

# TILASTOTIETEEN PERUSTEET, kl 2011

## Luku 3: TILASTOLLINEN KUVAILU

Tässä luvussa esittelemme menetelmiä, joilla havaintomatriisin yhden sarakkeen eli yhden muuttujan havaintojen sisältämä informaatio kuvaillaan taulukoiden, graafisten esitysten ja tunnuslukujen avulla.

### 3.1 Frekvenssijakauma

Muuttujan  $X$  yksiulotteisella **frekvenssijakaumalla** tarkoitetaan seuraavanlaisen taulukon sisältöä

| $X$ :n mahdolliset arvot t. luokat | absoluuttinen frekvenssi | suhteellinen frekvenssi (%) |
|------------------------------------|--------------------------|-----------------------------|
| $A_1$                              | $m_1$                    | $p_1$                       |
| $A_2$                              | $m_2$                    | $p_2$                       |
| $\vdots$                           | $\vdots$                 | $\vdots$                    |
| $A_r$                              | $m_r$                    | $p_r$                       |
| yhteensä                           | $n$                      | 100                         |

**Absoluuttinen frekvenssi**  $m_k$  kertoo arvon  $A_k$  saaneiden tai luokkaan  $A_k$  kuuluvien havaintoyksikköjen lukumäärän;  $k = 1, 2, \dots, r \in \mathbb{N}$ .

**Suhteellinen frekvenssi**  $p_k = m_k/n$  kertoo luokkaan  $A_k$  kuuluvien suhteellisen osuuden, joka tavallisesti ilmaistaan prosentteina.

Tämä esitystapa käy kaikilla mitta-asteikoilla. Luokittelu- ja järjestysasteikon muuttujilla luokat ovat yleensä valmiina. Välimatka- ja suhdeasteikon muuttujille joutuu yleensä ensin suorittamaan arvojen luokittelun (ks. 3.6).

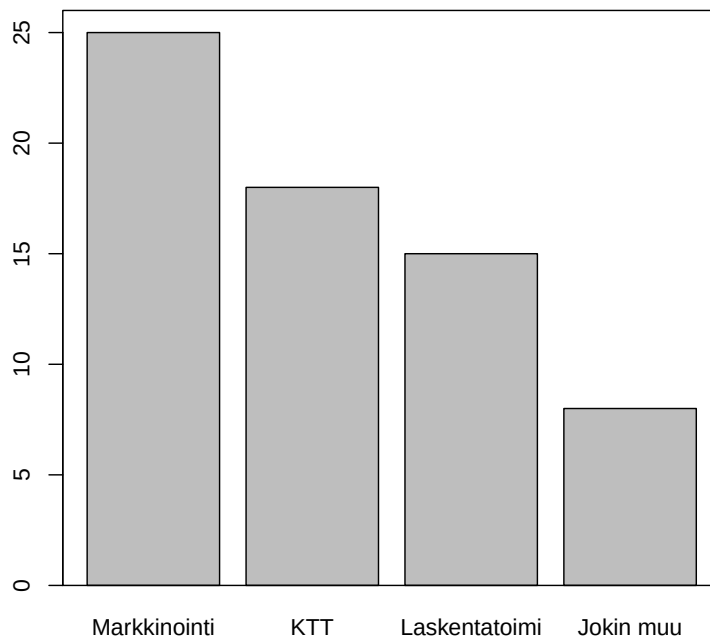
**Esim. 3.1.** Eräälle tilastotieteen kurssille osallistuneiden TaTK:n opiskelijoiden koulutusohjelman jakauma oli seuraavanlainen:

| koulutusohjelma   | lkm | %   |
|-------------------|-----|-----|
| markkinointi      | 25  | 38  |
| kansantaloustiede | 18  | 27  |
| laskentatoimi     | 15  | 23  |
| jokin muu         | 8   | 12  |
| yhteensä          | 66  | 100 |

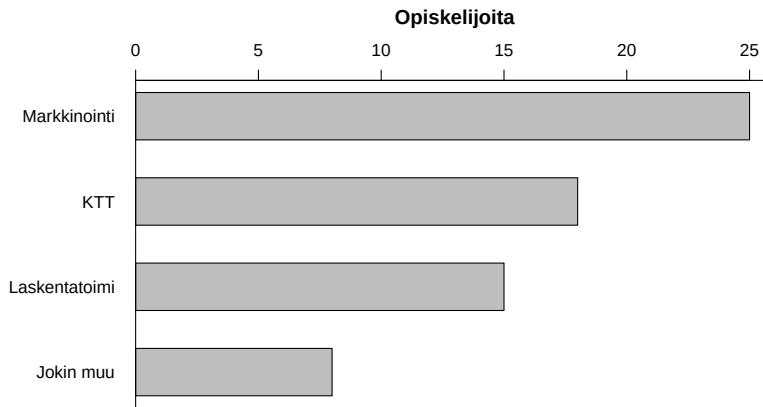
Jakauman **moodi** eli **tyyppiarvo**, merkitään  $Mo$ , on se arvo tai luokka, jonka frekvenssi on suurin.

Esimerkissä 3.1.koulutusohjelman jakauman moodi on “markkinointi”.

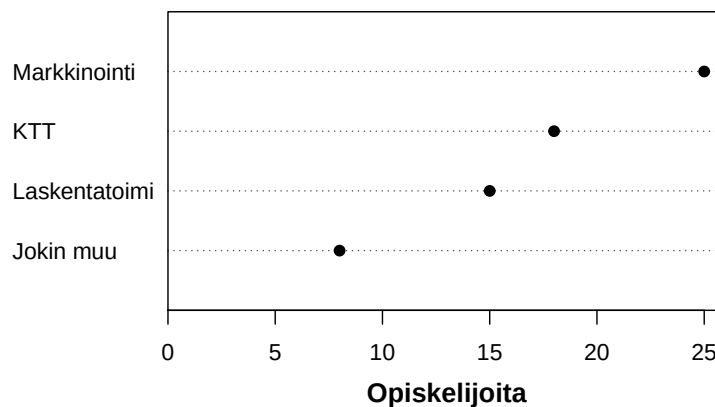
Seuraavassa esitetään **pylväskuvio** koulutusohjelman jakaumasta.



Pystyssä olevien pylväiden hahmottaminen ja niiden suuruuksien vertailu ei välttämättä toimi yhtä hyvin kuin vastaavan **palkkikuvion** palkkien, jotka ovat alekkain vaakasuorassa olevia pylväitä.



Huomattavasti tehokkaampi **data-mustesuhde** (*data-ink ratio*) luokitellun muuttujan jakauman graafiselle esitykselle saadaan **pisteviivakuvion** (*dot-chart*) avulla:



Kaikissa em. kuvioissa voidaan absoluuttisten frekvenssien asemesta kuvata eri luokkien suhteelliset frekvenssit.

**Piirakkakuviota** (*pie chart*) pidetään perustellusti huonoimpana niistä graafisista esityksistä, joita yleisesti käytetään luokitellun frekvenssijakauman kuvaamiseen (ks. esim. <http://www.stat.auckland.ac.nz/~ihaka/courses/787/lectures-presentation-graphics.pdf>).

Lopuksi todettakoon, että kun luokkia on näinkin vähän kuin tässä esimerkissä, niin tutkimusraportissa ison tilan vievän kuvion asemesta säästeliäämpää ja vähintään yhtä informatiivista on tyytyä yksinkertaiseen taulukkoon tai sen sisältämän muutaman luvun esittämiseen tekstissä.

### 3.2. Jatkuvat muuttujat: luokittelematon aineisto

Tarkastelemme nyt luonteeltaan jatkuvan välimatka-asteikon muuttujan jakauman kuvaamista erilaisia graafisia esitystapoja ja tunnuslukuja käyttäen. Esimerkkiaineiston tarjoavat muuttujan PITUUSM mittaustulokset kl 2009 pidetyn kurssin “Tilastotieteen perusteet A” osallistujista ( $n = 50$ ). R-ohjelmassa funktiokutsulla `sort(PITUUSM)` suuruusjärjestykseen lajitellut havainnot (cm) olivat:

```
[1] 157 159 160 160 160 160 160 160 161 161 163 164 165 165 165 165 166 166 166
[20] 167 167 168 168 170 170 170 171 172 172 174 175 175 175 176 176 176 177 177
[39] 178 180 182 182 182 182 183 185 186 188 188 193
```

Kun muuttujan  $X$  alkuperäiset havainnot  $x_1, \dots, x_n$  järjestetään pienimmästä suurimpaan, niin käytetään merkintää

$$x_{(i)} = \text{järjestyksessä } i\text{:nneksi pienin havaintoarvo, } i = 1, \dots, n,$$

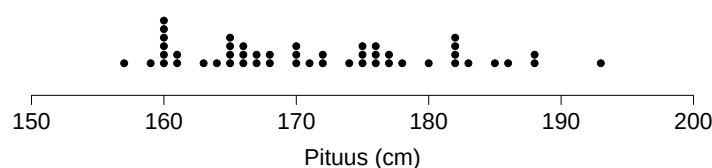
jolloin  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Näillä merkinnöillä mm.  $x_{(1)}$  on muuttujan  $X$  pienin arvo eli **minimi** ja  $x_{(n)}$  suurin arvo eli **maksimi**, ja lukupari  $(x_{(1)}, x_{(n)})$  on **vaihteluväli**. Tässä aineistossa vaihteluväli on (157 cm, 193 cm), joka saadaan R-funktion `range()` avulla argumenttina PITUUSM.

Pituusarvojen **runko-lehtikuvio** saadaan funktiolla `stem()` ja on seuraavan näköinen

```
The decimal point is 1 digit(s) to the right of the |
```

```
15 | 79
16 | 0000001134
16 | 55556667788
17 | 0001224
17 | 555666778
18 | 022223
18 | 5688
19 | 3
```

Sitten piirrämme **pistekuvion** eli **yksiulotteisen parvikuvion**



Tämä kuvio noudattaa erinomaisesti kuuluisan tilastograafikon Edward R. Tuften periaatetta: *Above all else, show the data!* Tätä iskulausetta on sittemmin jatkettu: *... all the data, and nothing but the data.*

### 3.3. Sijaintiluvut

Yksiulotteisen empiirisen jakauman **tunnuslukuja** ovat mm. sijainti- eli keskiluvut, hajontaluvut, vinousluvut ja huipukkuusluvut.

Sijaintilukuja ovat mm. moodi, mediaani sekä aritmeettinen keskiarvo. Hajontalukuja ovat mm. vaihteluväli, kvartiiliväli ja keskihajonta.

Pituuden moodi näyttäisi olevan 160 cm, koska tämä yksittäinen pituusarvo näyttää esiintyvän aineistossa muita arvoja suuremmalla frekvenssillä. Jatkuvan muuttujan tapauksessa moodi kannattaa mieluummin määrätä luokitellusta aineistosta (ks. 3.6-3.8), koska varsinkin pienillä aineistoilla yksittäisistä havainnoista laskettu moodi voi olla hyvinkin satunnainen.

Jakauman **mediaani**, merk.  $M_d$ , määritellään

- (a)  $M_d = x_{(n+1)/2}$  kun  $n$  on pariton ja
- (b)  $M_d = (x_{(n/2)} + x_{(n/2+1)})/2$ , kun  $n$  on parillinen.

Esimerkkiaineistossamme  $n = 50$  on parillinen, joten mediaani on havaintoarvojen  $x_{(25)} = 170$  ja  $x_{(26)} = 170$  keskiarvo eli  $M_d = 170$  cm.

Mediaani jakaa järjestyksessä olevat havainnot keskeltä kahtia niin, että puolet jakaumasta on sen vasemmalla ja puolet oikealla puolella.

Mediaani on sekä arvoltaan että tulkinnaltaan erittäin stabiili tunnusluku. Sen arvoon ei vaikuta jakauman muoto (esim. vinous tai paksuhäntäisyys) eivätkä jotkin muista havainnoista huomattavasti poikkeavat havainnot.

Jakauman **kvartiilit** jakavat järjetyt havainnot neljään yhtä suureen osaan:

- **alakvartiili**  $Q_1$  on sellainen  $X$ :n mahdollinen arvo, jonka vasemmalle puolelle jää 25% jakaumasta.
- **keskikvartiili**  $Q_2 = M_d$  eli mediaani, ja
- **yläkvartiili**  $Q_3$  on sellainen  $X$ :n mahdollinen arvo, jonka vasemmalle puolelle jää 75% jakaumasta.

Esimerkkiaineistossamme alakvartiili on  $Q_1 = x_{(13)} = 165$  cm ja yläkvartiili  $Q_3 = x_{(38)} = 177$  cm.

Kvartiilit ovat erikoistapaus laajemmasta sijaintilukujen luokasta, joita yhteisesti kutsutaan fraktiileiksi eli **kvantiileiksi**. Jakauman  **$p$ -fraktiili** on sellainen muuttujan  $X$  mahdollinen arvo  $q_p$ , jota pienempiä arvoja havaintoaineistossa on  $p$  %. Tässä  $p$ :n ei tarvitse olla kokonaisluku.

**Huom.** Laboratoriolääketieteessä käytetään usein 2.5%:n ja 97.5% fraktiileja rajaamaan jonkin kliniskemiallisen tai fysiologisen muuttujan 95% **viiteväli** eli sellainen mahdollisten laboratorioarvojen väli, jolle 95% terveen t. "normaalin" väestön **viitearvot** sijoittuisivat.

Fraktiilien määrittäminen empiirisessä jakaumassa ei ole yksikäsitteistä. Esimerkiksi R:n funktio `quantile()` tarjoaa 9 erilaista vaihtoehtoista laskukaavaa fraktiileille! Seuraavassa esittelemme R:n oletusarvoisesti tarjoaman laskutavan.

Kun otoskoko on  $n$ , ja halutaan  $p$ -fraktiili  $q_p$ , jossa  $0 < p < 100$ , niin toimitaan seuraavasti:

(1.) lasketaan  $k^* = (n - 1)p/100 + 1$ ,

(2.) merkitään  $k$ :lla luvun  $k^*$  kokonaisosaa

$$k = \max\{j \in \mathbb{N} \mid j \leq k^*\} = \lfloor k^* \rfloor,$$

(3.) lasketaan  $k^*$ :n desimaaliosasta painokerroin  $w = k^* - k$  ( $0 \leq w < 1$ ),

(4.) lasketaan  $p$ -fraktiili  $q_p$  havaintojen  $x_{(k)}$  ja  $x_{(k+1)}$  painotettuna keskiarvona

$$q_p = (1 - w)x_{(k)} + wx_{(k+1)}.$$

Jos  $k^*$  sattuu olemaan kokonaisluku, niin  $k = k^*$  ja  $w = 0$ , jolloin  $q_p = x_{(k)}$ .

Mediaanin tapauksessa tämä kuin myös kaikki muut vaihtoehtoiset laskutavat tiivistyvät kahteen edellä esitettyyn varianttiin, jotka määräytyvät aineiston koon  $n$  parillisuudesta tai parittomuudesta.

Välimatka- tai suhteasteikollisen muuttujan  $X$  **aritmeettinen keskiarvo**, lyhyemmin **keskiarvo**, merk.  $\bar{x}$ , on kaikkien havaintoarvojen  $x_i$  summa jaettuna aineiston koolla  $n$ .

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

Keskiarvo on teoreettisen jakauman odotusarvon  $\mu = \mathbb{E}(X)$  empiirinen vastine. Se kuvaa havaintojen jakauman tasapainopistettä.

Esimerkkiaineistossamme keskiarvo, joka saadaan R-funktion `mean()` kutsulla, on

$$\bar{x} = (157 + \dots + 193)/50 = 171.36 \text{ cm.}$$

Kun jakauma on likimain normaali-jakauman muotoinen eli symmetrinen, yk-sihuippuinen ja riittävän ”ohuthäntäinen” jakauman, niin keskiarvo on opti-maalinen tunnusluku kuvaamaan jakauman sijaintia.

Jos jakauma on kovin ”paksuhäntäinen” tai vino jompaan kumpaan suun-taan, jolloin siinä on joitakin hyvin suuria tai hyvin pieniä arvoja, niin täl-laiset muusta aineistosta poikkeavat arvot vaikuttavat keskiarvoon suurella painolla. Niinpä keskiarvoa ei pidetä jakauman muodon suhteen yhtä **robustina** eikä yksittäisten poikkeavien havaintojen suhteen yhtä **resistanttina** tunnuslukuna kuin esimerkiksi mediaania.

Geometrisen  $G$  keskiarvo määritellään

$$G = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n} = \exp \left\{ \frac{1}{n} \sum_{i=1}^n \log(x_i) \right\}.$$

Se edellyttää vähintään suhdeasteikon mittaustasoa ja kaikkien havaintoarvo-jen  $x_i$  positiivisuutta. Geometrisen keskiarvoa käytetään sijaintilukuna tyy-pillisesti oikealle vinojen jakaumien yhteydessä. Se on optimaalinen tunnus-luku kun havainnot noudattavat likimain log-normaali-jakaumaa, jolloin  $G \approx$  Md. Yleisesti pätee epäyhtälö  $G \leq \bar{x}$ . – Pituusjakaumassa  $G = 171.12 \text{ cm}$ .

### 3.4. Hajontaluvut

Empiirisessä jakaumassa havaintoarvojen vaihtelua kuvaavia tunnuslukuja kutsutaan **hajontaluvuiksi**. Näitä ovat mm. edellä määritelty vaihteluväli  $R_a = (x_{(1)}, x_{(n)})$  kuin myös kvartiiliväli, kvartiilivälin pituus sekä keskiha-jonta.

**Kvartiiliväli**  $QI = [Q_1, Q_3]$  kattaa aineiston keskipuolikkaan eli keskimmäi-set 50% muuttujan  $X$  havaituista arvoista.

