiilin erottava piste on sama kuin jakauman mediaani. Yleisemmin voidaan puhua *kvantiileista* tai *fraktiileista* (englanniksi myös *percentiles*) joita voidaan havainnoista määrittää suoraan yhtä monta kuin otoksessa on havaintopisteitä. Otosjakauman leveyttä voidaan kuvata esim. vaihteluvälillä (suurin arvo – pienin arvo), puoliarvoleveydellä tai kvartiilivälillä ($Q_3 - Q_1$). Kvartiilipoikkeama on puolet kvartiilivälistä.

Kolmanteen momenttiin liittyvä tilastollinen tunnusluku, vinous, saadaan laskettua keskiarvon $\overline{x}$ ja keskihajonnan σ avulla

$$vinous = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^3}{N\sigma^3}. \tag{1.9}$$

Vastaavasti havaintopisteiden jakauman häntien voimakkuutta kuvaava kurtosis on

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^4}{N\sigma^4}. \tag{1.10}$$

Kahden muuttujan välistä riippuvuutta voidaan kuvata näiden kovarianssilla. Tämä voidaan laskea kaavasta, joka on analoginen keskihajonnan kaavan kanssa

$$s_{xy} = Cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y}). \tag{1.11}$$

Asettamalla $x = y$ saadaan tuttu yhden muuttujan varianssin kaava. Jos kovarianssi lisäksi normitetaan x ja y keskihajonnoilla, saadaan muuttujien välinen korrelaatio

$$r_{xy} = \frac{s_{xy}}{s_x s_y}. \tag{1.12}$$

Tämä Pearsonin korrelaatiokerroin on auki kirjoitettuna

$$r_{\mathrm{xy}} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}\,. \tag{1.13}$$

Korrelaatiokertoimen arvo on välillä [-1,1]. Arvo +1 vastaa täydellistä lineaarista relaatiota, jossa y kasvaa samaa tahtia x:n kasvaessa. Korrelaatio on 0 jos muuttujat ovat täysin toisistaan riippumattomia. Kun y useimmiten pienenee x:n kasvaessa, korrelaatio on negatiivinen. Aina on muistettava, että korrelaatiokerroin testaa ainoastaan lineaarista riippuvuutta – muuttujien välillä voi olla täydellinen funktionaalinen riippuvuus (esim. toisen asteen yhtälö tai vaikkapa pisteet ympyrän kehällä) ja korrelaatiokerroin on silti nolla.

Useamman muuttujan tapauksessa voidaan muuttujat koota vektoriksi $[x_1, x_2, \ldots, x_N]^T$. Tässä x_i tarkoittavat siis eri muuttujia, eivät yhdestä muuttujasta tehtyjä havaintoja. Muttujien kovarianssit voidaan kirjoittaa neliömatriisiksi Σ. Matriisin diagonaalielementit ovat yksittäisten muuttujien keskihajonnan neliöitä σ_x^2 eli variansseja. Muut alkiot Σ_{ij} ovat muuttujien x_i ja x_j välisiä kovariansseja σ_{ij}.

Kovarianssimatriisin laskeminen on helppoa matriisioperaatioiden avulla. Merkitään havaintojen muodostamaa matriisia X. Tässä kukin rivi vastaa eri muuttujaa ja sarakkeilla ovat kyseisestä muuttujasta tehdyn havainnot. Matriisin sarakkeiden keskistämisen jälkeen kovarianssimatriisi saadaan kaavasta

$$S = \frac{1}{N-1} X^T X. \tag{1.14}$$

Jos kovarianssimatriisin jokainen sarake standardoidaan ennen kertolaskua, saadaan tulokseksi korrelaatiomatriisi. Sen diagonaalielementit ovat kaikki ykkösiä (korrelaatio muuttujan itsensä kanssa) ja muut elementit muuttujien x_i ja x_j välisiä korrelaatioita.

Keskiarvoa ja muita momentteihin perustuvia tunnuslukuja kutsutaan myös keskiluvuiksi. Vastaavasti kvartiilit ja mediaani ovat järjestystunnuslukuja ja moodi ja jakauman puoliarvoleveys tiheyslukuja.

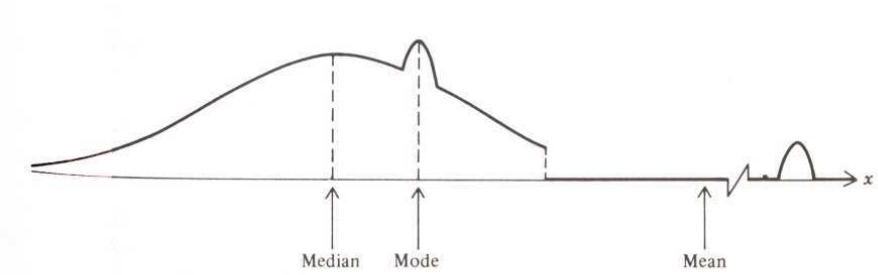


Figure 5.1-3. Measures of location for an altered $N(\mu, \sigma^2)$.

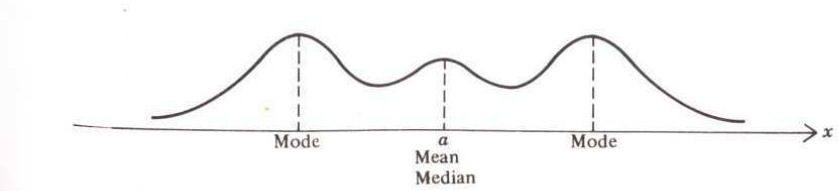


Figure 5.1-4.

Kuva 1.1. Esimerkkejä siitä, miten todennäköisyysjakauman keskiarvo, moodi ja mediaani voivat erota huomattavasti toisistaan (Kuva: Dudewicz & Mishra: Modern Mathematical Statistics)

Havaintoaineistossa saattaa esiintyä suuresti keskiarvosta poikkeavia arvoja, joko havaintovirheiden vuoksi (*outliers*) tai itse muuttujan jakauman vuoksi. Keskiarvon lasku on herkkä poikkeaville arvoille, joten esim. kohinaisen datan tapauksessa mediaani on luotettavampi tapa arvioida jakauman paikkaa. Hyvälaatuisesta aineistosta laskettu keskiarvo vaihtelee kuitenkin vähemmän kuin samoista mittauksista laskettu mediaani, eli sen virhe on pienempi. Tällöin kannattaa käyttää keskiarvoa. Jos aineistossa on mahdollisesti myös poikkeavia arvoja (=virheellisiä) voi keskiarvon arvioida ottamalla mukaan vain pisteet väliltä $[\overline{x} - k\delta x, \overline{x} + k\delta x]$, missä $\overline{x}$ ja δx ovat koko datasta laskettu keskiarvo ja keskihajonta ja k esimerkiksi 3. Näin laskettua arvoa kutsutaan Winsorin keskiarvoksi (*Winsorized mean*). Tätä voidaan käyttää myös iteratiivisesti: keskiarvo ja keskihajonta lasketaan uudestaan poikkeavien pisteiden eliminoinnin jälkeen, ja operaatiot toistetaan kunnes uusia hylättyjä pisteitä ei enää ole. *Trimmattu keskiarvo* voidaan toteuttaa hieman eri tavalla, poistamalla jakauman molemmista päistä sama määrä havaintoja. On myös kehitetty erityisiä suodattimia otoksen keskiarvon laskemiseksi silloin, kun jakauman tiedetään sisältävän keskiarvosta huomattavasti poikkeavia arvoja. *Myriad filtteri* on yksi esimerkki tällaisista, α - *stabiileista* suodattimista.

Esimerkki 1.1. Arvioidaan 100 luvun muodostaman havaintoaineiston jakauman keskikohtaa eri menetelmillä. Aineisto generoidaan toistuvasti (tuhat kertaa) normaalijakaumasta N(0,1). Jokaisesta generoidusta aineistosta lasketaan tilastolliset tunnusluvut, ja saadut arvot merkitään muistiin. Lopuksi tunnusluvuille itselleen laskettiin niiden keskiarvot, keskihajonnat sekä vaihteluvälit:

```
keskiarvo -0.00156 +- 0.10046    -0.34489 0.32998
mediaani  -0.00333 +- 0.12521    -0.39541 0.37671
trimmed   -0.00181 +- 0.10217    -0.33551 0.33673
windsor   -0.00135 +- 0.10524    -0.35525 0.32998
```

Edellä Winsorin keskiarvo on laskettu hylkäämällä yli 2.5 sigman verran poikkeavat arvot. Normaalijakauman tapauksessa tämä vastaa 1.24 prosenttia koko aineistosta. Huomaa mediaanin keskiarvoa (vain) hieman suurempi vaihtelu. 'Trimmed' tarkoittaa trimmattua keskiarvoa, joka on laskettu poistamalla pienimmät ja suurimmat arvot - 5% havaintoaineiston muodostaman jakauman molemmista päistä. Seuraavassa samankaltainen vertailu, kun luvut generoidaan Cauchyn jakaumasta, jolloin aineistossa on enemmän keskiarvosta selvästi poikkeavia arvoja. Luvut ovat jälleen peräisin jakaumasta, jonka μ=0.0.

```
keskiarvo -0.05694 +- 1.17191    -16.6339 11.0969
mediaani   0.00010 +- 0.01568     -0.0677  0.0540
trimmed   -0.00228 +- 0.03342     -0.2042  0.1166
windsor   -0.00205 +- 0.08650     -0.6152  1.6017
```

Tällä kertaa siis esim. mediaani toimii huomattavasti keskiarvoa paremmin, ja keskiarvoa käytettäessä on suuri riski saada selvästi 'oikeasta' arvosta poikkeava tulos.

Keskiarvo μ on luku, jolle keskineliöpoikkeama $\overline{(x-\mu)^2}$ on pienin, ja neliöllisestä luonteesta johtuen se on herkkä keskiarvosta paljon poikkeaville arvoille. Keskipoikkeama $\overline{|x-m|}$ on pienin puolestaa silloin, kun m on aineiston mediaani. Trimmattu keskiarvo, Winsorin keskiarvo ja mediaani ovat esimerkkejä tilastollisesti vankoista (*robust*) menetelmistä, jotka eivät riipu voimakkaasti virhejakauman muodosta eivätkä siten myöskään poikkeavista arvoista. Samaten keskihajonnan laskemisen sijasta voidaan käyttää *mad*-estimaattia (*Median Absolute Deviation*),

$$\mathrm{mad} = 1.4826\ \mathrm{mediaani}[\,|\,\mathrm{mediaani}(x) - x\,|\,]$$

(vrt. *R*-rutiini `mad`).

Edellä oletettiin, että kaikki havainnot ovat samanarvoisia. Käytännössä mittauksille voidaan usein arvioida erikseen niiden luotettavuus (so. virhearviot). Näissä tapauksissa pitää esim. aritmeettisen keskiarvon sijasta laskea painotettu keskiarvo (*weighted mean*) – trimmattu keskiarvo ja Winsorin keskiarvo voidaan yleistää myös näihin tapauksiin (pisteiden lukumäärän sijasta voidaan puhua esim. trimmattujen pisteiden suhteellisesta painosta). Painotettu keskiarvo on

$$\hat{x} = \frac{\sum_{i=1}^{N} w_i\, x_i}{\sum_{i=1}^{N} w_i}\,. \tag{1.15}$$

Painokertoimet voivat vaihdella, mutta jos havaintojen virheet ovat normaalijakautuneita, on oikea painokerrroin $w_i = [\sigma(x_i)]^{-2}$ – pisteen paino on kääntäen verrannollinen sen arvioituun varianssiin. Tämä voidaan yleistää usean muuttujan tapaukseen, jolloin varianssi korvautuu kovarianssimatriisilla ($\sigma^2 \to \Sigma$)

$$\hat{x} = \Sigma \sum_{i=1}^{N} \Sigma_i^{-1}\, x_i\ , \qquad \Sigma_{\hat{x}} = (\sum_{i=1}^{N} \Sigma_i^{-1})^{-1}\,.$$

Tässä Σ_i on i:nnen mittauksen kovarianssimatriisi ja $\Sigma_{\hat{x}}$ on keskiarvon kovarianssimatriisi.

Valintaefektien vuoksi havaintoaineisto saattaa sisältää esimerkiksi ainoastaan tiettyä rajaa suurempia arvoja. Havaitusta jakaumasta puuttuu siis toinen häntä ja esim. suoraan laskettu keskiarvossa on (*bias*-)virhettä. Tässä tapauksessa myös esim. Winsorin keskiarvo antaisivat luotettavampia arvoja. Valintaefektit voivat aiheuttaa paljon vaikeammin huomattavia virheitä, jolloin virheestä pääsee eroon ainoastaan mallintamalla havaittu jakauma huolellisesti ottaen (toivottavasti) tunnetut valintaefektit huomioon.

Valintaefektien käsittelyyn on kehitetty omia tilastollisia menetelmiään (terminä *censored* tai *truncated*). Esim. statlib (`lib.stat.cmu.edu`) sisältää algoritmeja normaalijakauman tunnuslukujen ja lineaaristen mallien laskemiseen silloin, kun tiedetään, että havaintoaineistosta puuttuu osa (esim. havainnot kattavat vain tietyn arvovälin). Vastaavia rutiineja on useissa internetistä löytyvissä *R*–kirjastoissa.

Eräs yleinen bias-efekti tulee esille kohteiden lukumääristä tehdyissä tutkimuksissa. Tarkastellaan tähtien tai galaksien lukumäärää taivaalla lähteiden kirkkauden funktiona. Tulokset esitetään histogrammina, jonka kukin pylväs kertoo tietyllä kirkkausvälillä havaittujen kohteiden lukumäärän. Himmeitä kohteita on enemmän kuin kirkkaita, joten pylväiden korkeus nousee ensin kirkkauden pienetessä. Lopulta kohteet ovat niin himmeitä, ettei niitä enää nähdä ja havaittu lähteiden lukumäärä kääntyy laskuun. Rajaa, jonka yläpuolella kaikki lähteet voidaan vielä havaita kutsutaan termillä 'täydellisyysraja' (*completeness limit*). Kaikissa kohteiden kirkkauden mittauksissa on virhettä, jonka suhteellinen arvo kasvaa kirkkauden pienetessä. Tarkastellaan kahta vierekkäistä kapeaa kirkkausintervallia. Mittauskohinan ansiosta tietty osa ylempään intervalliin kuuluvista kohteista tulee lasketuksi alempaan intervalliin ja päinvastoin. Koska alemmassa intervallissa on enemmän kohteita (ja ehkä suurempi mittausten epätarkkuus), on muutos suhteellisesti suurempi ylemmälle intervallille. Sen kohteiden lukumäärä kasvaa näennäisesti, ja saatu relaatio kohteiden kirkkauden ja niiden lukumäärän välillä loivenee. Efektin nimi on Malmquist-bias. Virhe on läsnä kaikissa intervalleissa, mutta erityisesti 'täydellisyysrajan' lähellä laskettu kohteiden määrä on liian suuri ja tulokset on korjattava. Malmquist-bias voidaan arvioida esim. simuloimalla lähteitä ja näistä tehtyjä havaintoja.

1.2 Todennäköisyysjakaumista

Tapahtuman a todennäköisyys määritellään havainnoista sen suhteellisena osuutena kaikista havainnoista, $P(a) = n(a)/n$. Varsinainen tilastollinen todennäköisyys on tämän raja-arvo, kun otoksen koko kasvaa äärettömän suureksi. Varsinainen tilastollinen todennäköisyys jää siis lopulta aina tuntemattomaksi, mutta suurempi havaintoaineisto antaa siitä paremman kuvan. Jos tapahtumat a ja b ovat täysin erillisiä, pätee näiden yhteiselle todennäköisyydelle

$$P(a \cup b) = P(a) + P(b). \tag{1.16}$$

Jos taasen on mahdollista, että a ja b tapahtuvat samanaikaisesti, on todennäköisyys sille, että vähintään toinen tapahtuu

$$P(a \cup b) = P(a) + P(b) - P(a \cap b). \tag{1.17}$$

Viimeinen termi on todennäköisyys molempien tapahtumien, a:n ja b:n, samanaikaiselle toteutumiselle. Todennäköisyys sille, että *riippumattomat* tapahtumat a ja b sattuvat samanaikaisesti on

$$P(a \cap b) = P(a)P(b). \tag{1.18}$$

Jos tapahtumat riippuvat toisistaan, on todennäköisyys

$$P(a \cap b) = P(a)\,P(b|a), \tag{1.19}$$

eli 'tapahtuman a todennäköisyys' kertaa 'tapahtuman b todennäköisyys, kun a on tapahtunut'. Tapahtumat a ja b ovat komplementtisiä, jos niistä toteutuu aina toinen, mutta vain toinen. Komplementtisille tapahtumille $P(a) = 1 - P(b)$.

Eri tapahtumien todennäköisyydet voidaan esittää todennäköisyysjakaumana. Jatkossa puhutaan lähinnä jatkuvista suureista, jolloin myös todennäköisyysjakauma on jatkuva funktio.

Todennäköisyysjakauma $p(x)$ määrittelee todennäköisyyden ($\mathcal{P}$), että satunnaismuuttujan x arvo on infinitesimaalisella välillä $d\,x$

$$\mathcal{P}(x_0 < x < x_0 + d\,x) = p(x)d\,x. \tag{1.20}$$

Tässä siis muuttuja x on jatkuva. Jos kyseessä on diskreetti suure, eri tapausten todennäköisyydet määrittelevät suoraan todennäköisyysjakauman. Merkintää $x \sim g(\mu)$ käytetään osoittamaan, että satunnaismuuttuja (havainto) noudattaa todennäköisyysjakaumaa g, jonka parametrit on lueteltu sulkeissa. Esimerkiksi merkintä $x \sim N(1, 0.5)$ kertoo, että x on normaalijakautunut, jakauman keskiarvo on 1 ja keskihajonta 0.5. Funktiota $p(x)$ kutsutaan todennäköisyyden tiheysfunktioksi. Koska x:llä oletetaan aina olevan jokin arvo, pätee kokonaistodennäköisyydelle

$$\int_{-\infty}^{\infty} p(x)d\,x = 1. \tag{1.21}$$

Tämä normitus sisältyy todennäköisyyden tiheysfunktion määritelmään. (Diskreeteille jakaumille integroinnin korvaa summaus eri tapahtumien yli.). Yleisesti todennäköisyys, että muuttujan arvo on annetulla välillä $[a,\ b]$ saadaan integroimalla tämän välin yli

$$\mathcal{P}(a < x < b) = \int_a^b p(x)d\,x. \tag{1.22}$$

Todennäköisyystiheyden kertymäfunktio $P(x)$ määritellään todennäköisyytenä, että satunnaismuuttujan arvo on pienempi kuin funktion argumentti. Se on siis integraali

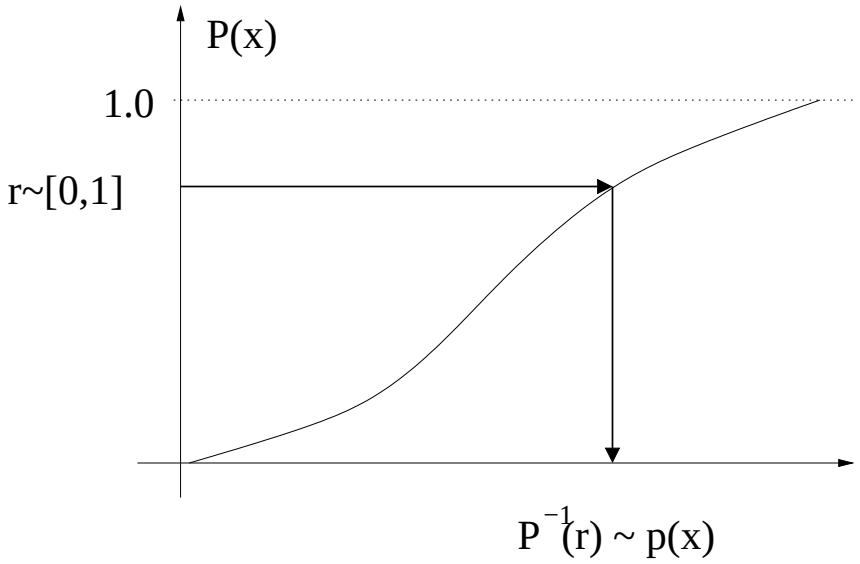
$$P(x) = \int_{-\infty}^{x} p(x')d\,x'. \tag{1.23}$$

Kertymäfunktiolle pätee aina $P(-\infty)$=0 ja $P(\infty)$=1. Edellinen todennäköisyys voidaan nyt esittää myös kertymäfunktion avulla

$$\mathcal{P}(a < x < b) = P(b) - P(a). \tag{1.24}$$

Diskreetin jakauman todennäköisyyden kertymäfunktio määritellään vastaavasti osasummana yli eri tapahtumien. Tämä edellyttää sovittua järjestystä eri tapahtumille.