Kun lasketaan ylä- ja alakvartiilin erotus, saadaan hajontaluku nimeltään **kvartiilivälin pituus**  $Q^* = Q_3 - Q_1$ .

Pituusjakauman kvartiiliväli on  $QI = [165 \text{ cm}, 177 \text{ cm}]$  ja kvartiilivälin pituus  $Q^* = 177 - 165 = 12 \text{ cm}$

**Keskihajonta** eli lyhyemmin **hajonta**, merk.  $s$  (tai  $s_x$  kun halutaan koros-

taa, että kyse on muuttujan  $X$  hajonnasta) lasketaan kaavalla

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

Hajonnan neliö  $s^2$  on jakauman empiirinen **varianssi** joka on teoreettisen jakauman varianssin  $\sigma^2 = D^2(X) = \text{var}(X)$  empiirinen vastine. Varianssi on additiivisuutensa vuoksi teoreettisissa tarkasteluissa tärkeä tunnusluku. Keskihajonta on kuitenkin käytännössä käyttökelpoisempi ja selkeämmin tulkitavissa, koska sen mittayksikkö on sama kuin muuttujan  $X$  itsensä.

Esimerkissämme pituusdatan havaintojen keskihajonta on R-funktion `sd()` kutsun tuloksena

$$s = \sqrt{\frac{1}{50-1} [(157 - 171.36)^2 + \dots + (193 - 171.36)^2]} = \sqrt{83.50} = 9.14 \text{ cm}.$$

Keskihajonta on parhaiten tulkittavissa, jos jakauma muistuttaa likimain normaalijakaumaa eli on yksihiippuinen sekä riittävän symmetrinen, ja “ohut-häntäinen”. Ideaalitilanteessa eli puhtaassa normaalijakaumassa väli  $[\bar{x} - s, \bar{x} + s]$  sisältää n. 68% jakauman keskeltä, väli  $[\bar{x} - 2s, \bar{x} + 2s]$  kattaa n. 95%, ja välin  $[\bar{x} - 3s, \bar{x} + 3s]$  sisälle jää miltei koko jakauma.

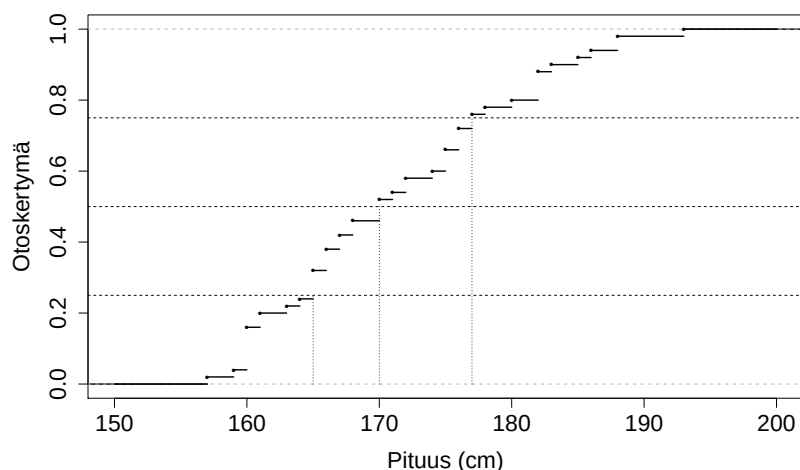
Jos muuttuja  $X$  on suhdeasteikollinen, jolloin sen kaikki arvot ovat ei-negatiivisia, ja kovin vino oikealle, em. tulkinnat eivät päde. Jo yksi keskihajonta keskiarvosta vasemmalle saattaa viedä reaaliakselin negatiiviselle puolikkaalle. Tällöin väleillä  $[\bar{x} - ks, \bar{x} + ks]$ , jossa  $k = 1, 2, \dots$ , ei ole mielekästä tulkintaa. Sen sijaan esim. kvartiileihin tai muihin fraktiileihin perustuvat välit säilyttävät tulkintansa minkämuotoisessa jakaumassa tahansa.

### 3.5 Otokertymäfunktio ja laatikko-janakuvio

Järjestettyjen havaintojen pohjalta voidaan myös tarkastella muuttujan **otokertymäfunktio**  $\tilde{F}(x)$ . Tämä funktio määritellään

$$\tilde{F}(x) = \frac{\#\{i \mid x_{(i)} \leq x\}}{n} = \sum_{i=1}^n I_{(-\infty, x]}(x_{(i)}),$$

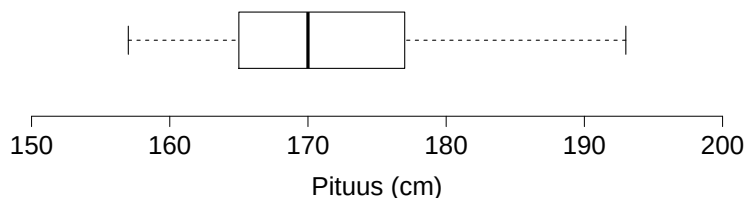
jossa  $I_A : \mathbb{R} \rightarrow \{0, 1\}$  tarkoittaa joukon  $A \subset \mathbb{R}$  indikaattorifunktiota s.e.  $I_A(a) = 1$ , kun  $a \in A$  ja  $I_A(a) = 0$  kun  $a \notin A$  jok.  $A \subset \mathbb{R}$ . Toisin sanoen  $\tilde{F}(x)$  on kaikissa pisteessä oikealta jatkuva porraskuvio, joka lähtee tasolta 0, kun  $x < x_{(1)}$ , hyppää kullakin  $i = 1, 2, \dots, n-1$  tasolle  $i/n$ , kun  $x = x_{(i)}$ , pysyen tällä tasolla, kunnes arvolla  $x = x_{(i+1)}$  nousee tasolle  $(i+1)/n$ , jne. ja päättyy lopulta tasolle 1, kun  $x \geq x_{(n)}$ . – Seuraavassa pituuden jakauman otokertymäfunktion kuvaaja.



R-funktio `ecdf(X)` laskee argumenttina annetun vektorin  $X$  otoskertymäfunktion arvot, ja `plot(ecdf(X))` piirtää ao. kuvaajan.

Kuviossa on lisäksi tasoja 25%, 50% ja 75% merkitty katkoviivoin, ja niiden avulla voidaan määrätä graafisesti jakauman kvartiilit, jotka siis ovat  $Q_1 = 165$  cm,  $Md = Q_2 = 170$  cm ja  $Q_3 = 177$  cm.

**Laatikko-janakuvion** tai lyhyemmin **laatikkokuvion** perusversio tiivistää jakauman kuvauksen viiteen tunnuslukuun (*five-number summary*): minimi, alakvartiili, mediaani, yläkvartiili ja maksimi.



Nykyään suositumpi ja mm. R-ohjelmassa funktion `boxplot()` oletusarvoisesti soveltama laatikkokuvion piirtämisperiaate on hieman mutkikkaampi. Siinä määritellään ensin ns. **sisäaidat**

$$\begin{aligned} \text{alempi sisäaita} &= Q_1 - 1.5 \times Q^*, \\ \text{ylempi sisäaita} &= Q_3 + 1.5 \times Q^* \end{aligned}$$

Laatikosta lähtevät janat tai “viikset” piirretään nyt siten, että alemman janan toinen päätepiste on pienin havaintoarvo  $x_i$ , joka on alempaa sisäaitaa suurempi, kun taas yläkvartiilista lähtevän jana päättyy suurimpaan havaintoon  $x_i$ , joka jää ylempää sisäaitaa pienemmäksi. Sisäaitaa pienemmät ja ulkoaitaa suuremmat havaintoarvot ovat mahdollisia **oudokkeja** (*outlier*)

eli vieraita havaintoja, jotka merkitään yksitellen esim. pienellä ympyrällä 'o'. Jos jakauma on kovin vino tai huipukas suhteessa normaalijakaumaan, niin tällaisia havaintoja tulee hyvinkin helposti ilman, että niissä sinänsä on mitään outoa.

Joissakin laatikkokuvion versioissa otetaan käyttöön vielä lisäksi ns. **ulkoaidat**, jotka ovat  $3 \times Q^*$  mittayksikön päässä ala- ja yläkvartiilista. Näiden aitojen ulkopuolelle sijoittuvia vahvasti poikkeavia arvoja (*extremes*) merkitään ympyrän sijaan esim. tähdellä eli asteriksilla '\*'.

### 3.6. Jatkuvan muuttujan luokiteltu jakauma

Luokittelemme nyt pituuden käyttäen 5 cm luokkavälejä. Seuraavassa taulukossa on yksityiskohtaisesti esitetty luokittelun tulokset: luokkarajat sekä frekvenssijakauman erilaiset frekvenssit.

| Pyöris-<br>tetyt<br>luokka-<br>rajat | Todelliset<br>luokkarajat<br>$a_{k-1}, a_k$ | Luo-<br>kan<br>kes-<br>kus | Abs.<br>frek-<br>venssi<br>$m_k$ | Suht.<br>frek-<br>venssi<br>$p_k$ | Summa-<br>frek-<br>venssi<br>$M_k$ | Suht.<br>summa-<br>frekv.<br>$P_k$ |
|--------------------------------------|---------------------------------------------|----------------------------|----------------------------------|-----------------------------------|------------------------------------|------------------------------------|
| 155-159                              | 154.5, 159.5                                | 157                        | 2                                | 4                                 | 2                                  | 4                                  |
| 160-164                              | 159.5, 164.5                                | 162                        | 10                               | 20                                | 12                                 | 24                                 |
| 165-169                              | 164.5, 169.5                                | 167                        | 11                               | 22                                | 23                                 | 46                                 |
| 170-174                              | 169.5, 174.5                                | 172                        | 7                                | 14                                | 30                                 | 60                                 |
| 175-179                              | 174.5, 179.5                                | 177                        | 9                                | 18                                | 39                                 | 78                                 |
| 180-184                              | 179.5, 184.5                                | 182                        | 6                                | 12                                | 45                                 | 90                                 |
| 185-189                              | 184.5, 189.5                                | 187                        | 4                                | 8                                 | 49                                 | 98                                 |
| 190-194                              | 189.5, 194.5                                | 192                        | 1                                | 2                                 | 50                                 | 100                                |
| Yhteensä                             |                                             |                            | 50                               | 100                               |                                    |                                    |

Taulukon vasemman laidan 1. sarakkeessa on annettu eri luokkien **pyöristetyt luokkarajat** olettaen, että mittaustulokset on pyöristetty lähimpään kokonaiseen senttimetriin.

Tämän pohjalta kunkin luokan  $A_k$  ( $k = 1, \dots, r$ ) **todelliset luokkarajat** määräytyvät suraavasti: todellinen alaraja  $a_{k-1}$  on 0.5 cm pyöristettyä alarajaa pienempi ja todellinen ylärajaa  $a_k$  on pyöristettyä ylärajaa 0.5 cm suurempi.

Luokassa  $A_k$  **luokkavälin pituus** (tai leveys)  $\Delta_k$  on todellisen ylärajan ja todellisen alarajan erotus:  $\Delta_k = a_k - a_{k-1}$ . Luokkakeskus  $\dot{x}_k$  on todellisen alarajan ja todellisen ylärajan keskiarvo:  $\dot{x}_k = (a_k + a_{k-1})/2$ .

Kahdessa oikeanpuoleisessa sarakkeessa olevat luvut saadaan seuraavalla periaatteella:

- **absoluuttinen summafrekvenssi** eli **kumulatiivinen frekvenssi**  
 $M_k = \sum_{j=1}^k m_j,$
- **suhteellinen summafrekvenssi**  $P_k = \sum_{j=1}^k p_j = M_k/n.$

Jos havaintoaineisto on valmiiksi luokiteltu ja taulukoitu edellä olevaan tapaan, ja alkuperäisiä havaintoarvoja ei ole käytettävissä, niin muuttujan  $X$  aritmeettinen ja geometrinen keskiarvo sekä keskihajonta saadaan likimääräisesti lasketuksi eri luokkien frekvenssien ja luokkakeskusten pohjalta seuraavasti.

$$\bar{x} = \frac{1}{n} \sum_{k=1}^r m_k \dot{x}_k, \quad G = \exp \left\{ \frac{1}{n} \sum_{k=1}^r m_k \log(\dot{x}_k) \right\},$$

$$s = \sqrt{\frac{1}{n-1} \sum_{k=1}^r m_k (\dot{x}_k - \bar{x})^2}$$

Tässä aineistossa näiden tunnuslukujen arvot olivat

$$\bar{x} = \frac{1}{50} (2 \cdot 157 + 10 \cdot 162 + \dots + 1 \cdot 192) = 172 \text{ cm},$$

$$\bar{G} = \exp \left\{ \frac{1}{50} [2 \cdot \log(157) + \dots + 1 \cdot \log(192)] \right\} = 171.78 \text{ cm},$$

$$s = \sqrt{\frac{1}{50-1} [2 \cdot (157 - 172)^2 + \dots + 1 \cdot (192 - 172)^2]} = 8.86 \text{ cm}.$$

### 3.7. Histogrammi ja frekvenssimonikulmio

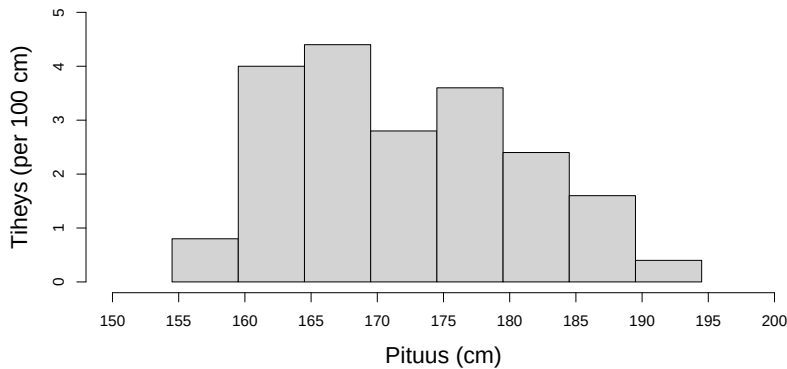
**Histogrammissa** kunkin luokan  $A_k$  ( $k = 1, \dots, r$ ) absoluuttista tai suhteellista frekvenssiä kuvaa pylväs, jonka korkeus  $f_k$  määräytyy seuraavasti:

$$f_k = \text{korkeus} = \frac{\text{suhteellinen frekvenssi}}{\text{luokkaleveys}} = \frac{p_k}{\Delta_k}, \quad k = 1, \dots, r.$$

Niinpä pylvään korkeus  $f_k$  on esim. toiseksi alimmassa luokassa  $8\%/5 \text{ cm} = 1.6$  per 100 cm, ja luokassa 175-179 cm se on  $18\%/5 \text{ cm} = 3.6$  per 100 cm. Pylvään korkeus kuvaa siis ao. luokan **frekvenssitiheyttä**. Pylvään sijainnin ja leveyden  $x$ -akselilla määräävät ao. luokan todelliset luokkarajat. Kun kyseessä on jatkuva muuttuja, niin pylväät piirretään kiinni toisiinsa.

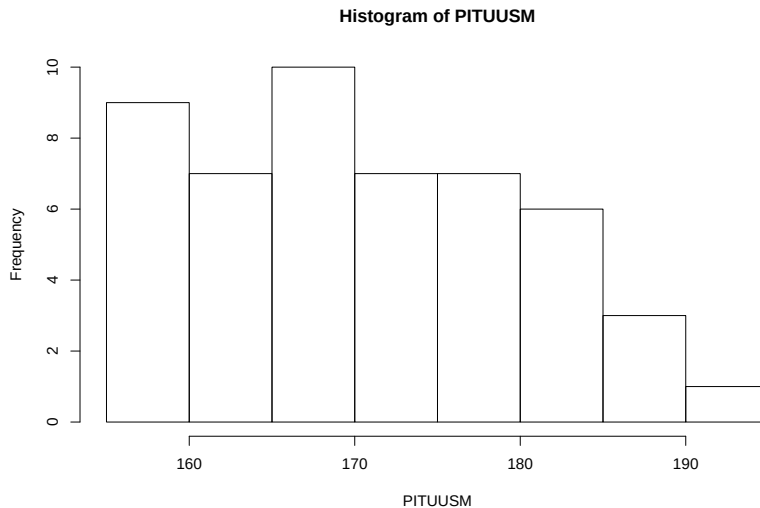
Näiden periaatteiden mukaan on R-funktion `hist()` avulla piirretty pituuden histogrammi, joka esitetään seuraavassa kuviossa. Jakauman moodiksi

voidaan nyt määrätä  $Mo = 167$  cm eli korkeinta pylvästä vastaava luokkakeskus.



Histogrammin ajatellaan usein toimivan parametrittomana estimaattina taustalla olevassa perusjoukossa ao. muuttujalle oletetun jakauman jatkuvalle tiheysfunktiolle  $f(x)$ . Niinpä  $y$ -akselin asteikoksi onkin luontevaa ottaa kunkin luokan suhteellinen frekvenssitiheys. Näin normitettujen pylväiden kokonaispinta-ala summautuu arvoon 1 eli 100%.

Todettakoon, että R-ohjelman histogrammifunktion `hist()` oletusarvoinen luokittelutapa ei noudata edellä esitettyä ”oikeaoppista” piirtämisperiaatetta, jonka mukaan pylväiden sijainnin  $x$ -akselilla määräävät olettamamme pyöristysperiaatteen (eli pyöristys lähimpään pyöreään lukuun) mukaiset tarkat luokkarajat. Niinpä mm. tässä esimerkissä funktiokutsulla `hist(PITUUSM)` ilman lisäargumentteja saadaan histogrammi, jonka pylväiden leveydet ovat kylläkin kaikki 5 cm, mutta esim. alimpaan luokkaan tulevat sisällytetyksi kaikki pyöristetyt pituusarvot väliltä 156-160 cm, ja kuitenkin vastaava pylväs on sijoitettu  $x$ -akselilla välille  $]155, 160]$  eli käytännössä peräkkäisten luokkien pyöristettyjen ylärajojen väliin.