Todennäköisyyden kertymäfunktio on erittäin hyödyllinen erilaisissa *simulaatioissa*, sillä sen avulla voidaan tuottaa satunnaislukuja annetusta jakaumasta. Menetelmässä generoidaan tasaisesti satunnaislukuja r välille [0,1]. Annetun jakauman mukainen satunnaismuuttujan arvo luetaan siitä x-akselin pisteestä, jossa kertymäfunktion arvo on r. Itse asiassa tarvitaan siis kertymäfunktion käänteisfunktio, jolloin generoidut satunnaismuuttujat ovat $x_i = P^{-1}(r_i)$. Jos käänteisfunktion laskeminen on hankalaa, voidaan käyttää taulukkoa kertymäfunktion arvoista ja lukea käänteisfunktion arvot siitä interpoloimalla.



Kuva 1.2. Käänteisfunktiomenetelmä satunnaislukujen generoimiseksi annetusta todennäköisyysjakaumasta. Satunnaisluvut r generoidaan tasaisesti välille [0,1], jolloin todennäköisyyden kertymäfunktion käänteisfunktion avulla lasketut arvot $P^{-1}(r)$ noudattavat jakaumaa $p(x)$.

Jos todennäköisyysjakauma tunnetaan, voidaan sille laskea keskiarvo, keskihajonta jne. kuten edellä kerrottiin. Esimerkiksi jakauman keskiarvo (satunnaismuuttujan x odotusarvo) on

$$\begin{aligned} \overline{x} &= \int_{-\infty}^{\infty} x\, p(x) d\, x \Big/ \int_{-\infty}^{\infty} p(x) d\, x \\ &= \int_{-\infty}^{\infty} x\, p(x) d\, x, \end{aligned}$$

sillä integraali todennäköisyysjakauman yli on aina yksi. Vastaavasti varianssi on

$$\sigma^2(x) = \int_{-\infty}^{\infty} (x - \overline{x})^2 p(x) d\, x, \tag{1.25}$$

ja jakauman keskihajonta on varianssin neliöjuuri, σ.

Todennäköisyys voi riippua useammasta kuin yhdestä muuttujasta. Yleisesti todennäköisyyden tiheysfunktio voi olla mikä tahansa ei-negatiivinen funktio, jonka integraali yli koko parametriavaruuden on yksi.

$$p(x, y, ...) \geq 0, \quad \iint ...p(x, y, ...)d\,x\,d\,y... = 1. \tag{1.26}$$

Moniulotteisesta todennäköisyysjakaumasta voidaan laskea *reunajakauma* integroimalla tiheysfunktio halutun muuttujan suhteen. Esimerkiksi, jos muuttuja saa tasaisesti arvoja neliössä $-1 < x < 1, -1 < y < 1$, on kaksiulotteinen todennäköisyyden tiheysfunktio $p(x, y)$=0.25 neliön sisällä ja nolla muualla. Mikäli halutaan tietää muuttujan riippuvuus ainoastaan parametristä x, integroidaan $p(x, y)$ muuttujan y suhteen. Tulos on tässä tapauksessa triviaalisti $p(x) = 0.5$ välillä $-1 < x < 1$ ja nolla ulkopuolella. Nähdään, että myös funktion $p(x)$ normalisointi on oikea. Muuttujan eliminoimista integroimalla kutsutaan myös todennäköisyysjakauman *marginalisoimiseksi*. Monen muuttujan todennäköisyysjakauma (yhteisjakauma) saadaan selville ainoastaan tekemällä samanaikaiset mittaukset kaikista muuttujista. Yksittäisen muuttujan jakauma saadaan joko marginalisoimalla moniulotteinen jakauma, tai suoraan mittaamalla ainoastaan kyseistä muuttujaa. Reunajakaumista ei voida päätellä yhteisjakauman muotoa. Yhteisjakauma voidaan kuitenkin laskea Bayesin kaavasta

$$p(x, y) = p(x)\,p(y|x), \tag{1.27}$$

missä jälkimmäinen termi on y:n *ehdollinen todennäköisyysjakauma* annetulla x:n arvolla. Tähän perustuu myös nk. Bayesilaisessa päättelyssä usein käytetty identiteetti

$$p(x)\,p(y|x) = p(y)p(x|y). \tag{1.28}$$

Tavallisin vastaan tuleva todennäköisyysjakauma on **normaalijakauma** $N(\mu, \sigma)$, joka on Gaussin kellokäyrän muotoinen

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1.29}$$

Normaalijakautuneita satunnaismuuttujia kutsutaankin myös gaussisiksi. Kaavassa μ on jakauman keskiarvo (odotusarvo) ja σ keskihajonta. Normaalijakauman vinous on nolla ja kurtosis 3 (tai nolla, määritelmästä riippuen!). Normaalijakauma on erityisasemassa *keskeisen raja-arvolauseen* (*central limit theorem*) ansiosta: kun lasketaan yhteen mistä tahansa todennäköisyysjakaumasta otettuja satunnaismuuttujia x_i, lähestyy summan jakauma normaalijakaumaa. Käytännössä jakaumat ovat kuitenkin harvoin *täsmälleen* normaalijakauman muotoisia, ja tämä mahdollisuus on otettava myös havaintojen analyysissä huomioon. Seuraavassa taulukossa on esitetty rajat, joiden välille osuu annettu prosenttiosuus kaikista tapahtumista silloin, kun todennäköisyysjakauma on $x \sim N(0, \sigma)$.

Tapausten osuus	muuttujan väli
68%	$\lvert x\rvert < 1\ \sigma$
95%	$\lvert x\rvert < 1.96\sigma$
99%	$\lvert x\rvert < 2.58\sigma$
99.9%	$\lvert x\rvert < 3.30\sigma$

Jos $x \sim N(\mu, \sigma)$, pätee vastaavalle standardoidulle muuttujalle

$$\frac{x-\mu}{\sigma} \sim N(0,1). \tag{1.30}$$

Yksiulotteisen normaalijakauman kertymäfunktiota merkitään usein erf(x). Kertymäfunktiolle ei ole yksinkertaista lauseketta, mutta sen likiarvo (absoluuttinen tarkkuus parempi kuin 0.005) voidaan laskea välille $0 \leqslant x \leqslant 2.2$ lausekkeesta

$$\mathrm{erf}(x) \approx \frac{x(4.4-x)}{10} + \frac{1}{2}. \tag{1.31}$$

Kaksiulotteinen normaalijakauma voidaan kirjoittaa

$$p(x,y) = \frac{1}{2\pi\sqrt{\det\Sigma}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}. \tag{1.32}$$

Tässä Σ on muuttujien x ja y välinen kovarianssimatriisi ja μ odotusarvojen muodostama vektori. Jos muuttujat x ja y eivät korreloi (matriisi Σ on diagonaalimatriisi) ovat jakauman pääakselit koordinaattiakselien suuntaiset. Jos muuttujat korreloivat, on jakauma vinossa koordinaattiakseleihin nähden. Kovarianssimatriisin determinantti on tässä tapauksessa

$$\det\Sigma = \left|\begin{pmatrix} \sigma_x^2 & \sigma_{\mathrm{xy}} \\ \sigma_{\mathrm{yx}} & \sigma_y^2 \end{pmatrix}\right| = \sigma_x^2\sigma_y^2 - \sigma_{\mathrm{xy}}\sigma_{\mathrm{yx}} = \sigma_x^2\,\sigma_y^2\,(1-\rho^2),$$

kun muuttujien välistä korrelaatiokerrointa merkitään ρ:lla. Pienellä vaivalla voidaan laskea myös eksponentissa esiintyvä matriisi Σ^{-1}, minkä jälkeen sijoitus antaa lausekkeen

$$N(\mu,\sigma) \sim \frac{1}{2\pi\,\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{(\frac{x-\mu_x}{\sigma_x})^2 - 2\,\rho\,(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2}{2(1-\rho^2)} \right]. \tag{1.33}$$

Useampiulotteisessa normaalijakaumassa on huomattavaa myös se, että sen kaikki reunajakaumat ovat myös normaalijakaumia. Edellinen kaava yleistyy usemman muuttujan tapaukseen ainoastaan normitustekijää muuttamalla,

$$p(X) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}, \tag{1.34}$$

missä n on muuttujien lukumäärä eli vektorin X komponenttien lukumäärä.