Tässä histogrammissa pylväiden korkeudet ovat itse asiassa samat kuin luokkien frekvenssit, mikä käy päinsä, kun luokittelu on tasavälinen.

Pienissä aineistoissa, joiden havaintoarvot ovat vahvasti pyöristettyjä, voidaan nähdä isoja satunnaisia eroja eri luokitusten pohjalta piirrettyjen histogrammien välillä. Kun aineisto on kooltaan suuri, eikä muuttujan arvoja ole voimakkaasti pyöristetty, luokittelun ja luokkarajojen yksityiskohdilla ei ole enää kovin suurta merkitystä lopputuloksen muotoon, vaikka käytettäisiin suhteellisen kapeitakin luokkia.

Huomattakoon silti, että funktiolla `hist()` saadaan piirretyksi haluttuja tarkkoja luokkarajoja noudattava histogrammi, kun funktiokutsussa lisäargumenttina annetaan vastaavien katkaisupisteiden (`breaks`) vektori esim. tyyliin `breaks = 154.5 + 5*(0:10)`. Näin on menetelty mm. ensimmäisessä histogrammissa edellä.

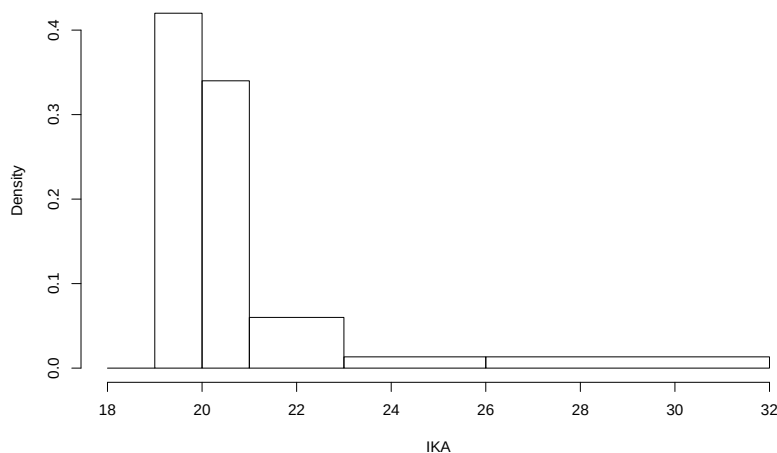
### 3.8. Vaino jakauma ja pyöristys alaspäin

Tarkastelemme nyt luokittelua ja histogrammin piirtämistä jatkuvalla muuttujalle, jonka arvot on tapana pyöristää lähimpään pienempään pyöreään lukuun (kuten ikä täysinä vuosina) ja jonka luokittelussa on syytä käyttää vaihtelevansuuruisia luokkaleveyksiä  $\Delta_k$ . Oheisessa taulukossa esitetään saman opiskelijaryhmän ikien luokiteltu frekvenssijakauma.

| Pyöris-<br>tetyt<br>luokka-<br>rajat (v) | Todelliset<br>luokkarajat<br>$a_{k-1}, a_k$ | Luo-<br>kan<br>kes-<br>kus | Abs.<br>frek-<br>venssi<br>$m_k$ | Suht.<br>frek-<br>venssi<br>$p_k$ | Summa<br>frek-<br>venssi<br>$M_k$ | Suht.<br>summa-<br>frekv.<br>$P_k$ |
|------------------------------------------|---------------------------------------------|----------------------------|----------------------------------|-----------------------------------|-----------------------------------|------------------------------------|
| 19                                       | 19, 20                                      | 19.5                       | 21                               | 42                                | 21                                | 42                                 |
| 20                                       | 20, 21                                      | 20.5                       | 17                               | 34                                | 38                                | 76                                 |
| 21-22                                    | 21, 23                                      | 22.0                       | 6                                | 12                                | 44                                | 88                                 |
| 23-25                                    | 23, 26                                      | 24.5                       | 2                                | 4                                 | 46                                | 92                                 |
| 26-31                                    | 26, 32                                      | 29                         | 4                                | 8                                 | 50                                | 100                                |
| Yhteensä                                 |                                             |                            | 50                               | 100                               |                                   |                                    |

Nyt kahden alimman luokan luokkaleveydet ovat 1 v, kolmannen leveys 2 v, neljännen 3 v, ja ylimmän luokan leveys on 5 v. Niinpä histogrammipylväiden korkeudet eli frekvenssitiheydet  $f_k$  ovat luokittain

$$\begin{aligned}
 f_1 &= 42\%/1 \text{ v} = 42 \text{ per } 100 \text{ v}, & f_2 &= 34\%/1 \text{ v} = 34 \text{ per } 100 \text{ v}, \\
 f_3 &= 12\%/2 \text{ v} = 6 \text{ per } 100 \text{ v}, & f_4 &= 4\%/3 \text{ v} = 1.33 \text{ per } 100 \text{ v}, \\
 f_5 &= 8\%/6 \text{ v} = 1.33 \text{ per } 100 \text{ v}.
 \end{aligned}$$

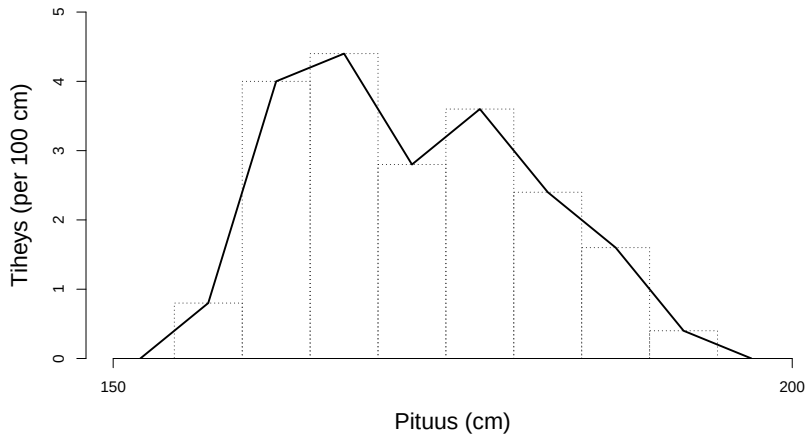


### 3.9. Frekvenssimonikulmio ja summakäyrä

Tilastotieteen sekä yliopistotason että lukion oppikirjoissa pidetään edelleen tärkeänä opeteltavana kuviotyypinä **frekvenssimonikulmiota**, vaikka sitä ei oppikirjojen ulkopuolella nykyään käytännössä juurikaan kohtaa. Kuvion ytimen muodostaa tasokoordinaatistoon piirretty murtoviiva, jonka kään-  
nepisteiden koordinaatit ovat  $(\hat{x}_k, f_k) = (\text{luokkakeskus}, \text{pylvään korkeus})$ ;  $k = 1, \dots, r$ . Murtoviiva alkaa kuitenkin alinta luokkaa edeltävän tyhjän luo-

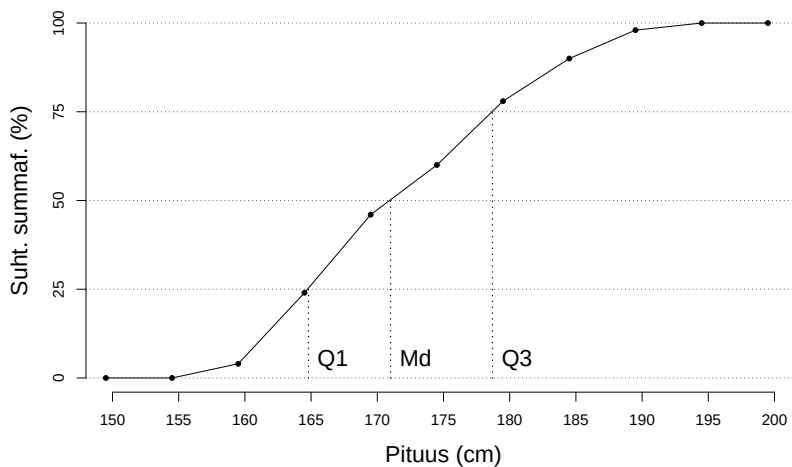
kan luokkakeskuksesta  $\hat{x}_0$  tasolta 0 ja päättyy lopulta tasolle 0 ylintä luokkaa seuraavan tyhjän luokan luokkakeskuksen  $\hat{x}_{r+1}$  kohdalla.

Ohessa pituusarvoista piirretty frekvenssipolygoni, jonka periaatetta havainnollistaa taustaksi pisteviivoin hahmotettu vastaava histogrammi.



**Summakäyrä** on murtoviivakuvi, jolla havainnollistetaan jatkuvan muuttujan summafrekvenssien kehitystä luokasta toiseen. Tämä on alkuperäisistä havainnoista piirrettävän otoskertymäfunktion porraskuvion vastine luokitellulle aineistolle, ja sitä voidaan myös pitää taustalle oletetun jatkuvan teoreettisen jakauman kertymäfunktion  $F(x)$  graafisena estimaattina.

Summakäyrän murtoviiva lähtee tasolta 0 (%) ja päättyy tasolle 1 = 100%, ja sen käännepisteet ovat  $(a_k, P_k) = (\text{todellinen yläraja, suhteellinen summafrekvenssi})$ ;  $k = 1, \dots, r$ . Kuvion avulla voi myös arvioida graafisesti medianin ja kvartiilien arvoja luokitellusta aineistosta samalla periaatteella kuin otoskertymäfunktion kuvaajastakin.



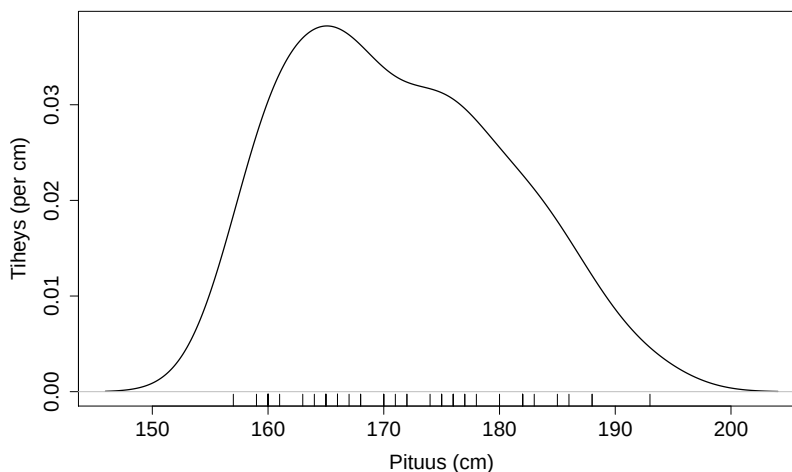
Mediaanin ja kvartiilien graafiset arviot ovat (sentin tarkkuudella)  $Md = 171$  cm,  $Q_1 = 165$  cm ja  $Q_3 = 179$  cm. Mediaanin ja yläkvartiilin arviot siis hieman poikkeavat niistä tarkoista luvuista, jotka laskettiin alkuperäisistä havainnoista.

### 3.10. Histogrammien silotus

Histogrammin (kuin myös frekvenssimonikulmion) ongelmana on kuvion muodon riippuvuus luokkarajoja ja luokkien leveyksiä koskevista valinnoista, jotka usein ovat varsin mielivaltaisia. Erityisesti pienissä ja keskisuurissa aineistoissa tiheä luokitus, joka sisällöllisesti voi olla hyvinkin perusteltavissa, tuottaa helposti satunnaisia lopputuloksia. Riittävän harvalla luokituksella päästään ehkä stabiilimpaan kuvaan, mutta tällöin jakauman jotkin olennaiset yksityiskohdat saattavat jäädä pimentoon.

Toisaalta jatkuvan muuttujan luokittelu merkitsee myös histogrammin väkivaltaista diskretoimista, ts. tiheyden muutokset luokkarajojen kohdalla näytävät epäjatkuvilta hyppäyksiltä. Tämä pykälittäisyys haittaa myös kahden eri ryhmän histogrammien vertailua samassa kuvassa.

Tietokoneiden ja ohjelmien kehittyminen on luonut mahdollisuuden piirtää histogrammin sijaan sopivalla menetelmällä **silotettu** tai **tasoitettu** (*smoothed*) histogrammi, jota voidaan pitää tavanomaista histogrammia luontevampana graafisena estimaattina taustalle oletetulle jatkuvalle tiheysfunktiolle  $f(x)$ .



Ohessa **ydinmenetelmällä** (*kernel method*) silotettu histogrammi, joka esittää pituuden tiheysfunktion  $f(x)$  ydinestimaattoria  $\hat{f}(x)$ . Ydinestimaattorin

arvo jokaisella  $x$  lasketaan havainnoista  $x_1, \dots, x_n$  seuraavasti

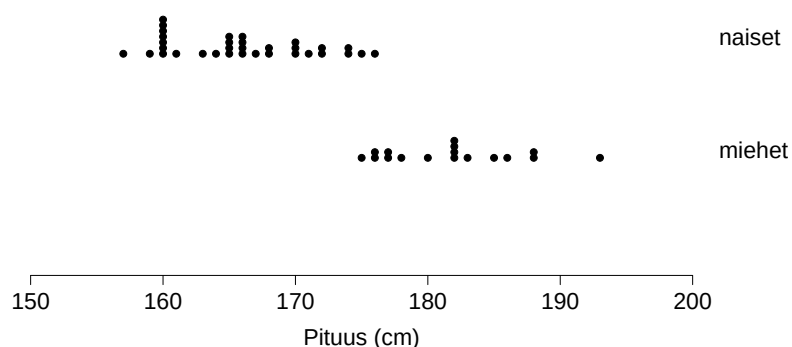
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right),$$

jossa  $K(z)$  on valittu **ydinfunktio**, ja  $h$  on **silotusparametri** tai “kaistanleveys” (*bandwidth*). Tässä käytetään Gaussin ydintä eli  $K(z) = \varphi(z)$  eli  $N(0, 1)$ -jakauman tiheysfunktioita ja kaistanleveytenä on  $h \approx 3.7$  cm. Huomaa myös  $x$ -akselin “matonrimpsut” (*rugplot*). Ne kuvaavat pistekuvion tapaan aineiston havaittuja arvoja, joiden pohjalta silotettu histogrammi on piirretty.

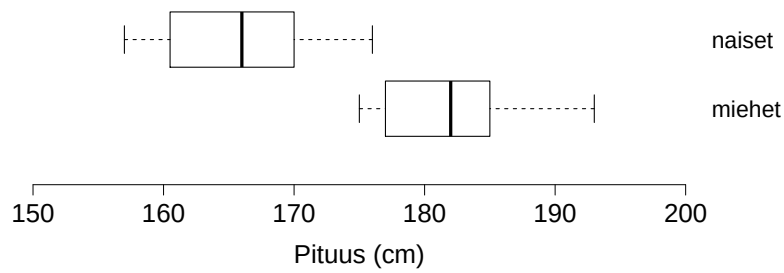
Todettakoon, että niin histogrammi ja frekvenssimonikulmio kuin myös silotettu histogrammi antavat sen vaikutelman, että pituuden jakauma tässä opiskelijajoukossa olisi sekä kaksihuippuinen että jossain määrin oikealle vino. Tällaiset piirteet saattavat syntyä pienten ja keskisuurten aineistojen pohjalta piirrettyihin kuvioihin pelkästään sattumalta, vaikka “todellinen” jakauma olisi symmetrinen ja yksihuippuinen. Tässä tapauksessa kaksihuippuisuuteen ja vinouteen on kuitenkin selkeä selitys: aineistossa on 33 naista ja 17 miestä. Kummallakin sukupuolella pituuden luonnollinen jakauma lienee varsin lähellä normaalijakaumaa.

### 3.11. Kahden ryhmän vertailu

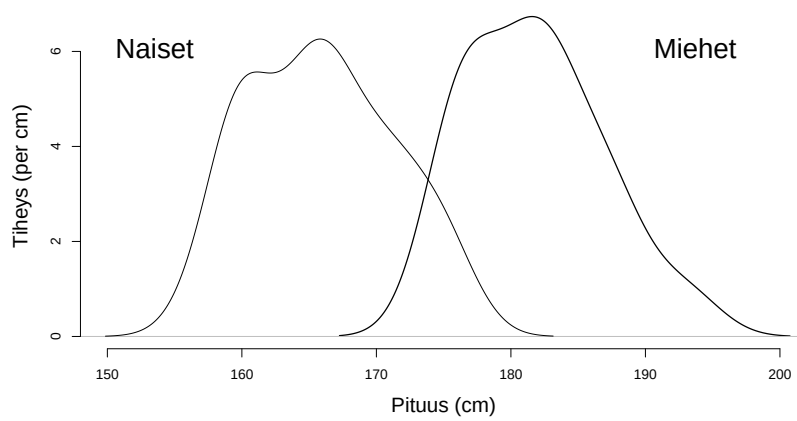
Esitämme nyt rinnakkain – tai alekkain – naisten ja miesten pituusjakaumat eri kuviotyyppejä käyttäen. Ensin ryhmittäinen pistekuvio



Sitten ryhmittäinen laatikkokuvio



Lopuksi silotetut histogrammit molemmille sukupuolille



Luokittelu- ja järjestysasteikon muuttujille havainnollisen tavan esittää prosenttijakaumia eri ryhmien välillä tarjoavat alekkain asetetut osiin jaetut palkkikuviot.