Toinen erittäin tärkeä todennäköisyysjakauma on χ^2–jakauma, joka määritellään kaavalla

$$\chi^2_\nu(0) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{(\nu/2)-1} e^{-x/2} \tag{1.35}$$

positiivisille x:n arvoille. Jakaumalla on yksi parametri, ν, joka on jakauman vapausasteiden lukumäärä. Todennäköisyys on 0 negatiivisille x:n arvoille. Jakaumalla on läheinen yhteys normaalijakaumaan. Esimerkiksi, jos $x \sim N(0,1)$, pätee $x^2 \sim \chi^2_1(0)$. Yleisestikin, studentoitujen normaalijakautuneiden satunnaislukujen neliösumma on χ^2–jakautunut satunnaismuuttuja, ja jakauman vapausasteiden lukumäärä on yhteenlaskettujen termien lukumäärä. χ^2-jakauma esiintyy usein mm. testattaessa mallien sopivuutta havaintoihin. Erityisesti, jos sovituksen jäännöspoikkeamien oletetaan olevan gaussisia, noudattaa jäännöspoikkeamien neliösumma χ^2–jakaumaa. Usein tarvittava χ^2-jakauman kertymäfunktio arvot löytyvät jälleen useista tilastotieteen taulukkokirjoista ja tilastollisista ohjelmistoista. Arvot voi myös laskea kaavasta

$$Q(\chi^2|\nu) = \frac{1}{\Gamma(\nu/2)}(\chi^2/2)^{\nu/2-1}e^{-\chi^2/2} \tag{1.36}$$

Tässä todetaan vielä kahden jatkuvan todennäköisyysjakauman, F–jakauman sekä t–jakauman olemassaolo. Jakaumafunktiot löytyvät useista taulukkokirjoista (sekä esim. R–ohjelmasta). F– ja t– jakaumiin palataan tilastollisia testejä käsittelevässä luvussa.

Oletetaan, että on havaittu diskreettejä tapahtumia, joiden väliajan odotusarvo Δt on vakio. Aikavälillä T havaittu tapahtumien lukumäärä noudattaa tällöin **Poisson-jakaumaa**. Merkitään tapahtumien määrän odotusarvoa $\mu = T/\Delta t$, jolloin Poisson-jakauman todennäköisyystiheysfunktio on

$$p(k)_{Poisson} = \frac{\mu^k}{k!}e^{-\mu}. \tag{1.37}$$

Jakauman odotusarvo on suoraan parametri μ, ja jakauman keskihajonta on $\sqrt{\mu}$. Poisson-jakauma on nollasta poikkeava ainoastaan ei-negatiivisilla arvoilla (tapahtumien lukumäärä). Kyseessä on diskreetti jakauma ja k on aina kokonaisluku. Jakauma on sitä selvemmin epäsymmetrinen, mitä lyhyempi aikaväli on. Jos μ on riittävän pieni, havaitaan aina vähintään nolla tapausta, mutta on myös pieni todennäköisyys sille, että annetulle aikavälille sattuu huomattavasti suurempi määrä tapahtumia. Pidempien aikavälien (suurempi μ) tapauksessa epäsymmetria pienenee, ja jakauman muoto alkaa muistuttaa normaalijakaumaa. Tämä on selvää myös keskeisen raja-arvoteoreeman perusteella: pitkä aikaväli koostuu joukosta lyhyitä aikavälejä, joita vastaavat satunnaismuuttujat lasketaan yhteen. Summan jakauma lähestyy asymptoottisesti normaalijakaumaa $N(\mu, \sqrt{\mu})$. Tyypillinen esimerkki Poisson-jakaumasta on laskettujen fotonien määrä vaikkapa radioaktiivisen säteilyn ilmaisimessa tai CCD-kamerassa. Toisaalta myös diskreettien kohteiden (esim. tähtien) havaittu lukumäärä pienessä taivaan osassa voi noudattaa likimain Poisson-jakaumaa.

Binomijakauma on toinen esimerkki diskreetistä jakaumasta. Oletetaan, että on olemassa todennäköisyys p että tietyllä kohteella on annettu ominaisuus. Havaitaan kaikkiaan n kohdetta. Todennäköisyys sille, että havaituista N kohteesta k:llä on kyseinen ominaisuus on

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1.38}$$

Tämä yhtälö määrittelee binomijakauman todennäköisyystiheysfunktion. Jakauma on määritelty vain ei-negatiivisille kokonaisluvuille k. Binomijakauman odotusarvo on $E(k) = np$, keskihajonta $\sigma(k) = \sqrt{np(1-p)}$ ja kertymäfunktio

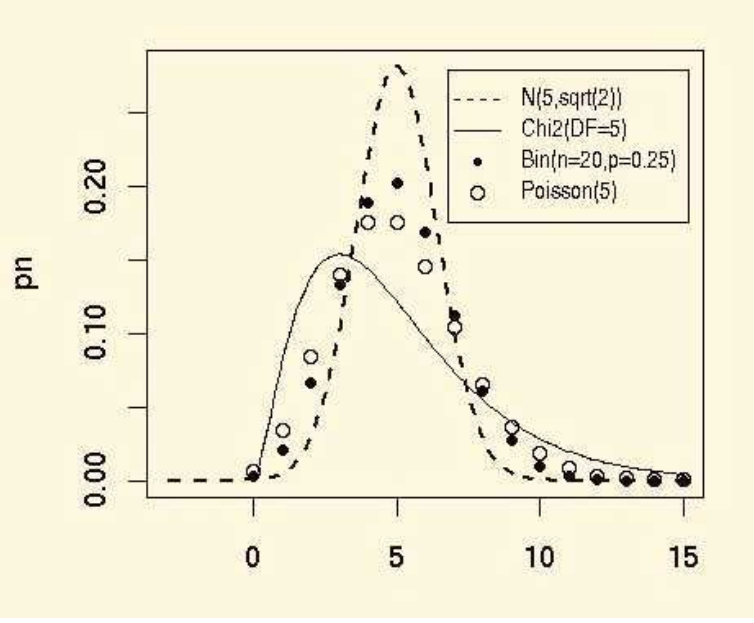
$$P(k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}. \tag{1.39}$$

Otoksen kasvaessa myös binomijakauman muoto lähestyy normaalijakaumaa – tällä kertaa $N(np, \sqrt{\mathrm{np}(1-p)})$. Jakauma on tietysti määritelty edelleen ainoastaan kokonaisluvuille.

Esimerkki 1.2. Havaitaan 1000 tähteä, joista 17 todetaan O-spektriluokan edustajiksi. Mikä on todennäköisyys, että todellisuudessa O-tähtien suhteellinen lukumäärä on alle 1% ? Kysytty todennäköisyys on tapausten 'p<1% ja $k = 17$' suhde tapahtumien 'k=17 ja $p \in [0,1]$' lukumäärään, eli

$$\int_0^{0.01} p(k=17\,|\,p')\,\mathrm{d}p' \;/\; \int_0^{1} p(k=17\,|\,p')\,\mathrm{d}p'\,.$$

Tässä $p(k|p)$ periaatteessa samaa kuin yhtälö 1.38, mutta muuttujana onkin p – oikeampi merkintä olisi $p(p|k = 17)$. (Huom. pelkästään muuttujaa vaihtamalla saatu uusi tiheysfunktio ei olisi välttämättä oikein normitettu - tässä skaalatekijä kuitenkin häviää osamäärää laskettaessa). Millä oletuksilla operaatio on oikeutettu (vrt. kaava 1.28).



Kuva 1.3. Todennäköisyystiheysfunktioita, joiden odotusarvo on 5. Kuvassa ovat normaalijakauma N(5,$\sqrt{5}$), χ^2_5-jakauma, binomijakauma (otos 20, p=0.25) sekä Poisson jakauma ($\mu = 5$). Näistä ainoastaan normaalijakauma on symmetrinen.

Esimerkki 1.3. Yksinkertaistenkin todennäköisyyslaskujen kanssa on syytä olla huolellinen, sillä intuitio voi helposti johtaa harhaan. Tästä esimerkkinä kaksi probleemaa, joissa on helppo päätyä väärään ratkaisuun. (Esimerkit on otettu *NPR*:n *Car Talk*-radio-ohjelmasta, joka tuskin on kuitenkaan alkuperäinen referenssi).

- Silmänkääntäjällä M on pussissa kolme korttia. Ensimmäisen kortin molemmat puolet ovat vihreät, toisen kortin molemmat puolet ovat punaiset, ja kolmannen kortin toinen puoli on vihreä ja toinen punainen. M ottaa pussista yhden kortin ja asettaa sen pöydälle. Kortin yläpuoli on vihreä. M sanoo, "Lyön vetoa 6 vastaan 5 että myös kortin toinen puoli on vihreä!". Kannattaako vetoon suostua?

- Oletetaan, että lapsen syntyessä on yhtä todennäköistä että lapsi on tyttö kuin että se on poika. Perheessä X on kaksi lasta, joista *toinen* on poika. Mikä on todennäköisyys, että myös toinen lapsista on poika.

1.3 Virhearvioiden laskemisesta

Seuraavassa tarkastellaan, miten arvioidaan lasketun suureen tilastollinen virhe, kun lähtöarvojen virheet tunnetaan. Olkoon x_i jokin havaittu tai laskettu suure ja $\sigma(x_i)$ sille annettu virhearvio. Aluksi oletetaan, että virheet noudattavat normaalijakaumaa, jolloin virhejakaumat voidaan määritellä yksikäsitteisesti käyttäen suureen keskihajontaa. Esimerkiksi merkintä 1.80 ± 0.15 tarkoittaa, että suureen odotusarvo on 1.80 ja esim. suureesta tehdyissä mittauksissa saatujen arvojen keskihajonta on 0.15. Virheraja ei suinkaan tarkoita, etteikö ilmoitettu arvo voisi olla enemmänkin väärässä. Normaalijakaumassa todennäköisyys, että arvo on enemmän kuin keskihajonnan verran arvioitua suurempi on n. 16% - kaikkiaan siis miltei kolmasosassa tapauksista oikea arvo on ilmoitetun välin ulkopuolella! Seuraavassa taulukossa on lueteltu todennäköisyydet erisuuruisille poikkeamille. Jos erillisiä datapisteitä on luokkaa sata kappaletta, ei edes 3σ poikkema yksittäisen pisteen kohdalla ole enää mitenkään yllättävä.

$x>\sigma$	15.9%
$x>2\sigma$	2.3%
$x>3\sigma$	0.13%

$\|x\|>\sigma$	31.7%
$\|x\|>2\sigma$	4.6%
$\|x\|>3\sigma$	0.27%

Jos suureen jakauma on selvästi vino, ei yhden virhearvion ilmoittaminen riitä. Yksi vaihtoehto on ilmoittaa virherajat erikseen ylös- ja alaspäin: $1.80^{+0.11}_{-0.20}$ – keskiarvo on 1.80 mutta on todennäköisempää, että todellinen arvo on selvästi keskiarvoa pienempi kuin selvästi sitä suurempi. Annetut rajat määrittelevät suureen *luotettavuusvälin*. Rajat voidaan valita niin, että todellinen arvo on 68% todennäköisyydellä annetussa välissä (olettaen, että todennäköisyysjakauma tunnetaan

riittävän hyvin). Näin merkintä on yhteensopiva tavallisen virherajamerkinnän kanssa. Vaihtoehtoisesti voidaan ilmoittaa esim. 90% tai 95% luotettavuusväli. Tässä tapauksessa pitää tietysti muistaa erikseen kertoa, mistä luotettavuusvälistä on kyse. Vielä yksi vaihtoehto on ilmoittaa virheeksi suoraan kvartiiliväli, jonka sisään jää 50% havainnoista. Eräs keino epäsymmetrisen virhevälin laskemiseksi on tehdä ensin muunnos (esim log, exp, $\sqrt{\ }$), niin että virhejakauma palautuu lähelle normaalijakaumaa. Muunnetulle aineistolle lasketaan keskiarvo ja keskihajonta, ja näin lasketut arvot $\mu - \sigma$, μ ja $\mu + \sigma$ muunnetaan takaisin alkuperäiselle asteikolle

Seuraavaksi lasketaan virhearviot lausekkeelle, jossa esiintyvien suureiden virhearviot tunnetaan. Tässä virhearviot tarkoittavat tilastollisia (statistisia) virheitä, joiden oletetaan olevan toisistaan riippumattomia eri tapauksissa, esim. perättäisissä mittauksissa.

Summan tai erotuksen virhe on sen termien statististen virheiden neliösumman neliöjuuri.

$$\sigma(\sum_{i=1}^{N} x_i) = \sqrt{\sum_{i=1}^{N} \sigma(x_i)^2}. \tag{1.40}$$

Tämä vastaa vektorin pituutta N-ulotteisessa avaruudessa. Kaava pätee ainoastaan, jos muuttujat eivät korreloi (N-ulotteisen avaruuden akselit ovat ortogonaaliset). Jos muuttujat korreloivat täydellisesti, on summan virhe suoraan $\sum \sigma(x_i)$ (muuttujien virittämä N-ulotteinen avaruus degeneroituu yksiulotteiseksi). Vastaavasti, jos muuttujien välinen korrelaatio on negatiivinen, muuttujien virheet kumoavat osittain toisensa ja kaava (1.40) yliarvioi todellisen virheen.

Muuttujien tulon tai osamäärän tapauksessa tuloksen suhteellinen virhe r on suhteellisten virheiden neliösumman neliöjuuri

$$r_{\Pi x_i} = \sqrt{\sum_{i=1}^{N} r_i^2}. \tag{1.41}$$

Tässä siis suhteelliset virheet ovat $r_i = (\sigma_i / x_i)$. Kaava pätee jälleen ainoastaan toisistaan riippumattomille virheille.

Jos muuttujien välisistä korrelaatioista on jonkinlaista tietoa, kannattaa sitä käyttää myös virheitä laskettaessa. Esimerkiksi kahden luvun summan tapauksessa todellinen virhe saadaan kaavasta

$$\sigma_{x+y}^2 = \sigma_x^2 + 2\,\sigma_{\mathrm{xy}} + \sigma_y^2\,,$$

jossa esiintyy myös muuttujien välinen kovarianssi σ_{xy}. Jos korrelaatio on nolla, summataan virheet neliössä yhteen, kuten edellä. Vastaavasti korrelaation ollessa täydellinen

$$\sigma_{x+y}^2 = \sigma_x^2 + 2\,\sigma_x \sigma_y + \sigma_y^2 = (\sigma_x + \sigma_y)^2,$$

eli virheet lasketaan suoraan yhteen. Vaikka kovarianssin tarkka arvo olisi tuntematon, voi kaavassa käyttää vaikkapa mutu-arvoa - ristitermin poisjättäminenhän johtaa tavallisesti enemmän väärään virhearvioon.

Monen muuttujan tapauksessa odotusarvo on vektori (μ) ja kovarianssit muodostavat matriisin (Σ). Yleisen lineaarisen muunnoksen, $y = L\,x$, tapauksessa pätee

$$\mu_y = L\,\mu_x\,, \quad \Sigma_y = L\,\Sigma_x\,L^T,$$

eli uuden muuttujan keskiarvo ja keskihajonta saadaan yksinkertaisilla matriisikertolaskuilla.

Esimerkki 1.4. Usein joudutaan arvioimaan virherajat painotetulle keskiarvolle

$$\mu = \sum_{i=1}^{N} w_i\,x_i \;/\; \sum_{i=1}^{N} w_i.$$

Tässä jokaiselle havaintopisteelle on annettu oma paino, w_i. Virhearvio voidaan johtaa seuraavaan tapaan

$$\sigma_\mu = \sigma(\frac{\sum w_i x_i}{\sum w_i}) = \frac{1}{\sum w_i}\sqrt{\sum \sigma^2(w_i x_i)} = \frac{\sqrt{\sum w_i^2\,\sigma_i^2}}{\sum w_i}\,.$$

Erityisesti, jos pisteet painotetaan tekijällä σ_i^{-2} (vrt. kaava 1.15), saadaan virhearvioksi

$$\sigma_\mu = \sqrt{\sum \sigma_i^{-2}} \;/\; \sum \sigma_i^{-2} = (\sum \sigma_i^{-2})^{-1/2}.$$

Jos kaikkien pisteiden virhearviot ovat yhtäsuuret, $\sigma_i = \sigma_0$, seuraa tästä virheen lauseke

$$\sigma_\mu = (\sum \sigma_i^{-2})^{-1/2} = (N\sigma_0^{-2})^{-1/2} = \frac{\sigma_0}{\sqrt{N}}$$

– normaali keskiarvon virheen kaava.

Epälineaarisessa tapauksessa lauseke voidaan linearisoida, jolloin lausekkeen $f(x_1, x_2, ...)$ virheeksi yksittäisen muuttujan suhteen saadaan

$$\frac{\partial f(\overline{x}\,)}{\partial x_i}\,dx_i\,, \tag{1.42}$$

ja tämän avulla voidaan periaatteessa laskea virhearviot mille tahansa lausekkeelle. On kuitenkin muistettava, että linearisointi pätee ainoastaan infinitesimaaliselle välille. Ei ole mitään takeita, että näin saadut virhearviot (esim. 67% luotettavuusvälin päätepisteet) olisivat edes oikeaa suuruusluokkaa. Halutessaan voi tietenkin ottaa mukaan lisää termejä virheen Taylorin kehitelmästä yksittäisen muuttujan suhteen

$$\frac{\partial f(\overline{x}\,)}{\partial x_i}\,dx_i + \frac{1}{2}\,\frac{\partial^2 f(\overline{x}\,)}{\partial x_i^2}\,(dx_i)^2 + \mathcal{O}((dx_i)^3),$$

mutta tämä on tavallisesti liian hankalaa. Lausekkeen kokonaisvirhettä laskettaessa on muistettava ottaa mukaan tietysti myös ristitermit $\frac{\partial^2 f}{\partial x_i \partial x_j}$ jne.

Jos virhe joudutaan arvioimaan monimutkaisemmalle lausekkeelle (esim. epälineaarinen funktio, jossa kaavassa esiintyy vaikkapa trigonometrisiä funktioita) on helpointa käyttää Monte Carlo simulointia lopullisen virheen arvioimiseen. Idea on yksinkertainen (mutta toteutus joskus tuskallisen hidas)

- toistetaan N kertaa
    1. simuloidaan muuttujien arvot niille oletetuista todennäköisyysjakaumista
    2. lasketaan lausekkeen arvo ja merkitään se muistiin
- lopuksi luetaan lausekkeen virherajat muistiin merkittyjen arvojen jakaumasta

Esimerkkejä Monte Carlo simuloinnin käytöstä seuraa myöhemmin. Menetelmän käyttö ei rajoitu mitenkään normaalijakaumaan, vaan muuttujien todennäköisyysjakaumat voivat olla mielivaltaisia - esim. empiirinen jakauma. Menetelmällä saadaan lisäksi simuloitua laskettavan lausekkeen koko todennäköisyysjakauma. Tulos voidaan tietenkin tiivistää keskiarvon ja keskihajonnan muotoon - olettaen, että jakauma on lähellä normaalijakaumaa. Menetelmää voidaan periaatteessa käyttää kaikissa tapauksissa, mutta se on usein hidas ja esim. muuttujien välisten korrelaatioiden oikea simulointi ei välttämättä ole helppoa.

Esimerkki 1.5. Lasketaan Monte Carlo menetelmällä virhejakauma lausekkeelle

$$y = \sin(x_1) + \exp(0.08\, x_2)$$

pisteessä $(x_1,\, x_2) = (1.0,\, 1.0)$. Oletetaan lähtömuuttujien jakaumiksi $x_1 \sim N(1.0,\, 0.2)$ ja $x_2 \sim \chi^2_{\nu=5}$. Generoidaan satunnaismuuttujat x_1 ja x_2 10000 kertaa ja merkitään y:n arvot muistiin. Seuraavassa simulaation suorittava R-ohjelma:

```
#----- y:n arvot talletetaan vektoriin 'y'
y _ 0
for (i in 1:10000) {
    x1   = rnorm(1, 1.0, 0.2) ;     # x1 ~ N(1.0,0.2)
    x2   = rchisq(1, 5.0) ;         # x2 ~ Chisq(v=5)
    y[i] = sin(x1)+exp(0.08*x2) ;
}
#-----  histogrammi
hist(y)
#-----  rajat, joiden väliin jää 67% havainnoista
```

```
cat("67% väli = ", quantile(y, c(0.1587, 0.8413)), "\n")
#-----  verrataan 'normaaleihin' virherajoihin
ka = mean(y) ;   kh = sd(y)
cat("keskiarvo = ", ka)
cat("keskiarvo+-keskihajonta = ", ka-kh, ka+kh, "\n") ;
```

Ohjelma tulostaa välit

```
67% väli                 =    2.0451  3.0408
keskiarvo                =    2.5587
keskiarvo+-keskihajonta =    1.8495  3.2679
y = 2.5587 - 0.5137 + 0.4821
```

67% y:n arvoista jää välille [2.045, 3.041] ja keskiarvo on 2.5587, joten tulos voidaan esittää muodossa y=$2.56^{+0.48}_{-0.51}$. Epäsymmetria kasvaa, kun jakaumassa siirrytään kauemmas keskiarvosta. Esimerkiksi 95% luotettavuusväliä käyttäen tulos olisi $y = 2.56^{+1.83}_{-0.74}$. Normaalia laajemman luotettavuusvälin tapauksessa olisi luontevampaa ilmoittaa suoraan, että "95% luotettavuusväli on [2.04, 3.07]".