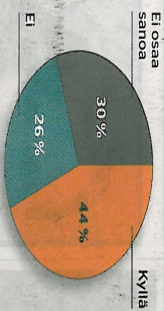
# Mielipiteet jalkautuvat

Kaleva 14.1.2011

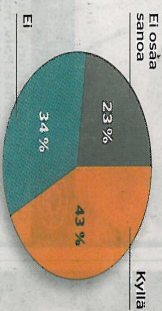
Noin puolet suomalaisista uskoo ex-pääministeri Vanhasen toimineen lainvastaisesti RAY-jupakassa Vasemmistoliittoa kannattavat kriittisimpiä. Perussuomalaisten kannat ylittävän lieviä

## Kolmanneksella ei kantaa lainvastaisuuteen

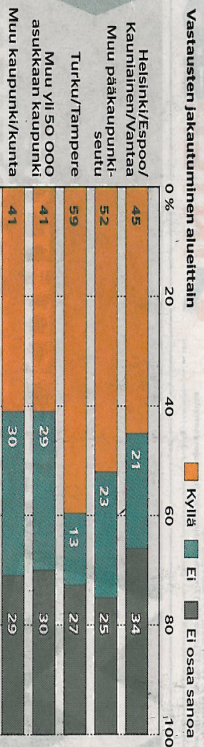
Onko Vanhasen toiminnut lainvastaisesti?



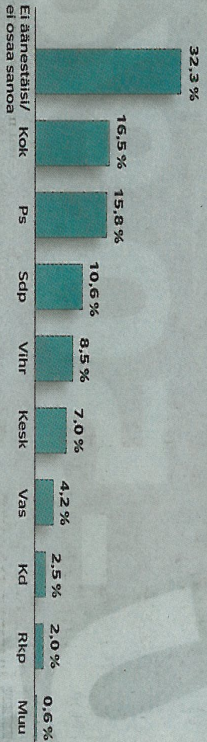
Pitäisikö eduskunnan viedä asia valtakunnallisuuden eteen?



Vastausten jakautuminen alueittain



Mitä puoluetta vastaaja äänestisi eduskuntavaaleissa?



## Yli tuhat vastasi kyselyyn

Maakuntalehden kyselyyn vastasi 1 051. Kyselytutkimuksen toteutti Taloustutkimus Oy. Gallup tehtiin puhelimitse ja internetissä tammikuussa. Tutkimuksen viitekehä on kolme prosenttia molempin suuntaan. Kysymykset olivat: Uskotteko, että Matti Vanhasen on toiminnut lainvastaisesti RAY:n tukirahoista päätettävissä? Pitäisikö eduskunnan viedä Matti Vanhasen asia valtakunnallisuuden tutkittavaksi?

### 3.12. Taulukkoja koskevat suositukset

Hyvä taulukko on sekä tiivis ja luettava että mahdollisimman paljon olennaista informaatiota välittävä. Näiden vaatimusten samanaikainen täyttäminen ei ole helppoa, ja taulukon laadinta onkin vaikeampi tehtävä kuin yleensä ajatellaan. Kaikkiin tilanteisiin yleispäteviä neuvoja ei voi antaa, mutta eräitä muodollisia periaatteita kannattaa pyrkiä noudattamaan.

- (T1) Taulukon on oltava mahdollisimman yksinkertainen ja itsensä selittävä. Se on kyettävä ymmärtämään “Aineisto ja menetelmät”-lukuun perehtymisen jälkeen ilman tukeutumista tekstiin.
- (T2) Taulukossa on otsikko, joka kertoo, mistä aineistosta tai sen osasta on kysymys ja mitä tietoja taulukko sisältää.
- (T3) Lähde on mainittava, jos tiedot ovat peräisin muusta julkaisusta tai aineistosta, jota artikkelissa ei pääsääntöisesti käsitellä.
- (T4) Jokaisella rivillä ja sarakkeella on tiivis yksikäsitteinen nimi tai otsikko (esim. ikävuosien luokitus 0-15, 15-25, 25-35, ... ei ole yksikäsitteinen mutta 0-14, 15-24, 25-34, ... on).
- (T5) Käytetyt mittayksiköt, koodit ja lyhenteet on selvitettävä, tarvittaessa alareunaan sijoitettavissa viitteissä.
- (T6) Sarakkeiden otsikot ja alaviitteet erotetaan muusta taulukosta vaakasuorin viivoin. Eri hierarkiatasojen sarakeotsikot erotetaan myös toisistaan vaakaviivoin siten, että ylemmän tason otsikon alle tuleva yhtäjaksoinen viiva kattaa vain tälle otsikolle alisteiset sarakkeet.
- (T7) Sarakkeita ei nykyisin vallitsevan käytännön mukaan erotella pystyviivoilla. Tässä tosin julkaisusarjojen käytännöt hieman vaihtelevat.
- (T8) Vertailtavat suureet kannattaa sijoittaa mahdollisimman lähelle toisiinsa. Lukujen vertailu alekkain on helpompaa kuin rinnakkain, joten jos muut näkökohdat antavat kohtuudella myöten, kiinnostavin luokittelutekijä kannattaa usein sijoittaa sisimmäksi rivimuuttujaksi.
- (T9) Tyhjiä soluja ei jätetä. Jos tieto puuttuu, se korvataan kahdella pisteellä. Jos lukua ei ole mielekkäästi olemassa (esim. eturauhassyövän tapausmäärä naisilla), merkitään yksi piste. Yhdysviiva tarkoittaa, että ao. lukumäärä tai siihen perustuva suhteellinen osuus on nolla, mutta luku 0 kuvaa arvoa, jonka suuruus on alle puolet pyöristystarkkuudesta (esim. alle 0.5 %).
- (T10) Taulukoitaessa luokiteltuja muuttujia ristiin prosenttilaskujen suunnan yleisperiaate on se, että esitetään kohde- tai vastemuuttujan arvojen prosenttijakaumat selittävän tai ennustavan muuttujan luokissa, jos otanta-asetelma sallii. Prosenttien nimittäjien täytyy myös näkyä esim. omilla riveillään tai sarakkeillaan (usein suluissa).

- (T11) Erityisesti lukumäärä- ja prosenttijakaumien yhteydessä on hyvä laskea näkyviin myös summarivejä ja -sarakkeita. Periaatteessa kaikkien yhteenlaskettavien lukujen tulisi olla näkyvissä, eli myös luokan “tieto puuttuu” (joskaan tälle luokalle ei aina välttämättä tarvitse varata omaa riviä tai saraketta, kunhan ainakin alaviitteessä ilmaistaan, kuinka moni puuttuvan tiedon omaava havaintoyksikkö sisältyi ao. rivi- tai sarakesummaan).
- (T12) Taulukon luettavuutta parantaa usein, jos rivit ja sarakkeet järjestetään niiden “suuruuden” mukaisesti (esim. reunajakaumien, summariivien tai sarakkeiden lukujen perusteella, suurinta lukua vastaava luokka ylimmälle riville tai vasemmanpuolimmaisiksi sarakkeeksi, jne.), ellei luokkien järjestystä täysin määrää painavampi sisällöllinen kriteeri.

### 3.13. Kuvioesityksiä koskevia suosituksia

Hyvien ja informatiivisten tilastollisten kuvien piirtäminen on vielä vaativampi taiteenlaji kuin taulukointi. Kuvaesitysten visuaalinen voima on suuri erityisesti muuttujien välisten yhteyksien, trendien ym. hahmottamisessa, mutta yksi kuva voi helposti valehdella enemmän kuin miljoona sanaa. Datatekniikan kehitys on tuonut jokaisen tutkijan ulottuville helppokäyttöisiä grafiikkaohjelmia, joissa on runsas valikoima yleisimpiä tyyppikuvapohjia. Monet ns. bisnesgrafiikan kuvatyypit ovat kuitenkin tieteelliseltä, tilastolliselta ja havaintopsykologiselta kannalta ala-arvoisia (esim. piirakkakuviot ja monet kolmiulotteisen vaikutelman antavat kuvat).

Kuvaesitysten osalta seuraavaa osin lähteistä Tufte (1983), Cleveland (1985), Hakulinen ja Teppo (1988) poimittua ohjelistaa voi pitää minimaalisena. (Em. ohjeet (T1) – (T3) ovat suoraan käännettävissä vastaaviksi kuvaohjeiksi.)

- (K4) Kuvan tyypin valinnassa ja sen yksityiskohtien toteutuksessa kannattaa tähdätä **data/muste-suhteen** (*data/ink ratio*) maksimointiin eli käyttää viivaa, symboleita, varjostuksia ym. säästeliäästi olennaisen informaation välittämiseksi.
- (K5) Kuva-alue täytetään mahdollisimman tehokkaasti. Akselin asteikko ulotetaan hieman ao. muuttujan havaitun vaihteluvälin ulkopuolelle. Tehokas täyttäminen voi joskus vaatia asteikon katkaisua, jota pitää kuitenkin käyttää varoen.
- (K6) Vaaka- ja pystyakseliin kuvaamat suureet mittayksiköineen on nimettävä selvästi. Asteikon jakopisteitä merkitään riittävästi (minimi on kolme) mutta ei liian tiheästi. Havaitun vaihteluvälin ulkopuolisia jakopisteitä on yksi kummallakin puolella. Jakoviivat merkitään kuva-alueesta ulospäin.

- (K7) Lukumääriä, suhteellisia osuuksia ym. ei-negatiivisia suureita kuvaava lineaariasteikko alkaa periaatteessa nolasta (ks. kuitenkin (K5)).
- (K8) Kuva-alue kannattaa usein kehystää kopioimalla vaaka-akseli alueen yläreunaan ja pystyakseli oikeaan reunaan, jakoviivat edelleen ulospäin. Jomman kumman akselin muutamia tärkeitä viitearvoja voi paikallistaa toisen akselin suuntaisilla ohuilla viivoilla yli koko kuva-alueen.
- (K9) Käytettävät symbolit, viivatyytit, rasterit, varjostukset, koodit, lyhenneet, tunnukset ym. on selitettävä lyhyesti joko otsikon yhteydessä tai tiiviillä avaimella, joka on selvästi erillään datan täyttämän alueesta.
- (K10) Eroteltaessa eri luokkia, ryhmiä ym. toisistaan täytyy käyttää hyvin toisistaan erottuvia symboleita, viivoja ym.
- (K11) Erityisen kriittinen pitää olla kolmiulotteisen vaikutelman antavien kuvien suhteen; niitä ei varsinkaan pidä käyttää silloin, kun kolmas ulottuvuus ei pidä sisällään erillistä sisällöllistä ulottuvuutta eli uutta muutujaa.

# TILASTOTIETEEN PERUSTEET, kl 2011

## Luku 4: TILASTOLLINEN PÄÄTTELY

### 4.1 Päättelytilanteita ja -kohteita

**Esim. 1.** Eri puolueiden  $k \in \{\text{kok, sd, kes, ps, vih, vas, rkp, kd, muu}\}$  kannatusosuudet  $\theta_k$  (jossa  $\sum_k \theta_k = 1$  eli 100%) äänestäjäkunnassa tällä hetkellä?

**Esim. 2.** Narkolepsiaan sairastuvien lasten ja nuorten lukumäärän  $Y_k$  odotusarvo  $\theta_k = \mathbb{E}(Y_k)$  vuosina  $k + 2005$ ,  $k = 1, 2, 3, 4$ ? Poikkeako vuoden 2010 arvo  $\theta_5$  muiden vuosien tasosta ja kuinka paljon?

**Esim. 3.** Miesopiskelijoiden pituuden  $Y_i$  ( $i = 1, 2, \dots, N$ ) jakauman keskiarvo  $\theta = \mathbb{E}(Y_i)$  ja keskihajonta  $\sigma = \sqrt{\text{var}(Y_i)}$  yliopisto- ja korkeakouluopiskelijain populaatiossa lukuvuonna 2010-11?

**Esim. 4.** Flunssaepisodien keskimääräinen kesto (vrk)  $\theta_C$ , jos ottaa säännöllisesti 1 g C-vitamiinilisän joka päivä, ja  $\theta_B$ , jos ei nauti C-vitamiinilisää päivittäin, sekä näiden erotus  $\delta = \theta_C - \theta_B$  tai suhteellinen erotus  $\rho = (\theta_B - \theta_C)/\theta_B$  nuorilla kilpauimareilla?

Suureet  $\theta, \theta_k, \theta_C, \theta_B, \sigma, \delta$  ja  $\rho$  ovat tutkijoita kiinnostavia mutta arvoltaan tuntemattomia **parametreja**. Niiden arvoista halutaan informaatiota asianmukaisen tilastollisen tutkimuksen avulla.

Kuvailevia kysymyksiä ovat esim. 1.–3. ja esim. 4. on syy-seuraussuhdetta koskeva.

### 4.2 Päätelyn kysymykset

- Millainen havaintoaineisto antaisi mahdollisimman luotettavaa (harhautonta ja tarkkaa) informaatiota tuntemattomista parametreista?
- Mikä on  $\theta$ :n ( $\theta_k$ :n, jne.) oikea tarkka arvo? Mikä on havaintoaineiston pohjalta paras yksittäinen arvio  $\theta$ :lle?
- Onko havaintoaineisto hyvin sopusoinnussa jonkin  $\theta$ :lle oletetun arvon  $\theta_0$  kanssa, vai antaako se näyttöä  $\theta_0$ :aa vastaan?
- Kuinka laaja tai kapea on sellaisten  $\theta$ :n arvojen joukko, joiden kanssa havainnot ovat hyvin sopusoinnussa?

- Miten parhaiten ennustaa saman populaation jonkin uuden havaintoyksikön saamaa arvoa kohdemuuttujalla?
- Onko analyysissä sovellettu tilastollinen malli riittävän hyvässä sopuinnussa havaintojen kanssa?

### 4.3 Tilastollinen malli

**Esim. 5.** Nastanheitto: Tapahtuma  $A_i =$  “nasta jää selälleen  $i$ . heitossa”. Heitot  $i = 1, \dots, n$  ovat riippumattomia, kussakin heitossa  $\theta = \mathbb{P}(A_i)$ . Olkoon  $Y_i = 1_{A_i}$  tapahtuman  $A_i$  ilmaisin:

$$Y_i = 1, \text{ kun } A_i \text{ toteutuu,} \quad Y_i = 0, \text{ kun } A_i^c \text{ toteutuu.}$$

Kunkin  $Y_i$  todennäköisyysjakauma on  $\text{Bern}(\theta)$ . Riippumattomuudesta seuraa, että summa

$$M = \sum_{i=1}^n Y_i = \text{“selälleen päättyneiden lkm } n \text{ toistossa”}$$

noudattaa  $\text{Bin}(n, \theta)$ -jakaumaa, jonka pistetodennäköisyydet ovat

$$\mathbb{P}(M = m; \theta) = \binom{n}{m} \theta^m (1 - \theta)^{n-m}, \quad m = 0, 1, \dots, n,$$

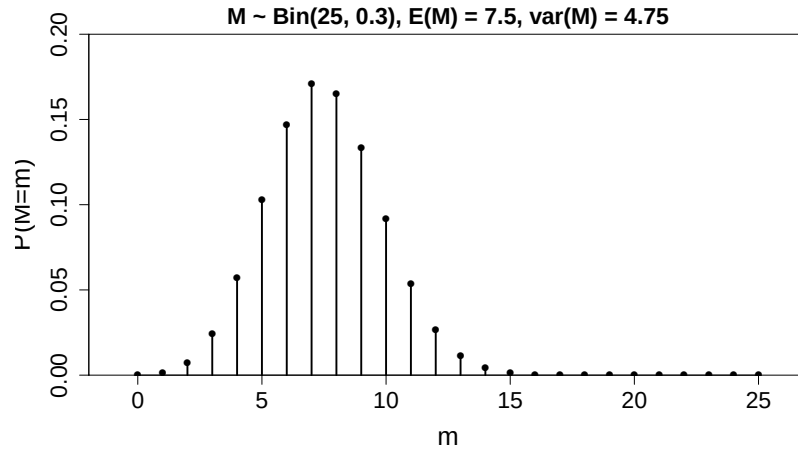
odotusarvo  $\mathbb{E}(M) = n\theta$  ja varianssi  $\text{var}(M) = n\theta(1 - \theta)$ .

Riittävän suurella  $n$ :n arvolla, samalla kun  $\theta$  ei ole liian lähellä arvoja 0 ja 1, pätee likimääräisesti (sivuuttaen ns. epäjatkuvuuskorjaus) normaaliapproksimaatio:  $M \sim N(n\theta, n\theta(1 - \theta))$  eli  $M/n \sim N(\theta, \theta(1 - \theta)/n)$ ; ts.

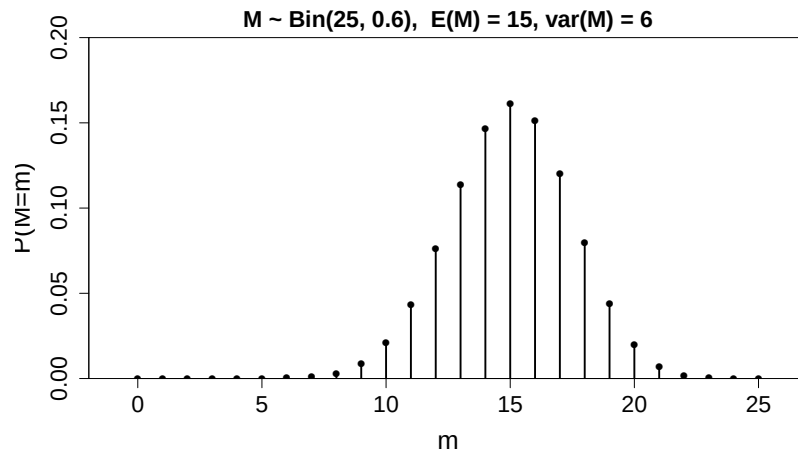
$$\mathbb{P}(M \leq m) \approx \Phi\left(\frac{m - n\theta}{\sqrt{n\theta(1 - \theta)}}\right) = \Phi\left(\frac{m/n - \theta}{\sqrt{\theta(1 - \theta)/n}}\right).$$

Todennäköisyyslaskennassa lähdetään siitä, että kun  $n \in \mathbb{N}_+$  ja  $\theta \in ]0, 1[$  ovat arvoltaan tunnettuja, niin satunnaismuuttujan  $M$  jakauma on täydellisesti määrätty, ja voidaan tarkasti laskea kaikki pistetodennäköisyydet, odotusarvo, varianssi ym. jakaumatunnusten arvot.

Esimerkiksi, kun  $n = 25$  ja  $\theta = 0.3$ , saadaan oheisenlainen pistetodennäköisyyksien  $\mathbb{P}(M = m; 0.3)$  kuvaaja, odotusarvona ja varianssina  $\mathbb{E}(M) = 7.5$  ja  $\text{var}(M) = 4.75$ .



Jos edelleen  $n = 25$  mutta nyt  $\theta = 0.6$ , niin jakauman painopiste siirtyy edelliseen verrattuna oikealle päin, ja odotusarvo ja varianssi ovat  $\mathbb{E}(M) = 15$ ,  $\text{var}(M) = 6$ .



Tilastollisessa ongelmassa lähtökohta on se, että tapahtuman  $A_i$  todennäköisyys  $\theta = \mathbb{P}(A_i)$  on tuntematon. Siitä pyritään hankkimaan informaatiota  $n$  toistosta saatavien  $A_i$ :n havaittujen esiintymisten  $y_1, \dots, y_n$  ja niiden summan  $m$  ja suhteellisen osuuden  $p = m/n$  avulla. Pienellä kirjaimella merkityt suureet ovat vastaavien isokirjaimisten satunnaismuuttujien, kuten  $M$  ja  $M/n$ , realisaatioita mallin  $\text{Bin}(n, \theta)$  puitteissa.

Tilastolliset tehtävät ja kysymykset koskien parametria  $\theta$  ovat seuraavia

- (i) **Piste-estimointi:** Mikä on se  $\theta$ :n mahdollinen arvo  $\hat{\theta}$ , joka parhaiten ennustaisi satunnaismuuttujan  $M$  havaitun arvon  $m$ ?
- (ii) **Testaus:** Onko havainto  $M = m$  riittävän hyvin sopusoinnussa jonkin  $\theta$ :a koskevan ennalta asetetun **nollahypoteesin**  $H_0 : \theta = \theta_0$  (kuten  $\theta_0 = 0.5$ ) kanssa, vai antaako aineisto näyttöä  $H_0$ :a vastaan?

- (iii) **Väliestimointi.** Mikä on sellainen  $\theta$ :n mahdollisten arvojen väli  $[\tilde{\theta}_a, \tilde{\theta}_b]$ , joiden kanssa havainto  $M = m$  on kohtuullisen hyvin sopuossuussa?

## 4.4 Uskottavuusfunktio

Tilastollisen päättelyn kysymyksiin vastaamisessa lähestymistapa perustuu keskeisesti parametrin  $\theta$  **uskottavuusfunktioon**, joka on omaksutun mallin sekä havaintojen määräämä.

Jos esim.  $\text{Bin}(n, \theta)$ -mallissa on havaittu  $M = m$ , niin uskottavuusfunktiossa  $L(\theta) = L(\theta; m)$  tarkastellaan juuri tämän arvon todennäköisyyttä erilaisilla parametrin arvoilla; ts.

$$L(\theta) = \mathbb{P}(M = m; \theta) = \binom{n}{m} \theta^m (1 - \theta)^{n-m}, \quad m = 0, 1, \dots, n,$$

Erotuksena pistetodennäköisyysfunktioon, jossa  $\theta$  on kiinteä ja tunnettu ja jossa vaihtelevana argumenttina on satunnaismuuttujan  $M$  mahdollinen realisaatio, on uskottavuusfunktiossa juuri tämä  $M$ :n havaittu arvo  $m$  kiinnitetty, mutta vaihtelevana argumenttina on parametri  $\theta$ .

**Esim.** Olli Opiskelija heitti nastaa  $n = 25$  kertaa, ja selälleen päätyneiden heittojen lukumääräksi hän sai  $m = 11$ . Tämän havainnon todennäköisyydet eräillä valituilla  $\theta$ :n arvoilla olisivat

$$\theta = 0.3 : \quad \mathbb{P}(M = 11; 0.3) = \binom{25}{11} 0.3^{11} (1 - 0.3)^{14} = 0.0536,$$

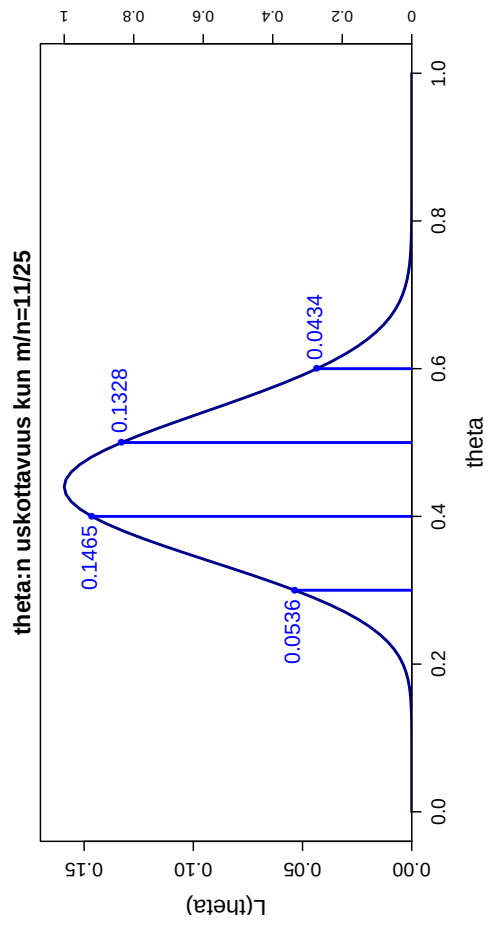
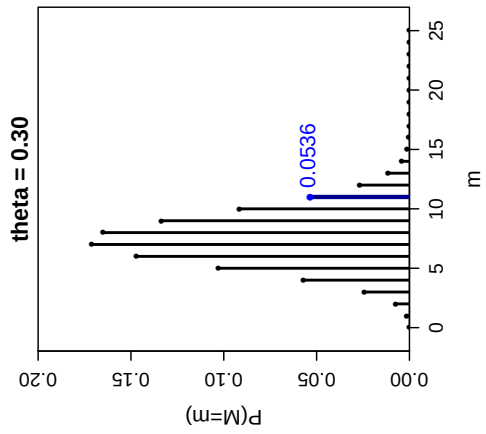
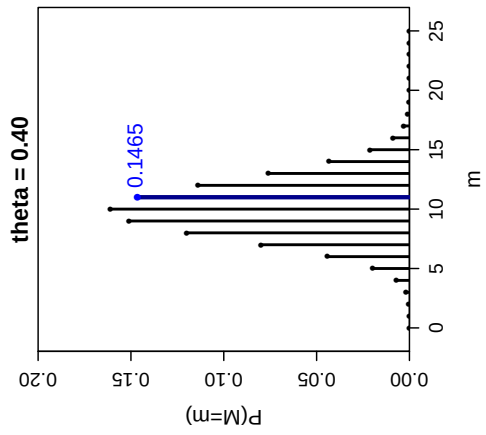
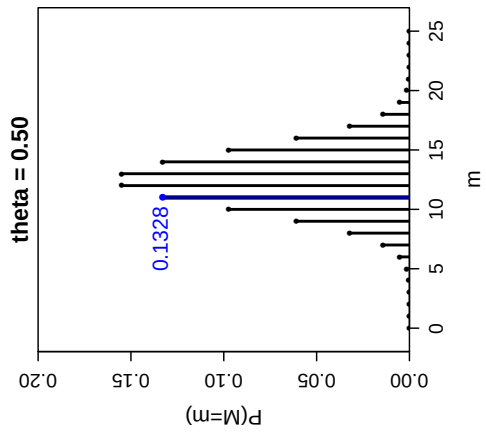
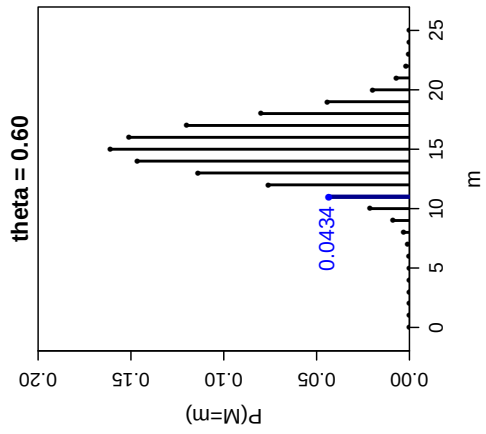
$$\theta = 0.4 : \quad \mathbb{P}(M = 11; 0.4) = \binom{25}{11} 0.4^{11} (1 - 0.4)^{14} = 0.1465,$$

$$\theta = 0.5 : \quad \mathbb{P}(M = 11; 0.5) = \binom{25}{11} 0.5^{11} (1 - 0.5)^{14} = 0.1328,$$

$$\theta = 0.6 : \quad \mathbb{P}(M = 11; 0.6) = \binom{25}{11} 0.6^{11} (1 - 0.6)^{14} = 0.0434.$$

Kun sama laskelma tehdään jokaiselle mahdolliselle  $\theta$ :n arvolle 0:n ja 1:n välillä, saadaan jatkuva uskottavuusfunktio  $L(\theta)$  täysin määräytyksi.

Seuraavalla sivulla on ensin neljän edellä käsitellyn binomijakauman  $\text{Bin}(25, \theta)$ ,  $\theta \in \{0.3, 0.4, 0.5, 0.6\}$  pistetodennäköisyysfunktioiden piikkikuvaajat. Niiden alla on tästä mallista saadun havainnon  $m = 11$  määräämän uskottavuusfunktion kuvaaja. Oikeanpuoleisella  $y$ -akselilla on suhteellisen uskottavuusfunktion  $L(\theta)/L(\hat{\theta})$  asteikko, jota pääsääntöisesti sovelletaan uskottavuusfunktion tarkasteluissa.



Uskottavuusfunktio sisältää kaiken informaation, mitä havainnot  $y_1, \dots, y_n$  ja summa  $m$  kertovat mallin  $\text{Bin}(n, \theta)$  tuntemattomasta parametrista  $\theta$ .

Kuvaajasta voi mm. arvioida, että havaintojen valossa kaikkein uskottavin  $\theta$ :n arvo on 0.44, joka on uskottavuusfunktion maksimipiste. Lisäksi varsin uskottavia arvoja ovat mm. 0.4 ja 0.5. Sen sijaan mitä kauemmaksi maksimipisteestä 0:n ja 1:n suuntaan edetään, sitä epäuskottavampiin  $\theta$ :n arvoihin päädytään tällä aineistolla.

Huomattakoon, että uskottavuusfunktiota  $L(\theta)$  ei voi tulkita parametrin  $\theta$  todennäköisyysstiheytenä, eikä sen integroiminen yli koko yksikkövälin tai sen minkään äärellisen osavälin tuota mitään tulkittavissa olevaa funktiota tai lukua.

Uskottavuusfunktion logaritmia  $l(\theta) = \log L(\theta)$  kutsutaan **log-uskottavuus-funktioksi**. Se sisältää kaiken saman informaation kuin  $L(\theta)$  mutta on monissa laskelmissa käyttökelpoisempi. Log-uskottavuusfunktion 1. derivaatta  $u(\theta) = l'(\theta)$  on parametrin  $\theta$  **pistemäärä-** tai **gradienttifunktio**, ja 2. derivaatan vastaluku  $j(\theta) = -l''(\theta)$  on puolestaan **informaatiofunktio**.

Mallin  $\text{Bin}(n, \theta)$  yhteydessä ja havainnon  $M = m$  määräämänä näiden funktioiden lausekkeet ovat

$$\begin{aligned} l(\theta) &= m \log(\theta) + (n - m) \log(1 - \theta) + \log \binom{n}{m}, \\ u(\theta) &= \frac{m}{\theta} - \frac{n - m}{1 - \theta} = \frac{m - n\theta}{\theta(1 - \theta)}, \\ j(\theta) &= \frac{m}{\theta^2} + \frac{n - m}{(1 - \theta)^2}. \end{aligned}$$

## 4.5 Piste-estimointi

Parametrin  $\theta$  **piste-estimaatti** on sopiva havaittujen arvojen  $y_1, \dots, y_n$  funktio  $t = t(y_1, \dots, y_n)$ . **Piste-estimaattori** on puolestaan vastaava satunnaismuuttuja  $T = t(Y_1, \dots, Y_n)$ . Estimaattori  $T$  on satunnaismuuttuja, jonka vaihtelua mahdollisesta samankokoisesta otoksesta toiseen kuvaa  $T$ :n teoreettinen **otantajakauma**. Tämän otantajakauman toivotaan keskittyvän parametrin oikean arvon  $\theta$  lähiympäristöön. Ero estimaattorin ja estimaatin välillä on analoginen sen kanssa, mikä on ero muuttujan ja muuttujan arvon välillä.

Jos erityisesti otantajakauman odotusarvolle pätee  $\mathbb{E}(T) = \theta$ , niin estimaattori  $T$  on **harhaton**; muussa tapauksessa harhainen. Lievä harha ei välttämättä ole ongelmallinen, kunhan estimaattori on **tarkentuva**, mikä tarkoittaa sitä, että havaintomäärän  $n$  kasvaessa odotusarvo suppenee kohti  $\theta$ :a ja otantajakauman varianssi  $\text{var}(T)$  suppenee kohti 0:aa.

Mallipohjaisessa tilastollisessa päättelyssä optimaalinen piste-estimointi perustuu **suurimman uskottavuuden menetelmään** (su-menetelmä). Se arvo  $\hat{\theta}$ , jolle pätee, että  $L(\hat{\theta}) \geq L(\theta')$  jokaisella mahdollisella arvolla  $\theta'$ , on parametrin  $\theta$  **suurimman uskottavuuden estimaatti**, lyhemmin su-estimaatti tai SUE.

Jos uskottavuusfunktio  $L(\theta)$  täyttää tietyt matemaattiset säännöllisyys ehdot, niin sillä on yksikäsitteinen maksimi. Se on samalla log-uskottavuusfunktion maksimi samoin kuin pistemääräfunktion nollakohta. Näin ollen SUE  $\hat{\theta}$  on piste, joka on estimointiyhtälön  $u(\theta) = 0$  yksikäsitteinen ratkaisu.

Bin( $n, \theta$ )-mallissa havainnon  $M = m$  pohjalta yhtälö  $u(\theta) = 0$  toteutuu täsmälleen arvolla  $\hat{\theta} = m/n$  eli "onnistuneiden toistojen" havaittu suhteellinen osuus, joka onkin tässä mallissa  $\theta$ :n su-estimaatti. Nastanheittoesimerkissämme SUE on siis  $\hat{\theta} = 11/25 = 0.44$ , kuten graafisesti voitiin jo havaita.

Yksittäisestä havaintosarjasta  $(y_1, \dots, y_n)$  estimoitu  $\hat{\theta} = \hat{\theta}(y_1, \dots, y_n)$  on siis realisaatio vastaavasta satunnaismuuttujasta, jonka arvot vaihtelisivat mahdollisesta samankokoisesta toistokoesarjasta tai otoksesta toiseen. Tätä satunnaismuuttujaa  $\hat{\theta} = \hat{\theta}(Y_1, \dots, Y_n)$  kutsutaan  $\theta$ :n **suurimman uskottavuuden estimaattoriksi**. Sitä on tapana merkitä samoin kuin yksittäistä estimaattia.

Bin( $n, \theta$ )-mallissa  $\theta$ :n su-estimaattori on  $\hat{\theta} = M/n$ , jossa siis satunnaismuuttuja  $M/n$  on "onnistuvien toistojen" havaittavissa oleva osuus.

Bin( $n, \theta$ )-mallissa su-estimaattorin  $\hat{\theta} = M/n$  otantajakauma on skaalattu binomijakauma, jonka odotusarvo on  $\mathbb{E}(M/n) = \theta$  ja varianssi on  $\text{var}(M/n) = \theta(1 - \theta)/n$ . Koska tässä tapauksessa parametrin  $\theta$  su-estimaattorin odotusarvo on  $\theta$  itse, on  $\hat{\theta}$  harhaton estimaattori  $\theta$ :lle.

Huomataan myös, että SUE:n otantajakauman varianssi lähenee 0:aa kun  $n$  kasvaa, joten estimaattorimme on myös tarkentuva. Estimoinnin satunnaisvirhe on siis sitä pienempi mitä suurempi otos on.

Keskeisen raja-arvolauseen ansiosta tätä otantajakaumaa on mahdollista approksimoida normaalijakaumalla  $N(\theta, \theta(1 - \theta)/n)$ .

Koska  $\theta$  on edelleen tuntematon, sen estimaattiin  $\hat{\theta}$  joudutaan turvautumaan mm. kun lasketaan estimoinnin satunnaisvirheen "keskimääräistä" suuruutta mittaava estimaattorin **keskivirhe** (*standard error*, SE). Tämä tunnusluku on estimaattorin otantajakauman estimoitu keskihajonta eli estimoidun varianssin neliöjuuri:  $\text{SE}(\theta) = \sqrt{\hat{\theta}(1 - \hat{\theta})/n}$ .

## 4.6 Tilastollinen testaus

Tarkastellaan edelleen mallia  $\text{Bin}(n, \theta)$ , jossa kiinnostus saattaa kohdistua siihen, onko havaintoaineisto  $(y_1, \dots, y_n)$  ja siitä laskettu SUE  $\hat{\theta}$  sopusoinnussa jonkin ennalta asetetun mahdollisen parametrin arvon  $\theta_0$  kanssa.