Edellä käsiteltyjen *statististen* virheiden lisäksi mittauksissa esiintyy myös *systemaattisia* virheitä, jotka toistuvat samanlaisina mittauksesta toiseen. Mitä enemmän havaintoja tehdään, sitä pienemmäksi muodostuu lopullinen statistinen virhe. Sen sijaan systemaattiseen virheeseen ei useampien havaintojen tekeminen vaikuta, vaan mittauksista saatu arvo on yhä virheellinen. Jos systemaattinen virheen suuruus pystytään arvioimaan, se pitää ilmoittaa erillään statistisesta virheestä, esim: $1.80 \pm 0.20 \pm 0.15$ – tässä toinen luku on statistinen ja viimeinen systemaattinen virhe. Merkintä ei luonnollisesti tarkoita, että tuloksessa olisi annetun suuruinen systemaattinen virhe, vaan että virheen suuruuden arvioidaan olevan $\sim N(0, 0.15)$. Virheitä yhdistettäessä systemaattiset virheet voidaan laskea neliöllisesti yhteen. Itse asiassa tämä on jälleen korrektia vain, jos systemaattiset virheet eivät korreloi keskenään. Toisaalta suora yhteenlasku johtaa todennäköisesti liian suuriin virhearvioihin. Jos korrelaatioita ei pystytä arvioimaan, on turvallisempaa laskea systemaattiset virhettä yhteen esim. korotettuina potenssiin 1.4. Edellisten virhetyyppien lisäksi laskettava estimaatti itsessään voi tuottaa virheellisiä arvoja, esimerkiksi koska havaintojen todellinen todennäköisyysjakauman poikkeaa oletetusta. Tällaista virhettä kutsutaan *harhaksi* (*bias*), ja se voi aiheutua esimerkiksi havaintoihin vaikuttavista valintaefekteistä.

Luku 2
Tilastollisia testejä

Seuraavassa luetellaan lyhyesti (miltei) jokapäiväisessä työssä tarvittavia tilastollisia testejä. Tilastolliset testit kuuluvat todentavaan data-analyysiin. Lähtökohtana on jokin malli tai oletus jonka todenmukaisuus (todennäköisyys) pyritään havaintojen avulla määrittämään.

Todennäköisyysjakaumasta voidaan suoraan lukea eri hypoteesien todennäköisyyksiä. Käytetty jakauma voi olla empiirinen jakauma tai teoreettinen jakauma, jonka parametrit on tiedossa (johdettu havainnoista tai teoriasta). Merkitään todennäköisyyden kertymäfunktiota $P(x)$. Seuraavassa on pari yksittäisiä jakaumasta generoituja satunnaislukuja koskevia hypoteeseja sekä niiden todennäköisyydet

$$\begin{aligned} \mathcal{P}(x > x_0) &= \text{1-}P(x_0) \\ \mathcal{P}(|x| > m) &= P(-m) + 1 - P(m) \\ \mathcal{P}(x > y) &= \int_{-\infty}^{\infty} p_x(x')\, P_y(x')\, dx' \end{aligned}$$

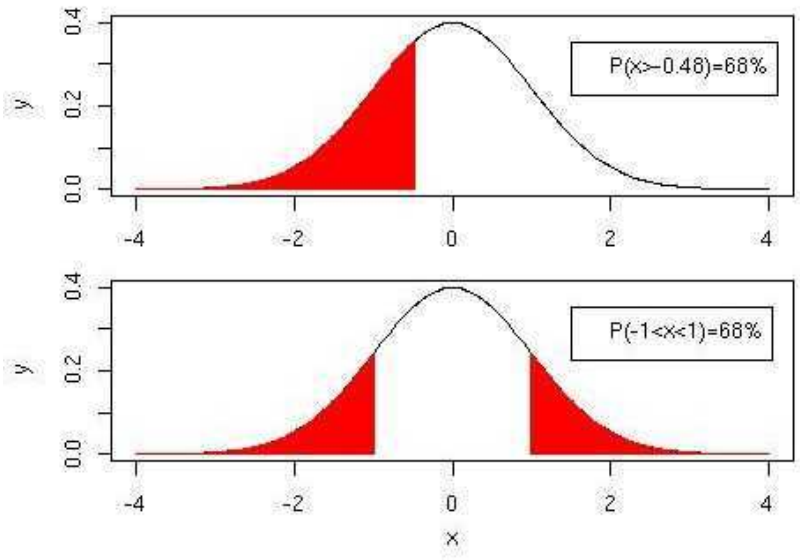
Parametrin *luottamusväliksi* luottamustasolla β kutsutaan väliä, jonka sisälle jää osuus β kaikista tapauksista:

$$\text{luottamusväli } [a, b]\text{: } \mathcal{P}(a \leqslant x \leqslant b) = \beta. \tag{2.1}$$

Tilastollinen testi aloitetaan asettamalla kaksi hypoteesia. Nollahypoteesi (*null hypothesis*) kuvaa oletettua asiantilaa. Havaintojen perusteella pyritään selvittämään, onko asetettu hypoteesi tosi. Nollahypoteesia merkitään *H0*. Testausta varten asetetaan myös toinen hypoteesi *H1*, joka toteutuu silloin, kun *H0* on väärä. Hypoteesien ei tarvitse olla komplementaarisia. Voidaan esimerkiksi tutkia oletusta *H0*, että kahden otoksen keskiarvot ovat samat. Vastahypoteesina voi olla (1) $\mu_1 > \mu_2$, (2) $\mu_1 < \mu_2$ tai (3) $\mu_1 \neq \mu_2$. Tapauksessa (1) uskotaan, että joko *H0* toteutuu tai vaihtoehtoisesti ensimmäisen otoksen keskiarvo on suurempi. Tapauksessa (3) periaatteessa kumman tahansa otoksen keskiarvo voi olla toista suurempi. Tapaukset (1)-(2) ovat esimerkkejä toispuoleisista testeistä, ja tapauksessa (3) testi on kaksipuoleinen.

Tilastolliset testit perustuvat sopivan testimuuttujan (*statistics*) laskemiseen havainnoista. Kun tämän muuttujan todennäköisyysjakauma tunnetaan, voidaan sen arvo muuntaa todennäköisyysarvoksi. Tätä kutsutaan muuttujan merkitsevyyden testaamiseksi (*significance testing*). Toinen vaihtoehto on alunperin asettaa määrätty *kriittinen alue*, jolla nollahypoteesin todennäköisyys tippuu alle annetun todennäköisyyden α. Jos testimuuttujan arvo osuu kriittiselle alueelle nollahypoteesi tulee hyljätyksi luotettavuustasolla β=100(1-α)%. Tätä menettelyä kutsutaan hypoteesin testaamiseksi (*hypothesis testing*). Kriittisen alueen sijainti vaihtelee sen mukaan, onko kyseessä toispuoleinen vai kaksipuoleinen testi.

Testauksessa voidaan tehdä kahdenlaisia virheitä. Ensinnäkin, hypoteesi H_0 voidaan hylätä, vaikka se itse asiassa olisikin tosi. Tätä kutsutaan tyypin I virheeksi (*type I error*). Vastaavasti hypoteesi H_0 voidaan hyväksyä vaikka se olisi virheellinen. Tämä on tyypin II virhe. Testin teho (*power*) on todennäköisyys, että hypoteesi H_0 hylätään, kun vaihtoehtoinen hypoteesi H_1 on tosi. Testi on tietenkin sitä parempi, mitä tehokkaampi se on. Alla käsitellään normaalijakaumaan perustuvia testejä sekä toisaalta testejä, joissa todennäköisyysjakaumasta ei tehdä mitään oletuksia. Normaalijakaumaan perustuvat testit ovat yleensä tehokkaampia – mutta vain sillä oletuksella, että satunnaismuuttujat todella ovat gaussisia.



Kuva 2.1. Kriittiset alueet testille H_0: $x = 0$ (α=0.32), kun vaihtoehtoinen hypoteesi on H_1: $x < 0$ (yläkuva) tai H_1: $x \neq 0$ (alakuva) ja jakauma on normaalijakauma $N(0, 1)$.

2.1 Todennäköisyysjakauman muodon testaus

Useat testit olettavan mittausten noudattavan normaalijakaumaa. Siispä ensin on hyvä testata tämän oletuksen paikkansapitävyyttä. Kuvien piirtäminen (*scatter plots*, histogrammit, jne) ovat tässäkin korvaamattomia, sillä ne paljastavat välittömästi poikkeavat arvot tai vaikkapa jakauman vinouden. Varsinaisesti kahta

todennäköisyysjakaumaa verrataan **Q-Q-piirroksella** (*quantile-quantile plot*), jossa havaitun jakauman kvantiilit piirretään joko toisen otoksen tai teoreettisen jakauman kvantiilien funktiona. Q-Q–pirros on helppo piirtää vaikka käsin, esimerkiksi kvartiileja käyttäen. Kuva vertaa siis itse asiassa jakaumien kertymäfunktioita. Jos todennäköisyysjakaumat ovat samat, osuvat pisteet kuvan nousevalle diagonaalille. Vaikka kvartiiliarvot olisivat kahdelle jakaumalle samat, voivat jakaumat silti poiketa huomattavasti toisistaan. Voi siis olla syytä piirtää Q-Q piirros käyttäen useampia kvantiileja. Kertymäfunktiosta voidaan tehdä vaihtoehtoisesti nk. **P-P–piirros**, jossa vertailtavat kertymäfunktiot piirretään samalle asteikolle.