Jos todellisuudessa olisi niin, että oletettu  $\theta_0$  olisi  $\theta$ :n oikea arvo, niin seuraava **testisuure**

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1 - \theta_0)/n}},$$

joka on su-estimaattorin  $\hat{\theta}$ :n standardoitu muunnos, noudattaisi likimain standardinormaalijakaumaa  $N(0, 1)$ .

Näin ollen, jos siis em. oletus eli **nollahypoteesi**  $H_0 : \theta = \theta_0$  pätsisi, niin  $Z$ :n odotusarvo olisi 0, varianssi ja keskihajonta 1. Edelleen, tässä tilanteessa todennäköisyys saada  $Z$ :n arvoja esim. väliltä  $[-1, 1]$ , olisi likimain 0.683, ja välille  $[-2, 2]$  osumisen todennäköisyys olisi n. 0.954. Sen sijaan todennäköisyys saada  $H_0$ :n vallitessa itseisarvoltaan suurempia  $Z$ :n arvoja kuin 2.5 olisi vain 0.0027.

Nastanheitossa kiinnostava nollahypoteesi voisi olla symmetriaoletus eli  $H_0 : \theta = 0.5$ . Lasketaanpa tällä oletuksella suureen  $Z$  toteutunut arvo  $Z_{\text{hav}}$ , kun heittoja oli  $n = 25$  ja selälleen jääneiden havaittu lukumäärä oli 11, jolloin  $\hat{\theta} = 0.44$ :

$$Z_{\text{hav}} = \frac{0.44 - 0.5}{\sqrt{0.5(1 - 0.5)/25}} = -0.6.$$

Todennäköisyys saada havaittua  $Z$ :n arvoa itseisarvoltaan vähintään yhtä suuria poikkeamia odotusarvosta 0, on nyt

$$\mathbb{P}(|Z| \geq |Z_{\text{hav}}|) = \mathbb{P}(Z \leq -0.6) + \mathbb{P}(Z \geq 0.6) \approx 0.55,$$

mikä on kohtalaisen suuri todennäköisyys. Voi siis päätellä, että havaittu tulos ja piste-estimaatti  $\hat{\theta} = 0.44$  on hyvin sopusoinnussa oletuksen  $\theta = 0.5$  kanssa, koska vähintään havaitunsuuruuden poikkeaman todennäköisyys nollahypoteesin vallitessa on näinkin suuri.

Tätä todennäköisyyttä  $P = \mathbb{P}(|Z| \geq |z_{\text{hav}}|)$  kutsutaan yleisesti nimellä **havaittu merkitsevyystaso** eli **P-arvo**. Sitä käytetään yleisesti mittana havaintoaineiston sisältämän näytön voimakkuudesta nimen omaan nollahypoteesia vastaan: mitä pienempi  $P$ , sitä vahvempi on näyttö  $H_0$ :a vastaan. Toisaalta, suuri  $P$ , joka viittaa sopusointuun  $H_0$ :n kanssa, ei sellaisenaan vielä tarkoita sitä, että havainnot antaisivat tukevaa näyttöä erityisesti  $H_0$ :n määrittelemän pistearvon  $\theta_0$  puolesta.

Esimerkiksi nastanheitossa saamamme P-arvo  $P = 0.55$  ei sinänsä tue nollahypoteesia  $H_0 : \theta = 0.5$  sen enempää kuin vaikkapa sitä epäsymmetristä

mahdollisuutta että  $\theta$ :n arvo olisi 0.38. Toisaalta havaittu tulos  $\hat{\theta} = 0.44$  ja  $P = 0.55$  eivät tässä tilanteessa eli näin pienellä toistojen määrällä  $n = 25$  myöskään sisällä lähestulkoonkaan riittävää näyttöä symmetriaoletusta  $\theta = 0.5$  vastaan.

P-arvojen tulkinta ei ole yksikäsitteistä, vaan siinä pitää ottaa huomioon useita tilannekohtaisia seikkoja. Monilla soveltavilla tieteenaloilla on silti edelleen hyvin tavallista, että P-arvoja tulkitaan ajattelematta ja mekaanisesti jonkin mielivaltaisen ennalta asetetun merkitsevyystason kuten 0.05 perusteella tyyliin: “Jos  $P < 0.05$ , niin tulos on ‘tilastollisesti merkitsevä’ ja jos taas  $P \geq 0.05$ , niin tulos ei ole ‘tilastollisesti merkitsevä’ ”. Pahimmillaan jälkimmäinen vaihtoehto tulkitaan usein virheellisesti niin, että  $H_0$  saa havainnoista tukea.

Tällä kurssilla pyrimme välttämään tämäntapaisia yksioikoisia tulkintoja P-arvoista. Seuraavassa hyvin karkea ohjenuora:

- Jos  $P \geq 0.1$ , niin aineisto on kohtalaisesti sopusoinnussa  $H_0$ :n kanssa,
- Jos  $0.01 \leq P < 0.1$ , niin havainnot antavat heikkoa näyttöä  $H_0$ :a vastaan,
- Jos  $P < 0.01$ , niin havainnot antavat näyttöä  $H_0$ :a vastaan,

Testitulosten tulkintaan vaikuttavia osin ulkotilastollisia tekijöitä ovat mm.

- tutkimusasetelma: käytettiinkö kontrolloitua sattumaa eli havaintoaineiston hankinnassa aitoa satunnaisotantaa tai ryhmiin jakamisessa satunnaistusta, joihin testisuureen otantajakauman nimelliset ominaisuudet perustuvat,
- etukäteen sisällöllisesti tärkeäksi katsottu ero tai poikkeama nollahypoteesin mukaisesta parametrin arvosta,
- aineiston koko: suurilla  $n$ :n arvoilla on suuri todennäköisyys saada pieni P-arvo, vaikka todellinen ero ei olisi sisällöllisesti tärkeä, kun taas  $n$ :n ollessa pieni jää  $P$  helposti suureksi eikä anna tarpeeksi näyttöä sitä vastaan, vaikka todellisuudessa ero olisikin merkittävä,
- ilmiötä koskeva taustatieto tai aikaisempi kokemus.

Monissa päättelytilanteissa luottamusvälit (ks. seuraava alaluku) ovat huomattavasti informatiivisempia kuin testaaminen ja P-arvot.

## 4.7 Luottamusväli

Nyt siirrymme tarkastelemaan ns. virhemarginaalia ja sen määrittämistä. Virhemarginaalilla tarkoitetaan parametrin  $\theta$  **luottamusväliä** joka lasketaan halutulla **luottamustasolla**  $100(1 - \alpha)\%$ , jossa  $\alpha$  on pieni todennäköisyystaso. Tavanomaisia luottamustasoja ovat esim. 95% ja 90%, jolloin  $\alpha$  on vastaavasti 0.05 tai 0.1.

Luottamusvälin  $CI = [\tilde{\theta}_a, \tilde{\theta}_b]$  ala- ja yläraja  $\tilde{\theta}_a$  ja  $\tilde{\theta}_b$  ovat ennen aineiston hankintaa ja mittauksia luonteeltaan satunnaismuuttujia. Niiden arvot vaihtelisivat mahdollisesta samankokoisesta ja samalla tavalla hankitusta satunnaisotoksesta toiseen  $\tilde{\theta}$ :n otantajakauman määrittämällä tavalla siten, että väli  $CI$  peittäisi tuntemattoman parametrin  $\theta$  oikean arvon todennäköisyydellä  $1 - \alpha$ .

Toteutuneesta havaintoaineistosta  $(y_1, \dots, y_n)$  laskettu luottamusväli on edellä luonnehditun satunnaisvälin realisaatio. Sitä ei voi tulkita frekventistisen tilastotieteen viitekehyksessä seuraavasti: "todennäköisyys sille, että laskettu luottamusväli sisältää  $\theta$ :n oikean arvon, on  $1 - \alpha$ ." Omaksuttu luottamustaso kuvaa vain luottamusvälin laskentamenetelmän ominaisuutta tuottaa ao. tason mukaisella osuudella oikean parametrin arvon peittäviä välejä, jos otantaa toistettaisiin samalla otoskoolla pitkänä otossarjana.

Luottamusväli konstruoidaan tavallisesti vastaavan testisuureen avulla. Nyt kuitenkin ei kiinnittäydytä yhteen yksittäiseen nolla-arvoon  $\theta_0$  ja lasketa vain sille, kuinka suuria ovat vastaava testisuureen arvo ja P-arvo. Sen sijaan kiinnitetään luottamustaso ja sen mukainen  $\alpha$  ja haetaan kaikki sellaiset  $\theta$ :n mahdolliset arvot  $\theta'$ , jotka ovat havaintojen kanssa sopusoinnussa niin, että hypoteesia  $H' : \theta = \theta'$  vastaava P-arvo on suurempi kuin  $\alpha$ . Tämä toteutetaan seuraavalla tavalla.

Olkoon  $\theta'$  parametrin  $\theta$  oikea arvo. Tarkastellaan testisuureen  $Z = Z(\theta')$  otantajakaumaa. Vähintään likimääräisesti pätee nyt, että  $Z(\theta') \sim N(0, 1)$ , jolloin mm. pätee, että

$$\mathbb{P}(z_{\alpha/2} \leq Z(\theta') \leq z_{1-\alpha/2}) = \mathbb{P}(|Z(\theta')| \leq z_{1-\alpha/2}) = 1 - \alpha,$$

jossa  $z_u$  tarkoittaa  $N(0, 1)$ -jakauman  $u$ -fraktiilia, eli

$$\mathbb{P}(Z \leq z_u) = \Phi(z_u) = u, \quad \text{jok. } u \in ]0, 1[.$$

Kun esimerkiksi luottamustaso on 95 %, niin  $\alpha = 0.05$  ja

$$z_{\alpha/2} = z_{0.025} = -1.96, \quad z_{1-\alpha/2} = z_{0.975} = 1.96.$$

Normaalijakauman symmetrisyyden ansiosta em. todennäköisyys voidaan kirjoittaa myös muotoon

$$\mathbb{P}(|Z(\theta')| \leq z_{1-\alpha/2}) = \mathbb{P}(Z(\theta')^2 \leq z_{1-\alpha/2}^2) = 1 - \alpha.$$

Nyt tehtävänä on ratkaista 2. asteen epäyhtälö  $Z(\theta')^2 \leq z_{1-\alpha/2}^2$  tuntemattoman argumentin  $\theta'$  suhteen, ja ratkaisuksi saadaan haluttu luottamusväli  $[\tilde{\theta}_a, \tilde{\theta}_b]$ .

Bin( $n, \theta$ )-mallissa tätä periaatetta sovellettaessa päädytään epäyhtälöön

$$Z(\theta')^2 = \frac{(\hat{\theta} - \theta')^2}{\theta'(1 - \theta')/n} \leq z_{1-\alpha/2}^2$$

Tämän epäyhtälön tarkka ratkaisu (harjoitustehtävä!) tuottaa ns. Wilsonin luottamusvälin parametrille  $\theta$ . Sen ala- ja ylärajat määrittelevä tarkka lauseke on kuitenkin tarpeettoman mutkikas.

Erittäin hyvän likiarvoisen ratkaisun em. epäyhtälölle tarjoaa seuraava Agrestin ja Coullin esittämä AC-luottamusväli. Lasketaan ensin piste-estimaatin ja keskivirheen muokatut arvot

$$\tilde{\theta} = \frac{M + 2}{n + 4}, \quad \text{SE}(\tilde{\theta}) = \sqrt{\frac{\tilde{\theta}(1 - \tilde{\theta})}{n + 4}}.$$

Tämän jälkeen luottamusvälin ala- ja yläraja määräytyvät kaavoista

$$\tilde{\theta}_a = \tilde{\theta} - z_{1-\alpha/2} \times \text{SE}(\tilde{\theta}), \quad \tilde{\theta}_b = \tilde{\theta} + z_{1-\alpha/2} \times \text{SE}(\tilde{\theta}).$$

Kaikkein yksinkertaisin ja edelleen ylivoimaisesti suosituin tapa laskea likimääräinen luottamusväli Bin( $n, \theta$ )-mallin parametrille  $\theta$  perustuu muokkaamattomaan estimaattiin  $\hat{\theta}$  ja sen keskivirheeseen  $\text{SE}(\hat{\theta}) = \sqrt{\hat{\theta}(1 - \hat{\theta})/n}$ , joiden mukaiset rajat määräytyvät kaavasta

$$\hat{\theta} \pm z_{1-\alpha/2} \times \text{SE}(\hat{\theta}).$$

Pienillä otoksilla ja  $\theta$ :n arvon ollessa lähellä 0:aa tai 1:tä, tämä kaava saattaa johtaa yli laitojen meneviin epäloogisiin väleihin; ts. joko alaraja on negatiivinen tai yläraja on suurempi kuin 1.

Nastanheittoesimerkissämme, jossa  $n = 25$  ja havaittiin  $\hat{\theta} = 11/25 = 0.44$ , saadaan muokattujen suureiden arvoiksi

$$\tilde{\theta} = \frac{11 + 2}{25 + 4} = 0.4483, \quad \text{SE}(\tilde{\theta}) = \sqrt{\frac{0.4483(1 - 0.4483)}{29}} = 0.09235,$$

joiden pohjalta 95% AC-luottamusväli on

$$0.4483 \pm 1.96 \times 0.09235 = [0.2673, 0.6293].$$

Tavanomaisella kaavalla laskettu keskivirhe on

$$\text{SE}(\hat{\theta}) = \sqrt{\frac{0.44(1 - 0.44)}{25}} = 0.09928$$

ja siitä luottamusväliksi tulee

$$0.44 \pm 1.96 \times 0.09928 = [0.2454, 0.6346].$$

Tässä tapauksessa, kun piste-estimaatti on lähellä arvoa 0.5, erot eri menetelmien tuottamien välien kesken eivät ole olennaisia.

Lisätietoja binomimallin parametrin erilaisista luottamusväleistä ks. esim. [http://en.wikipedia.org/wiki/Binomial\\_proportion\\_confidence\\_interval](http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval).

## 4.8 Kahden riippumattoman otoksen vertailu

Yksittäisten populaatioiden parametrit ovat joissakin erikoistilanteissa riittävän kiinnostavia niin, että on hyödyllistä estimoida niitä ja laskea niille luottamusvälejä, ehkä jopa testata relevantteja nollahypoteeseja. Useimmat kiinnostavat tutkimuskysymykset ovat kuitenkin vertailevia.

**Esim. 1.** (jatkuu). Puolueen  $k$  kannatusosuutta tällä hetkellä halutaan usein verrata vaikkapa puolta vuotta aikaisempaan kannatusosuuteen. Kysymys asetetaan esimerkiksi, onko näyttöä siitä, että kannatusosuudessa olisi tapahtunut jotain muutosta ajankohtien välillä ja mihin suuntaan? Vähintään yhtä mielenkiintoista on tietää kannatusosuuksien mahdollisista eroista eri puolueiden välillä samana ajankohtana.

Ensimmäisessä tilanteessa vastauksen hakeminen perustuu äänestäjäkunnasta eri ajankohtina poimittuihin otoksiin, joita voidaan pitää toisistaan riippumattomina. Eri puolueiden vertailu saman otoksen sisällä puolestaan merkitsee sitä, että havaittavat kannatusosuudet  $M_k/n$  ovat toisistaan tilastollisesti riippuvat, koska osuuksien summaa yli kaikkien puolueiden sitoo lineaarinen riippuvuus:  $\sum_k M/k = 1$  eli 100%. Tämä riippuvuus pitää asianmukaisesti ottaa huomioon osuuksien erotuksen virhemarginaalia laskettaessa.

Kahden riippumattoman toistokokeen vertailuun kohdistuva malli rakennetaan seuraavasti. Olkoot  $M_1$  ja  $M_2$  toisistaan riippumattomia satunnaismuuttujia, jotka kuvaavat "onnistuneiden tulosten" lukumääriä kahdessa eri toistokokeetilanteessa, joissa "onnistumistodennäköisyydet"  $\theta_1$  ja  $\theta_2$  voivat olla erilaiset samoin kuin toistojen määrät eli otokoot  $n_1$  ja  $n_2$ . Nyt oletamme, että kummallekin  $j = 1, 2$  pätee  $M_j \sim \text{Bin}(n_j, \theta_j)$ , ja lisäksi  $M_1 M_2$ . Kiinnostavana **vertailuparametrina** tarkastelemme todennäköisyyksien erotusta  $\delta = \theta_1 - \theta_2$ . Toisinaan todennäköisyyksien vertailuparametreina käytetään myös suhdetta  $\phi = \theta_1/\theta_2$  ja vetosuhdetta  $\phi = [\theta_1/(1 - \theta_1)]/[\theta_2/(1 - \theta_2)]$

Riippumattomuudesta ja muista mallioletuksista sekä normaalijakautuneiden satunnaismuuttujien summaa koskevista jakaumatuloksista seuraa suoraan riviivaisesti, että

- $\theta_1$ :n ja  $\theta_2$ :n su-estimaattorit ovat  $\hat{\theta}_j = M_j/n_j$ ,  $j = 1, 2$ , ja ne ovat toisistaan riippumattomat,

- su-estimaattorien  $\hat{\theta}_j$  likimääräiset otantajakaumat ovat  $N(\theta_j, \theta_j(1 - \theta_j)/n_j)$ ,  $j = 1, 2$ ,
- erotuksen  $\delta = \theta_1 - \theta_2$  su-estimaattori on  $\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_2$ ,
- erotuksen su-estimaattorin  $\hat{\delta}$  likimääräinen otantajakauma on  $N(\delta, \text{var}(\hat{\delta}))$ , jossa

$$\text{var}(\hat{\delta}) = \frac{\theta_1(1 - \theta_1)}{n_1} + \frac{\theta_2(1 - \theta_2)}{n_2}.$$

Näiltä pohjilta erotusparametria  $\delta$  koskevaa nollahypoteesia  $H_0 : \delta = \delta_0$  testaava testisuure sekä  $\delta$ :n likimääräinen luottamusväli tasolla  $100(1 - \alpha)\%$  voidaan konstruoida samoja periaatteita noudattaen kuin edellisissä alaluvuissa koskien yksittäistä  $\theta$ :a. Sivuumme nyt kuitenkin matemaattiset yksityiskohdat ja toteamme, että hyvät likimääräiset tulokset saavutetaan seuraavilla menetelmillä.