Kahden jakauman samuutta testaavista testeistä on **Kolmogorov-Smirnov** testi yleisimmin käytetty. Testissä piirretään verrattavien jakaumien kertymäfunktiot muuttujan funktiona, ja kuvasta luetaan käyrien välinen suurin y-suuntainen etäisyys

$$D_{\mathrm{obs}} = m\,a\,x_x |S_N(x) - P(x)|. \tag{2.2}$$

Tämän arvon ja otosten suuruuksien perusteella lasketaan todennäköisyys nollahypoteesille, että molemmat otokset ovat peräisin samasta jakaumasta (tai että otos on peräisin annetusta teoreettisesta jakaumasta). Testiparametrin arvot eri todennäköisyyksille löytyvät tilastotieteen taulukkokirjoista. Todennäköisyydet voidaan myös laskea likimääräisesti kaavasta

$$\mathcal{P}(D > D_{\mathrm{obs}}) = Q_{KS}\left(D[\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}}] \right), \tag{2.3}$$

missä funktio Q_{KS} määritellään

$$Q_{KS}(\lambda) = s \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}. \tag{2.4}$$

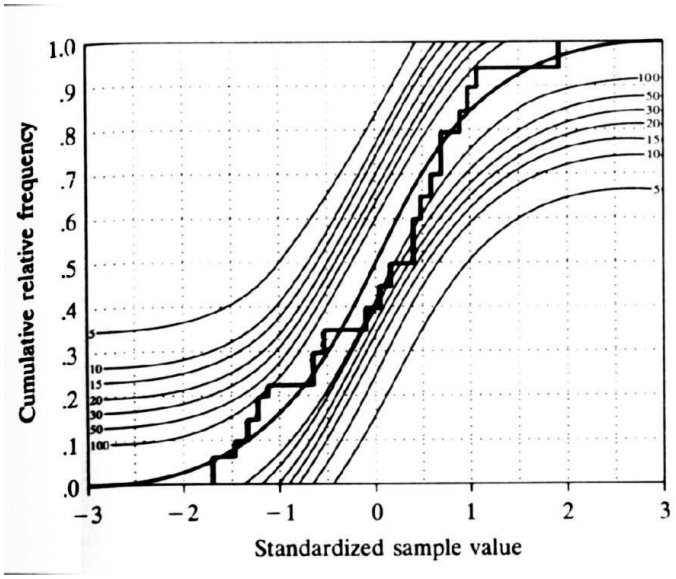
Lisäksi kaavassa esiintyy N_e, eli pisteiden efektiivinen lukumäärä. Jos N-pisteen otosta verrataan teoreettiseen kertymäfunktioon, $N_e = N$. Jos verrattavina ovat N_1 ja N_2 pisteen otokset, määritellään $N_e = (N_1 N_2)/(N_1 + N_2)$.

Lillieforsin normaalisuustesti on modifioitu Kolmogorov-Smirnov testi, joka testaa nimenomaan sitä, onko otos peräisin normaalijakaumasta. Data standardoidaan käyttäen otoskeskiarvoa ja otoskeskihajontaa. Uusista datapisteistä muodostetaan empiirinen kertymäfunktio $F_n(x)$, jota verrataan normaalijakauman kertymäfunktioon $\Phi(x)$. Testiparametri on

$$D_n = m\,a\,x_{\,x}\,|\,F_n(x) - \Phi(x)\,|\,, \tag{2.5}$$

aivan Kolmogorov-Smirnov testin tapaan. Tätä vastaava todennäköisyys luetaan taulukoista. Testiä varten on olemassa myös piirroksia, joissa standardoiduista mittauksista laskettua kertymäfunktiota verrataan vastaavankokoiselle otokselle laskettuihin, eri todennäköisyyksiä vastaaviin käyriin. Jos otoksesta saatu kertymäfunktio leikkaa esim. 90% vastaavan käyrän, voidaan nollahypoteesi (siis oletus jakauman normaaliudesta) hylätä tällä luotettavuustasolla. Testi on erityisen hyödyllinen pienten otosten tapauksessa.

Myös esim. **Shapiro & Wilk W-testillä** voidaan testata, onko otos peräisin normaalijakaumasta.



Kuva 2.2. Esimerkki Lillieforsin normaalisuustestistä. Histogrammi on havaittu kertymäfunktio. Muut käyrät ovat osoittavat kertymäfunktion rajat luotettavuustasolla α=0.05 erisuuruisille otoksille. (Kuva: Milton & Arnold: Introduction to Probability and Statistics)

R-kirjastosta `ctest` löytyy koko joukko 'klassisia' tilastollisia testejä, mukaanluettuina Kolmogorov-Smirnov testi (`ks.test`) ja W-testi (`shapiro.test`).

2.2 Normaalijakauman testejä

Seuraavissa testeissä otosten oletetaan olevan peräisin normaalijakaumasta. Kahden otoksen keskiarvojen normitettu *erotus noudattaa Studentin* t*-jakaumaa*, ja keskiarvojen samuutta testataan t-testillä. Testin toteutus eroaa sen mukaan, verrataanko jakaumaa toiseen otokseen vai teoreettiseen jakaumaan. Samoin laskentaan vaikuttaa se, ovatko jakaumien keskihajonnat ja/tai keskiarvot tunnettuja vai pitääkö ne ensin estimoida samasta aineistosta. Oletetaan, että otoksen keskiarvoa verrataan annettuun arvoon μ_0. Aluksi lasketaan testimuuttuja

$$T = (\bar{x} - \mu_0) \,/\, (s/\sqrt{n}), \tag{2.6}$$

missä $\bar{x}$ on otoskeskiarvo ja s samoin otoksesta laskettu keskihajonta. Suure T mittaa siis keskiarvon poikkeamaa annetusta luvusta μ_0, suhteutettuna keskiarvon keskivirheeseen (n on otoksen suuruus). T noudattaa Studentin t-jakaumaa, jossa vapausasteiden lukumäärä on n-1 – yksi vapausaste menetetään jälleen otoskeskiarvon laskuun. Testataan hypoteesia $H0$: $\mu = \mu_0$, missä μ on se tuntematon keskiarvo, jonka estimaatti $\bar{x}$ on. Seuraavassa on esitetty kolme tilannetta sen mukaan, mikä on vaihtoehtoinen hypoteesi $H1$.

hypoteesi $H1$	hylkää $H0$, jos	$H0$ todennäköisyys
$\mu < \mu_0$	$T < -t_{n-1}^{-1}(1-\alpha)$	$P(t_{n-1} < T)$
$\mu > \mu_0$	$T > t_{n-1}^{-1}(1-\alpha)$	$P(t_{n-1} > T)$
$\mu \neq \mu_0$	$\lvert T\rvert > t_{n-1}^{-1}(1-\alpha/2)$	$2P(t_{n-1} > \lvert T\rvert)$

Merkintä $t_{n-1}^{-1}(\alpha)$ määritellään siten, että $P(t_{n-1} \leqslant t_{n-1}^{-1}(\alpha))=\alpha$. R:ssä t-testi voidaan useimmissa tapauksissa suorittaa rutiinin `t.test` avulla.

Verrattaessa kahta otosta, on testiparametri

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\sigma_A^2/N_A + \sigma_B^2/N_B}}. \tag{2.7}$$

Kaavassa esiintyvät molempien otosten keskiarvot sekä niiden keskihajonnat. Muuttuja noudattaa Studentin t-jakaumaa, jossa vapausasteiden lukumäärä on

$$\nu = \frac{\left(\sigma_A^2/N_A + \sigma_B^2/N_B\right)^2}{(\sigma_A^2/N_A)^2/(N_A-1) + (\sigma_B^2/N_B)^2/(N_B-1)} \tag{2.8}$$

(Smith-Satterthwaite menetelmä kts. Milton & Arnold, s. 351). Keskiarvojen vertailu tapahtuu edelleen hieman eri lailla, jos otosten varianssit oletetaan samoiksi. Tällöin voidaan nimittäjän keskihajonta määrittää yhdistetystä aineistosta ($N_A + N_B$ pistettä), ja testin teho vastaavasti kasvaa. Jos mittaukset koostuvat pareista (x_i, y_i), saadaan otoskeskiarvojen erotuksen luotettavuusrajoiksi

$$\bar{D} \pm t_{\alpha/2}\, S_d / \sqrt{n}\ . \tag{2.9}$$

Tässä N on pisteiden lukumäärä. $\bar{D}$ ja S_d ovat pisteittäin laskettujen erotusten otoskeskiarvo ja keskihajonta.

Verrattaessa otoksen varianssia (s^2) annettuun arvoon (σ_0^2) on testimuuttuja

$$\chi^2 = (n-1)\,(\frac{s}{\sigma_0})^2\ . \tag{2.10}$$

Kuten merkintä antaa ymmärtää, tämä noudattaa χ^2- jakaumaa. Vapausasteiden lukumäärä on otoksen suuruus miinus 1. Yksi vapausaste häviää keskiarvon laskemiseen, mikä vaaditaan s:n laskemiseksi. Asetetaan nollahypoteesi $H0$: $\sigma^2 = \sigma_0^2$, missä σ tarkoittaa todellista keskihajontaa, jonka estimaatti on s. Voidaan erottaa kolme tilannetta sen mukaan, mikä on vaihtoehtoinen hypoteesi $H1$. Seuraava taulukko kertoo, milloin nollahypoteesi voidaan hylätä annetulla merkitsevyystasolla α, sekä miten lasketaan nollahypoteesin todennäköisyys.

hypoteesi $H1$	hylkää $H0$, jos	$H0$ todennäköisyys
$\sigma^2 < \sigma_0^2$	$\chi^2 < {\chi^2_{n-1,\alpha}}^{-1}(0)$	$P(\chi^2_{n-1}(0) < \chi^2)$
$\sigma^2 > \sigma_0^2$	$\chi^2 > {\chi^2_{n-1,1-\alpha}}^{-1}(0)$	$P(\chi^2_{n-1}(0) > \chi^2)$
$\sigma^2 \neq \sigma_0^2$	$\chi^2 < {\chi^2_{n-1,\alpha/2}}^{-1}(0)$ tai $\chi^2 > {\chi^2_{n-1,1-\alpha/2}}^{-1}(0)$	$2\ \min(P(\chi^2_{n-1}(0) > \chi^2), P(\chi^2_{n-1}(0) < \chi^2))$

(Taulukko on mukaeltu teoksesta Dudewics & Mishra, Modern Mathematical Statistics). Taulukossa esiintyvä merkintä ${\chi^2_{n-1,\alpha}}^{-1}(0)$ tarkoittaa jälleen lukua, jonka alapuolella $\chi^2_{n-1}(0)$ on todennäköisyydellä α.