Lasketaan ensin su-estimaattorin  $\hat{\delta}$  muokattu arvo sekä muokattu keskivirhe:

$$\begin{aligned} \tilde{\theta}_j &= \frac{M_j + 1}{n_j + 2}, \quad j = 1, 2, & \text{SE}(\tilde{\theta}_j) &= \sqrt{\frac{\tilde{\theta}_j(1 - \tilde{\theta}_j)}{n_j + 2}}, \\ \tilde{\delta} &= \tilde{\theta}_1 - \tilde{\theta}_2, & \text{SE}(\tilde{\delta}) &= \sqrt{\text{SE}(\tilde{\theta}_1)^2 + \text{SE}(\tilde{\theta}_2)^2}. \end{aligned}$$

Huomaa, että muokkausperiaate poikkeaa hieman siitä, mitä sovellettiin yksittäisen  $\theta$ :n analyysissä.

Nollahypoteesia  $H_0 : \delta = \delta_0$  voidaan nyt testata seuraavalla suureella:

$$Z = \frac{\hat{\delta} - \delta_0}{\text{SE}(\hat{\delta})}.$$

Siinä tilanteessa, että nollahypoteesi pätee, on tämänkin suureen likimääräinen otantajakauma standardinormaalijakauma, eli  $Z \sim N(0, 1)$ , kun  $H_0$  tosi. P-arvo haetaan myös samalla tavalla kuin alaluvussa 4.6.

Luottamusvälin ala- ja ylärajat tasolla  $100(1 - \alpha)\%$  lasketaan seuraavasta kaavasta

$$\tilde{\delta} \pm z_{1-\alpha/2} \times \text{SE}(\tilde{\delta}).$$

Yksinkertainen ja laajalti käytetty mutta pienillä otoksilla epätarkempi likiarvoinen testisuure ja luottamusväli saadaan soveltamalla muuten samoja kaavoja kuten yllä mutta korvaamalla muokatut estimaatit ja keskivirheet niiden tavanomaisilla versioilla, jossa erityisesti muokkaamaton keskivirhe erotukselle saadaan kaavoista

$$\begin{aligned} \text{SE}(\hat{\theta}_j) &= \sqrt{\frac{\hat{\theta}_j(1 - \hat{\theta}_j)}{n_j}}, \\ \text{SE}(\hat{\delta}) &= \sqrt{\text{SE}(\hat{\theta}_1)^2 + \text{SE}(\hat{\theta}_2)^2}. \end{aligned}$$

**Esim. 6.** Englannissa toteutettiin 1990-luvulla kliininen hoitokoe, jonka kohdepopulaation muodostivat kotonaan sepelvaltimotukoksen eli sydäninfarktin saaneet henkilöt. Kokeen tehtävänä oli selvittää, olisiko hyödyllistä aloittaa jo potilaan kotona tukoksen liotushoito, jonka suorittaisi paikalle kutsutun ambulanssin henkilökunta, ja voitaisiinko tällä tavalla vähentää potilaiden kuolemanvaaraa verrattuna tavanomaiseen käypään hoitoon, jossa ambulanssi vie ensin potilaan sairaalaan, ja vasta siellä aloitetaan aktiivinen hoito.

Kokeeseen satunnaistettiin yhteensä 311 potilasta, joista 163 koeryhmään (liotushoito aloitettiin jo kotona) ja 148 vertailuryhmään (aktiivinen hoito aloitettiin vasta sairaalassa). Vasteena tarkasteltiin kuolemaa 30 vrk kulussa. Tänä ajanjaksona koeryhmän kuolleisuus oli  $13/163 = 8.0\%$  ja vertailuryhmässä  $23/148 = 15.5\%$ .

Vertailuparametrin eli eri hoitovaihtoehtoihin liittyvien kuolemanriskien  $\theta_1$  ja  $\theta_2$  erotuksen  $\delta$  piste-estimaatti on  $\hat{\delta} = 0.0798 - 0.1554 = -0.0756$  eli n. 7.6 %-yksikköä. Sen keskivirhe on

$$SE = \sqrt{\frac{0.0798(1 - 0.0798)}{163} + \frac{0.1554(1 - 0.1554)}{148}} = 0.0366,$$

eli n. 3.7 %-yksikköä.

Tällaisissa vertailevissa tutkimuksissa tavanomainen nollahypoteesi ilmaistaan sanallisesti: “uusi hoitokäytäntö ei muuta potilaiden keskimääräistä ennustetta vanhaan käytäntöön verrattuna”, mikä voidaan pukea yksinkertaiseen väitteeseen  $H_0 : \delta = 0$ . Laskemme nyt tätä nollahypoteesia testaavan  $Z$ -suureen arvon sekä vastaavan kaksitahoisen  $P$ -arvon perustuen yksinkertaiseen “estimaatti/keskivirhe”-kaavaan:

$$Z = \frac{-0.0756}{0.0366} = -2.069,$$

josta  $P$ -arvoksi saadaan  $P = 2 \times (1 - \Phi(2.069)) = 0.0386$ . Kuolemanriskien erotuksen 95% likimääräinen luottamusväli kaavaa “estimaatti  $\pm$  z-fraktiili  $\times$  keskivirhe” noudattaen on

$$-0.0756 \pm 1.96 \times 0.0366 = [-0.1473, -0.0040],$$

eli  $-14.7\%$  -yksiköstä  $-0.4\%$  -yksikköön.

Koetulosten sisältämä näyttö uuden hoitokäytännön paremmuudesta on lupauksia herättävä mutta ei vielä kovin vakuuttava. Itse asiassa skeptisten asiantuntijoiden mielestä kuoleman todennäköisyyden näin suuri suhteellinen väheneminen eli  $100 \times (15.5 - 8)/15.5\% = 48\%$  alkuperäisestä tasosta ei ole tällaisenaan kovin uskottava. Kuolemanriskien erotuksen virhemarginaali on varsin leveä, ja sen oikea laita on hyvin lähellä 0:aa, joten epävarmuus kotona aloitetun liotushoidon todellisesta vaikuttavuudesta jää verrattain suureksi pelkästään tämän yksittäisen kokeen tulosten pohjalta.

Muokattuja kaavoja soveltaen saamme seuraavat tulokset:

$$\begin{aligned}\tilde{\theta}_1 &= \frac{13 + 1}{163 + 2} = 0.0848, & \tilde{\theta}_2 &= \frac{23 + 1}{148 + 2} = 0.1600 \\ \tilde{\delta} &= 0.0848 - 0.1600 = -0.0752, \\ \text{SE}(\tilde{\delta}) &= \sqrt{\frac{0.0848(1 - 0.0848)}{163 + 2} + \frac{0.16(1 - 0.16)}{148 + 2}} = 0.0370, \\ Z &= \frac{-0.0752}{0.0370} = -2.033, \\ P &= 2 \times (1 - \Phi(2.033)) = 0.042, \\ \text{CI} &= 0.0752 \pm 1.96 \times 0.0370 = [-0.1476, -0.0027].\end{aligned}$$

Eri laskentamenetelmillä saadut numeeriset tulokset eivät tässä tapauksessa poikkea olennaisesti toisistaan.

# TILASTOTIETEEN PERUSTEET, kl 2011

## Luku 5: JATKUVAN VASTEEN ANALYYSI

### 5.1 Johdanto

Tässä luvussa käsittelemme mitta-asteikoltaan jatkuvan (välimatka- tai suhdeasteikon) kohde- tai **vastemuuttujan**  $Y$  jakauman parametreja koskevaa päättelyä. Tavallisesti ensisijainen mielenkiinto kohdistuu muuttujan jakauman odotusarvoon  $\mu = \mathbb{E}(Y)$  yhdessä populaatiossa, odotusarvojen erotukseen  $\mu_1 - \mu_2$  kahden populaation tai vaihtoehtoisten olosuhteiden välillä, tai siihen, kuinka vasteen odotusarvo riippuu jostakin selittävästä muuttujasta  $X$ .

Tyypilliset mallioletukset, jotka näissä analyyseissä tehdään, ovat sellaiset, että

- vastemuuttujasta  $Y$  tehtävät havainnot  $Y_1, \dots, Y_n$  ovat toisistaan riippumattomat; ts. ne on hankittu yksinkertaisella satunnaisotoksella taustalla olevasta populaatiosta,
- kullakin havaintoyksiköllä  $i$  vasteen  $Y_i$  satunnaisvaihtelu noudattaa normaalijakaumaa odotusarvonsa ympärillä vakiovarianssilla  $\sigma^2$ .

Näiden oletusten realistisuutta ja yhteensopivuutta havaintoaineiston kanssa kannattaa arvioida kriittisesti itse kussakin sovellustilanteessa. Toisaalta keskeisen raja-arvolauseen (ks. todennäköisyyslaskennan peruskurssi) ansiosta odotusarvoja koskevassa päättelyssä muuttujan  $Y$  todellisen jakauman muodon merkitys laskelmien pätevyyyteen heikkenee sitä mukaa kuin otoskoko kasvaa. Normaalijakaumamalliin nojautuvat menetelmät ovat siis varsin **robusteja** mallioletuksista poikkeamille. Sen sijaan kriittisempiä oletuksia ovat havaintojen riippumattomuus ja oletus vakiovarianssista.

**Esim. 1.** Kuinka suuria ovat miesopiskelijoiden pituuden  $Y_i$  ( $i = 1, 2, \dots, N$ ) jakauman keski- tai odotusarvo  $\theta = \mathbb{E}(Y_i)$  ja keskihajonta  $\sigma = \sqrt{\text{var}(Y_i)}$  Oulun yliopiston opiskelijapopulaatiossa lukuvuonna 2010-11? Vuonna 2008 toteutetussa koko maan korkeakouluopiskelijoiden terveystutkimuksessa odotusarvoksi estimoitiin 180 cm. Onko odotusarvo OY:n opiskelijoiden joukossa tänä lukuvuonna mahdollisesti tästä poikkeava?

**Esim. 2.** Tarkastellaan seuraavaa satunnaistettua vertailevaa koetta. Koe-ryhmän rottia, joita oli 13 kpl, pidettiin tiety aika hyvin pölyisessä elintilassa. Vertailuryhmän 12 rotan asuinpaikan sisäilma sitä vastoin oli puhdas koko

ajan. Altistusjakson päätyttyä rotat tapettiin ja jokaisen keuhkot punnittiin. Seuraavassa keuhkojen painot (grammaa) ryhmittäin:

$$\begin{aligned} 5.4, 5.6, 5.4, 6.5, 6.8, 6.0, 4.2, 4.4, 4.8, 4.9, 6.9, 4.9, 5.2, & \quad \bar{y}_1 = 5.46, s_1 = 0.87, \\ 5.1, 3.8, 5.0, 6.4, 5.0, 4.0, 3.2, 4.4, 4.1, 5.6, 4.8, 4.5, & \quad \bar{y}_2 = 4.66, s_2 = 0.86. \end{aligned}$$

Olkoon ao. rottapopulaatiossa keuhkojen painon tuntematon odotusarvo  $\mu_2$ , kun rottia ei altisteta pölylle, ja  $\mu_1$  sen jälkeen kun ne ovat eläneet pölyn keskellä määrätyn ajan. Odotusarvojen erotus  $\delta = \mu_1 - \mu_2$  on kiinnostava vertailuparametri.

## 5.2 Yhden populaation odotusarvo, varianssi tunnettu

Olko  $y = (y_1, \dots, y_n)$  vastemuuttujasta  $Y$  tehtyjen havaintojen vektori  $n$  havaintoyksikön otoksesta. Päätelyn kohteena on vasteen  $Y$  tuntematon odotusarvo  $\mu$  taustalla olevassa populaatiossa.

Analyysi lähtee liikkeelle tavanomaisesta mallioletuksesta, jonka mukaan em. havaintosarja on realisaatio vastaavien satunnaismuuttujien jonosta tai vektorista  $(Y_1, \dots, Y_n)$ , jotka ovat toisistaan riippumattomia ja joista kukin  $Y_i \sim N(\mu, \sigma^2)$ . Tässä alaluvussa teemme sen yksinkertaistavan mutta tosielämässä harvoin realistisen oletuksen, että  $Y_i$ :den populaatiovarienssi  $\sigma^2$  on tunnettu.

Havaintosarjan  $(y_1, \dots, y_n)$  määräämä parametrin  $\mu$  uskottavuusfunktio on satunnaisvektorin  $Y = (Y_1, \dots, Y_n)$  yhteistiheysfunktion arvo havaitulla argumentin arvolla  $y = (y_1, \dots, y_n)$ , kun jakauman parametreina ovat  $\mu$  ja  $\sigma^2$ . Niinpä koko havaintosarjan määräämä uskottavuusfunktio voidaan kirjoittaa yksittäisten havaintojen  $y_i$  määräämien uskottavuuksien ( $y_i$ :den tiheyksien) tulona

$$L(\mu) = f_Y(y; \mu, \sigma^2) = \prod_{i=1}^n f(y_i; \mu, \sigma^2)$$

Koska  $Y_i \sim N(\mu, \sigma^2)$  on  $f(y_i; \mu, \sigma^2) = (1/\sqrt{2\pi\sigma^2}) \exp\{-(y_i - \mu)^2/(2\sigma^2)\}$  jok.  $i = 1, \dots, n$ , jolloin uskottavuusfunktio on tällaisten tekijöiden tulo:

$$\begin{aligned}
L(\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right\} \\
&= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \\
&= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu)^2 \right\} \\
&= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [(y_i - \bar{y})^2 + (\bar{y} - \mu)^2] \right\} \\
&= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} [n(\bar{y} - \mu)^2 + (n-1)s^2] \right\},
\end{aligned}$$

jossa havainnot tiivistyvät kahteen tyhjentävään tunnuslukuun, jotka ovat vanhat tutut otoskeskiarvo ja otosvarianssi:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Uskottavuusfunktion logaritmi on

$$l(\mu) = \log L(\mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} [n(\bar{y} - \mu)^2 + (n-1)s^2],$$

jonka ensimmäinen derivaatta  $\mu$ :n suhteen on sievennysten jälkeen muotoa

$$u(\mu) = l'(\mu) = -\frac{n}{\sigma^2} (\bar{y} - \mu).$$

Yhtälön  $s(\mu) = 0$  ainoa nollakohta on  $\bar{y}$ . Koska log-uskottavuusfunktio on  $\mu$ :n suhteen alaspäin aukeava paraabeli, on sen derivaatan nollakohta  $\hat{\mu}$  siten log-uskottavuuden ja samalla uskottavuusfunktion  $L(\mu)$  maksimoiva  $\mu$ :n arvo. Niinpä  $\mu$ :n suurimman uskottavuuden estimaatti  $\hat{\mu}$  on havainnoista laskettu otoskeskiarvo  $\bar{y}$ .

Havaintosarjasta  $y = (y_1, \dots, y_n)$  laskettu su-estimaatti  $\bar{y}$  on realisaatio vastaavasta satunnaismuuttujasta  $\bar{Y} = \sum_{i=1}^n Y_i/n$ , joka on siis  $\mu$ :n su-estimaattori ja jonka arvot vaihtelisivat mahdollisesta samankokoisesta otoksesta toiseen sen mukaan, mikä on keskiarvon  $\bar{Y}$  otantajakauma.

Todennäköisyyslaskennan peruskurssilla olemme oppineet, että riippumattomien satunnaismuuttujien  $Y_1, \dots, Y_n$  summan  $U = \sum Y_i$  odotusarvo  $\mathbb{E}(U)$  on yksittäisten  $Y_i$ :den odotusarvojen  $\mathbb{E}(Y_i)$  summa, ja summan varianssi  $\text{var}(U)$  vastaavasti varianssien  $\text{var}(Y_i)$  summa. Edelleen,  $Y_i$ :den keskiarvo on  $\bar{Y} = cU$ , jossa  $c = 1/n$  ja jolle niin muodoin pätee  $\mathbb{E}(\bar{Y}) = c\mathbb{E}(U)$  ja  $\text{var}(\bar{Y}) = c^2\text{var}(U)$ . Koska  $\mathbb{E}(Y_i) = \mu$  ja  $\text{var}(Y_i) = \sigma^2$  jok.  $i = 1, \dots, n$ ,

päädytään siihen, että  $\mathbb{E}(\bar{Y}) = \mu$  ja  $\text{var}(\bar{Y}) = \sigma^2/n$ . Lisäksi todennäköisyyslaskennassa on osoitettu, että normaalijakautuneiden muuttujien summa ja keskiarvo noudattavat myös normaalijakaumaa em. odotusarvoilla ja variansseilla. Loppupäätelmänä on siis, että näillä oletuksilla otoskeskiarvon otantajakaumalle pätee:  $\bar{Y} \sim N(\mu, \sigma^2/n)$ .