Edellä verrattiin otoksen varianssia annettuun arvoon. Entäpä jos meillä on kaksi otosta? Tässä tapauksessa otosvarianssien samuutta testataan F-testillä, jossa käyttöön tulee luvussa 1.2 ohimennen mainittu F-jakauma. Hypoteesit ovat kuten edellä, mutta σ_0:n korvaa toisen otoksen keskihajonta. Aluksi lasketaan otosvarianssien suhde

$$F = s_1^2 \,/\, s_2^2 \,, \tag{2.11}$$

ja sen jälkeen testi etenee seuraavan taulukon mukaisesti.

hypoteesi $H1$	hylkää $H0$, jos	$H0$ todennäköisyys
$\sigma_1^2 < \sigma_2^2$	$F < F^{-1}_{n_1-1,n_2-1,\alpha}$	$P(F_{n_1-1,n_2-1} < F)$
$\sigma_1^2 > \sigma_2^2$	$F > F^{-1}_{n_1-1,n_2-1,1-\alpha}$	$P(F_{n_1-1,n_2-1} > F)$
$\sigma_1^2 \neq \sigma_2^2$	$F < F^{-1}_{n_1-1,n_2-1,\alpha/2}$ tai $F > F^{-1}_{n_1-1,n_2-1,1-\alpha/2}$	$2\,\min\{P(F_{n_1-1,n_2-1} < F), P(F_{n_1-1,n_2-1} > F)\}$

Kuten taulukosta näkyy, F-jakauman parametreinä ovat molempien otosten keskihajontojen laskussa esiintyvät vapausasteiden lukumäärät. Tästä esimerkistä voidaan päätellä, että yleisemminkin normaalijakautuneiden satunnaislukujen neliöiden osamäärä noudattaa F-jakaumaa.

Aiemmin todettiin jo, että normaalijakautuneiden satunnaislukujen neliösumma noudattaa χ^2- jakaumaa. Jos kyseessä on funktion sovitus havaintoihin on χ^2-jakauman vapausasteiden lukumäärä datapisteiden lukumäärä vähennettynä mallin parametrien lukumäärällä. Lasketun χ^2-arvon vertaaminen vastaavaan todennäköisyysjakaumaan antaa siten periaatteessa todennäköisyyden sille, että datassa ei ole piirteitä, joita malli ei olisi selittänyt. Tämä tulkinta edellyttää kuitenkin, että pisteiden virhearviot ovat oikeita. Kääntäen voidaan tietysti myös käyttää χ^2 lukua sen arvioimiseen, olivatko virhearviot alunperin esim. liian suuret. Normalisoidulle χ^2-luvulle

$$\chi^2/N = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

pitäisi arvon olla ~ 1, jos kaikki (virhearviot+malli) on kunnossa. Jos arvo on alle yhden, ovat virhearviot olleet liian suuria.

χ^2–testillä voidaan yleisesti verrata havaittujen tapahtumien frekvenssejä teoreettisiin arvoihin – siis verrata periaatteessa havaittua histogrammia teoreettiseen. Olkoon havaitut tapahtumien frekvenssit y_i ja vastaavat teoreettiset arvot $\hat{y}_i$. Kaavasta

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \tag{2.12}$$

laskettu χ^2-arvo noudattaa χ^2–jakaumaa, jossa vapausasteiden lukumäärä on $k-1$. Jos nollahypoteesina on havaintojen yhteensopivuus teoreettisten arvojen kanssa, voidaan hypoteesi hylätä merkitsevyystasolla α, mikäli laskettu arvo on suurempi kuin χ^2_α (= raja, jonka yläpuolelle χ^2–jakaumassa jää osa α kaikista tapauksista). Kaava pätee itse asiassa vain, jos frekvenssit ovat riittävän suuria ($\gtrsim 5$ ja ehdottomasti $\geqslant 1$). Pienille frekvensseille on olemassa muunnoskaavoja, joilla testiä voidaan soveltaa vielä silloinkin, kun frekvenssit ovat pieniä useissa (mutta ei useimmissa) väleissä.

Monet jakaumat lähestyvät normaalijakaumaa, kun pisteiden määrä kasvaa riittävän suureksi. Näin käy myös binomijakaumalle, ja sen vuoksi suurten otosten tapauksessa testit voivat perustua normaalijakaumaan. Merkitään $\hat{p}$ tietyn ominaisuuden omaavien havaintojen suhteellista lukumäärää

$$\hat{p} = \frac{n}{N}. \tag{2.13}$$

Tämä on tietenkin samalla binomijakaumassa esiintyvän todennäköisyyden estimaatti. Parametrin luotettavuusrajat ovat

$$\hat{p} \pm z_{\alpha/2}\, \sqrt{\hat{p}\,(1-\hat{p})\,/\,N}. \tag{2.14}$$

Merkintä $z_{\alpha/2}$ tarkoittaa normaalijakauman pistettä, jossa kertymäfunktion arvo on $\alpha/2$. Luotettavuusrajat perustuvat keskeiseen raja-arvoteoreemaan, joten kaava pätee vain, jos N on riittävän suuri. Jos $\hat{p}$:n arvo tiedetään suurin piirtein, voidaan tarvittava otoksen vähimmäiskoko arvioida kaavasta

$$N_{\min} = \frac{z^2_{\alpha/2}\,\hat{p}\,(1-\hat{p})}{d^2}\,, \tag{2.15}$$

kun ensin asetetaan vaadittu tarkkuus, $d \sim p - \hat{p}$.

Jos Pearsonin korrelaatiokerroin lasketaan normaalijakautuneille otoksille, voidaan korrelaatiokertoimelle johtaa luotettavuusväli normaalijakaumaa käyttäen. Kun korrelaatiokerrointa merkitään ρ, noudattaa testimuuttuja

$$Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{2.16}$$

normaalijakaumaa

$$N\left(\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right),\ \sqrt{\frac{1}{n-3}}\right). \tag{2.17}$$

Otoksen suuruus on n. Testi tapahtuu esim. standardoimalla muuttuja Z ja vertaamalla arvoa normaalijakaumaan $N(0,1)$:

$$\mathcal{P}\left[\, -z_{\alpha/2} \leqslant \frac{\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}} \leqslant z_{\alpha/2} \,\right] = 1-\alpha. \tag{2.18}$$

Tästä saadaan ratkaistua 100(1-α) prosentin luotettavuusväli, kun arvot $z_{\alpha/2}$ luetaan normaalijakauman taulukoista. Ratkaistut ylä- ja alarajat ovat

$$\frac{(1+\rho) - (1-\rho)\exp[\,\pm 2\,z_{\alpha/2}\,/\,\sqrt{n-3}\,]}{(1+\rho) + (1-\rho)\exp[\,\pm 2\,z_{\alpha/2}\,/\,\sqrt{n-3}\,]}\,. \tag{2.19}$$

2.3 Parametrittomia testejä

Parametrittomat testit (*nonparametric, distribution free*) eivät oleta tiettyä todennäköisyysjakaumaa. Niitä on syytä käyttää silloin, kun jakauman normaaliudesta ei ole varmuutta. Toisaalta, silloin kun havainnot todella ovat normaalijakautuneita, ovat edelliset luvun testit tehokkaampia.

Wilcoxon Rank-Sum Test on testi, jota käytetään vertaamaan kahden otoksen mediaaneja. Menetelmä perustuu havaintojen 'rankkaukseen' niiden suuruuden mukaan. Olkoon otosten suuruudet N_1 ja N_2. Aluksi kaikki mittaukset järjestetään pienimmästä suurimpaan, $N_1 + N_2$ pituiseksi vektoriksi. Tämän jälkeen lasketaan suure W_m, joka on pienemmän otoksen mittausten järjestyslukujen summa tässä vektorissa. Vertaamalla muuttujan arvoa taulukoituihin arvoihin, voidaan selvittää asetetun hypoteesin (esim. H_0: $M_1 = M_2$, H_1: $M_1 < M_2$) todennäköisyys. Suurten otosten tapauksessa voidaan edellisen sijasta taas turvautua normaalijakaumaan, sillä studendoitu muuttuja

$$\frac{W_m - E(W_m)}{\sqrt{s^2(W_m)}} \tag{2.20}$$

noudattaa normaalijakaumaa $N(0,1)$. Odotusarvo ja varianssi lasketaan kaavoista

$$E(W_m) = \frac{N_1(N_1+N_2+1)}{2}, \quad s^2(W_m) = \frac{N_1 N_2 (N_1+N_2+1)}{12}\,. \tag{2.21}$$

Jos havainnot koostuvat pareista $(x_i,\ y_i)$, voidaan käyttää samantapaista järjestyslukutestiä. Nyt kuitenkin lasketaan ensin kullekin havainnolle $x_i - y_i$. Nämä asetetaan suuruusjärjestykseen, ja kunkin pisteen järjestysluvulle R_i annetaan vastaavan erotuksen $(x_i - y_i)$ etumerkki. Testimuuttujina lasketaan summat

$$W_+ = \sum_{R>0} R_i\,, \qquad W_- = \sum_{R<0} |R_i|\,. \tag{2.22}$$

Näistä summaa W_- käytetään silloin, kun vaihtoehtoinen hypoteesi on H_1: $M_x > M_y$. Jos W_- on tällöin liian suuri, tulee nollahypoteesi hylätä. Vastaavasti W_+ käytetään testissä, jossa H_1: $M_x < M_y$. Tiettyä parametriarvoa vastaavat todennäköisyydet saadaan valmiiksi lasketuista taulukoista (*Wilcoxon Signed-Rank test for Paired Observations*).