Tästä jakaumatuloksesta seuraa välittömästi, että kun  $\mu$ :n oikea arvo on  $\mu_0$ , niin keskiarvon standardoitu eli normitettu muunnos

$$Z = Z(\mu_0) = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

noudattaa jakaumaa  $N(0, 1)$ . Tämä on siten testisuureen  $Z$  otantajakauma nollahypoteesin  $H_0 : \mu = \mu_0$  vallitessa. Kaksitahoinen P-arvo saadaan samaan tapaan kuin alaluvussa 4.6:

$$P = \mathbb{P}(|Z| \geq |Z_{\text{hav}}|) = 2 \times (1 - \Phi(|Z_{\text{hav}}|)).$$

Edelleen, ratkaisemalla epäyhtälö  $Z(\mu')^2 \leq z_{1-\alpha/2}^2$  argumentin  $\mu'$  suhteen saadaan luottamustasolla  $100(1 - \alpha) \%$  olevan  $\mu$ :n luottamusvälin alaraja  $\tilde{\mu}_a$  ja yläraja  $\tilde{\mu}_b$  seuraavasti

$$\tilde{\mu}_a = \bar{Y} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \quad \tilde{\mu}_b = \bar{Y} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}.$$

**Esim. 1.** (jatkuu). OY:n opiskelijoista 2010-11 poimittiin 25 miehen otos. Heidän pituutensa mitattiin, ja keskiarvoksi saatiin 182 cm. Oletetaan populaatiossa varianssi tunnetuksi:  $\sigma^2 = 6^2 \text{ cm}^2$ . Nollahypoteesia  $H_0 : \mu = 180$  koskevan testisuureen havaittu arvo on

$$Z_{\text{hav}} = \frac{182 - 180}{6/\sqrt{25}} = 1.667.$$

Tätä vastaava P-arvo on

$$P = \mathbb{P}(Z \geq 1.667 \text{ tai } Z \leq -1.667) = 2[1 - \Phi(1.667)] = 2(1 - 0.952) = 0.096.$$

OY:n miesopiskelijain pituusjakauman odotusarvon  $\mu$  luottamusväliksi tasolla 95% saadaan

$$182 \pm 1.96 \times \frac{6}{\sqrt{25}} = [179.6, 184.4].$$

## 5.3 Yhden populaation varianssin estimointi

Tilanne, joissa varianssin  $\sigma^2$  arvo populaatiossa olisi täysin tunnettu, on lähinnä teoreettinen mutta yksinkertaisuudessaan pedagogisesti hyödyllinen. Käytännössä varianssi on odotusarvon lailla tuntematon, joten se täytyy estimoida. Varianssin estimointiin liittyvä satunnaisvirhe pitää myös ottaa huomioon odotusarvoa  $\mu$  koskevassa päättelyssä.

Jos edellisessä alaluvussa määriteltyä uskottavuusfunktiota tarkastellaan samanaikaisesti kahden parametrin eli  $\mu$ :n ja  $\sigma$ :n funktiona  $L(\mu, \sigma^2)$  ja maksimoidaan myös  $\sigma^2$ :n suhteen, päädytään su-estimaattiin  $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / n$ . Tämän asemesta on kuitenkin tapana estimoida varianssia jakamalla neliösumma luvulla  $n - 1$  eikä otoskoolla  $n$ , joten varianssin tavanomainen estimaatti  $s^2$  ja estimaattori  $S^2$  ovat

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Lukua  $n - 1$  kuvataan **vapausasteluvuksi**, joka on olemukseltaan lineaarialgebrallinen käsite. Se tarkoittaa, että kun  $n$  havainnosta  $y_1, \dots, y_{n-1}, y_n$  on laskettu keskiarvo  $\bar{y}$ , niin ns. vapaita havaintoja on enää  $n - 1$  kpl: Kun tunnetaan keskiarvo  $\bar{y}$  ja havaintoarvot  $y_1, \dots, y_{n-1}$ , niin viimeinen havainto  $y_n$  ei ole enää ”vapaa” vaan on näistä lineaarisesti riippuvainen:  $y_n = n\bar{y} - \sum_{i=1}^{n-1} y_i$ .

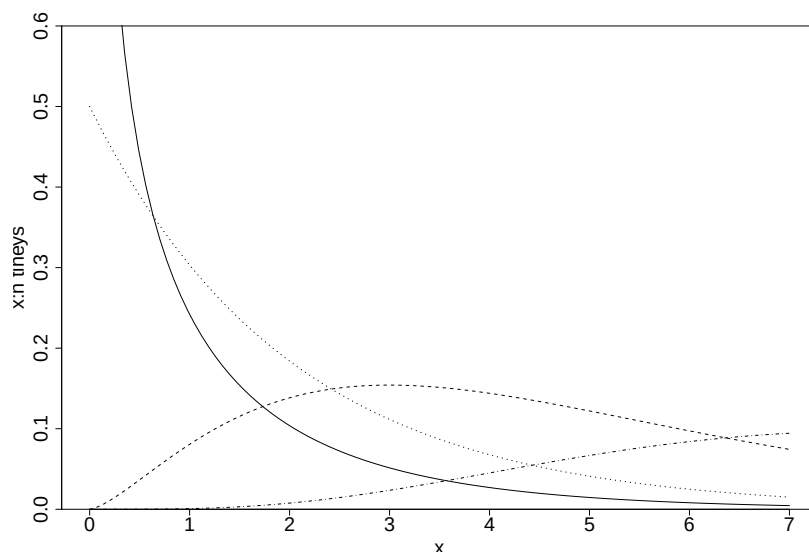
Tunnusluku  $S^2$  on varianssin  $\sigma^2$  harhaton estimaattori; ts.  $\mathbb{E}(S^2) = \sigma^2$ . Lisäksi, kun  $Y_1, \dots, Y_n$  ovat toisistaan riippumattomia satunnaismuuttujia normaalijakaumasta  $N(\mu, \sigma^2)$ , niin skaalatulle neliösummalle

$$\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

pätee, että se noudattaa **khiin neliöjakaumaa** vapausasteluvulla  $n - 1$ , jolloin merkitään  $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ .

Kaikilla mahdollisilla vapausasteluvuilla  $d = 1, 2, \dots$  khiin neliöjakauma  $\chi^2(d)$  on muodoltaan oikealle vino, sen odotusarvo on  $d$  ja varianssi  $2d$ .  $\chi^2$ -jakauman valittuja fraktiileja eri vapausasteluvuilla löytyy monista taulukkokokoelmista (mm. MAOL-taulukot). Nykyaikana käyttökelpoisempia ovat R -ohjelman funktiot `pchisq()` ja `qchisq()`, jotka antavat painettuja taulukkoja yksityiskohtaisempaa tietoa näiden jakaumien kertymäfunktioista ja fraktiileista.

Seuraavassa kuvassa khiin neliöjakauman tiheysfunktioiden kuvaajia vapausasteluvuilla  $d = 1$  (yhtenäinen viiva),  $2$  (katkoviiva),  $5$  (pisteviiva) ja  $10$  (pistekatkoviiva).



## 5.4 Yhden populaation odotusarvo, varianssi tuntematon

Kun varianssi  $\sigma^2$  ei ole tunnettu, ei voi suoraan soveltaa alaluvun 5.2 tuloksia odotusarvoa koskevassa testauksessa ja luottamusvälien laskennassa. Testisuureen  $Z = (\bar{Y} - \mu_0)(\sigma/\sqrt{n})$  sijaan voidaan kuitenkin käyttää hyvin samanlaista testisuuretta, jossa  $\sigma^2$  on korvattu sen estimaattorilla  $S^2$  eli tavanomaisella otosvarianssilla:

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}.$$

Testisuureen nimittäjässä olevaa lukua  $S/\sqrt{n}$  kutsutaan keskiarvon **keskivirheeksi** ja merkitään myös  $\text{SE}(\bar{Y})$ . Se on keskiarvon  $\bar{Y}$  otantajakauman  $N(\mu, \sigma^2/n)$  keskihajonnan  $\sigma/\sqrt{n}$  estimaattori.

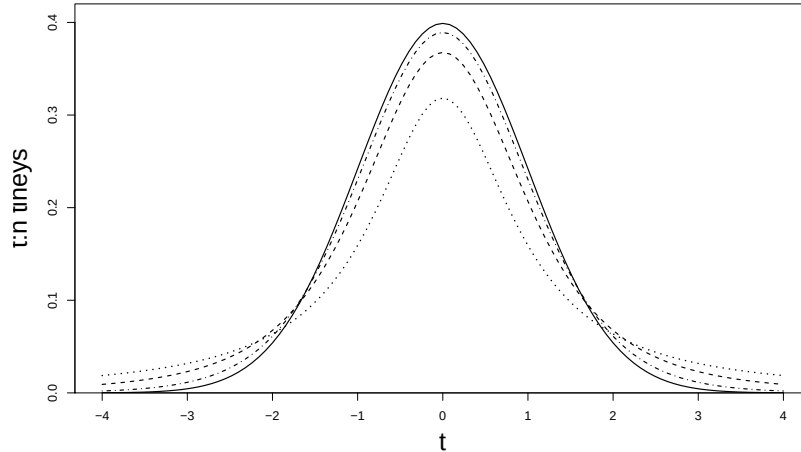
Voidaan osoittaa, että kun  $(Y_1, \dots, Y_n)$  on satunnaisotos normaalijakaumasta  $N(\mu, \sigma^2)$  ja nollahypoteesi  $H_0 : \mu = \mu_0$  pitää paikkansa, niin testisuure  $T$  noudattaa **Studentin jakaumaa** eli **t-jakaumaa** vapausasteluvulla  $n - 1$ , mitä asiantilaa voidaan merkitä esim. tyyliin  $T \sim \text{Stud}(n - 1)$ , riippumatta siitä, mikä on tuntemattoman varianssin  $\sigma^2$  arvo.

Studentin jakaumat  $\text{Stud}(d)$  kaikilla vapausasteluvuilla  $d = 1, 2, \dots$  ovat symmetrisiä ja yksihiippuisia. Niiden moodi ja mediaani ovat 0 kuten standardinormaalijakaumassakin. Toisaalta t-jakaumat ovat  $N(0, 1)$ -jakaumaa paksumuotoisempia. Jakaumalla ei esimerkiksi ole odotusarvoa, kun vapausasteluku on 1, eikä varianssia kun vapausasteita on korkeintaan 2. Kun odotusarvo ja varianssi ovat olemassa, niiden arvot ovat

$$\mathbb{E}(T) = 0, \quad \text{var}(T) = \frac{d}{d-2}.$$

Varianssi lähestyy siis arvoa 1, kun vapausasteluku kasvaa. Kun  $d \rightarrow \infty$ , Studentin jakauma yhtyy lopulta standardinormaalijakaumaan  $N(0, 1)$ .

Seuraavassa kuvassa on piirretty t-jakaumien  $\text{Stud}(d)$  tiheysfunktioiden  $f(t; d)$  kuvaajat vapausasteluvuilla  $d = 1$  (pisteviiva),  $d = 3$  (katkoviiva),  $d = 10$  (pistekatkoviiva) sekä  $d = \infty$ , jossa  $f(t; \infty) = \varphi(t)$  (yhtenäinen viiva).



Studentin jakauman valittuja fraktiileja  $t_u(d)$  eri vapausasteluvuilla löytyy monista painetuista taulukkokokoelmista (mm. MAOL-taulukot), erityisesti kun  $u \in \{0.975, 0.995, 0.9995\}$ , jotka vastaavat konventionaalisia “merkitsevyystasoja” 0.05, 0.01 ja 0.001. Käyttökelpoisia R-funktioita t-jakaumien kertymäfunktioiden arvojen sekä fraktiilien laskemiseksi ovat `pt()` ja `qt()`.

Jos siis testisuureen  $T$  havaittua arvoa merkitään  $T_{\text{hav}}$ , niin vastaava 2-tahoinen P-arvo saadaan samaan tapaan kuin testisuurelle  $Z$  (vrt. alaluku 5.2), eli

$$P = \mathbb{P}(|T| \geq |T_{\text{hav}}|) = 2 \times (1 - F(|T_{\text{hav}}|; d)),$$

jossa  $F(t; d) = \int_{-\infty}^t f(x; d) dx$  on  $\text{Stud}(d)$ -jakauman kertymäfunktion arvo pisteessä  $t$ . Edelleen, ratkaisemalla epäyhtälö  $T(\mu')^2 \leq t_{1-\alpha/2}(d)^2$  parametrin  $\mu'$  suhteen saadaan luottamustasolla  $100(1 - \alpha) \%$  olevan  $\mu$ :n luottamusvälin alaraja  $\tilde{\mu}_a$  ja yläraja  $\tilde{\mu}_b$  seuraavasti

$$\tilde{\mu}_a = \bar{Y} - t_{1-\alpha/2}(d) \times \frac{S}{\sqrt{n}}, \quad \tilde{\mu}_b = \bar{Y} + t_{1-\alpha/2}(d) \times \frac{S}{\sqrt{n}}.$$

**Esim. 1** (jatkuu). Arvioidaan edelleen OY:n miesopiskelijoiden pituusjakauman odotusarvoa  $\mu$  mutta ilman tarkkaa tietoa jakauman todellisesta varianssista  $\sigma^2$ . Hankitussa otoksessa ( $n = 25$ ) mitattujen pituuksien keskiarvoksi saatiin  $\bar{y} = 182$  cm ja keskihajonnaksi  $s = 6$  cm. Keskiarvon keskivirhe on siis  $\text{SE}(\bar{y}) = 6/\sqrt{25} = 1.2$  cm. Vapausasteluku on  $d = n - 1 = 24$ .

Testisuureen  $T$  havaituksi arvoksi saadaan näillä luvuilla

$$T_{\text{hav}} = \frac{182 - 180}{1.2} = 1.667,$$

jota vastaava P-arvo on

$$P = \mathbb{P}(T \geq 1.667 \text{ tai } T \leq -1.667) = 2[1 - F(1.667; 24)] = 0.109,$$

joka on hieman suurempi kuin jos varianssi  $\sigma^2$  olisi tunnettu ja samansuuruinen kuin otosvarienssi.

95% luottamusväliä varten tarvitaan  $\text{Stud}(24)$  jakauman 97.5 % fraktiili  $t_{0.975}(24)$  joka on 2.064. Luottamusväliksi saadaan siten

$$182 \pm 2.064 \times 1.2 = [179.52, 184.48],$$

joka on hieman leveämpi kuin populaatiovarienssin ollessa samansuuruinen mutta tunnettu.

## 5.5 Kahden populaation odotusarvojen vertailu

Kun halutaan verrata odotusarvojen  $\mu_1$  ja  $\mu_2$  erotusta  $\delta = \mu_1 - \mu_2$  kahden populaation välillä tai periaatteessa samassa populaatiossa kahden eri olosuhteen (esim. koekäsittely ja vertailukäsittely) välillä, niin tilastollinen analyysi nojautuu seuraavaan malliin.

Kummastakin populaatiosta  $r = 1, 2$  poimitaan toisistaan riippumatta satunnaisotos t. havaintoyksikköjen ryhmä kooltaan  $n_r$ , jossa vastemuuttujasta  $Y$  tehty havaintosarja on  $Y_r = (Y_{r1}, \dots, Y_{rn_r})$ . Tässä ensimmäinen indeksi  $r$  viittaa siis otokseen/ryhmään ja toinen indeksi  $i$  havaintoyksikköön ryhmän sisällä. Ryhmän  $r$  kullekin havaintoyksikölle  $i$  oletetaan malli

$$Y_{ri} \sim N(\mu_r, \sigma^2), \quad i = 1, \dots, n_r, \quad r = 1, 2.$$

Huomaa, että  $Y$ :n varianssi  $\sigma^2$  oletetaan samaksi kummassakin populaatiossa, mikä yksinkertaistaa laskukaavoja.

Mallioletuksista voidaan varsin suoraviivaisesti johtaa suurimman uskottavuuden estimaattorit odotusarvoille ja niiden erotuksille

$$\hat{\mu}_1 = \bar{Y}_1, \quad \hat{\mu}_2 = \bar{Y}_2, \quad \hat{\delta} = \bar{Y}_1 - \bar{Y}_2,$$

jossa ryhmäkeskiarvot  $\bar{Y}_1$  ja  $\bar{Y}_2$  määritellään kuten ennenkin:

$$\bar{Y}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} Y_{ri}, \quad r = 1, 2.$$

Mallioletuksista seuraa edelleen keskiarvojen otantajakaumia koskien, että  $\bar{Y}_r \sim N(\mu_r, \sigma^2/n_r)$ ,  $r = 1, 2$ . Koska otokset ovat toisistaan riippumattomia, pätee nyt keskiarvojen erotuksen  $\hat{\delta} = \bar{Y}_1 - \bar{Y}_2$  otantajakaumalle

$$\hat{\delta} \sim N\left(\delta, \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right).$$

Realistisissa sovellustilanteissa  $\sigma^2$  on tuntematon, joten se pitää estimoida. Harhaton estimaattori yhteiselle varianssille on muotoa

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

jossa  $S_r^2 = \sum_{i=1}^{n_r} (Y_{ri} - \bar{Y}_r)^2 / (n_r - 1)$  on otosvarianssi ryhmässä  $r$ ;  $r = 1, 2$ .

Erotuksen  $\hat{\delta} = \bar{Y}_1 - \bar{Y}_2$  otantajakauman keskihajonnan  $\sigma\sqrt{1/n_1 + 1/n_2}$  estimaattori on erotuksen keskivirhe

$$\text{SE}(\hat{\delta}) = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Testattaessa nollahypoteesia  $H_0 : \delta = \delta_0$  käytetään testisuurena seuraavaa  $T$ -suureen versiota

$$T = \frac{\hat{\delta} - \delta_0}{\text{SE}(\hat{\delta})}.$$

Jos mallioletukset ovat pätevät ja  $H_0$  on tosi, niin testisuure  $T$  noudattaa  $t$ -jakaumaa vapausasteluvulla  $d = n_1 + n_2 - 2$ . P-arvo katsotaan jakauman  $\text{Stud}(d)$  häntätodennäköisyyksistä kuten aiemminkin.

Luottamusväli  $\delta$ :lle luottamustasolla  $100(1 - \alpha) \%$  konstruoidaan samaan tapaan kuin ennenkin:

$$\tilde{\delta}_a = \hat{\delta} - t_{1-\alpha/2}(d) \times \text{SE}(\hat{\delta}), \quad \tilde{\delta}_b = \hat{\delta} + t_{1-\alpha/2}(d) \times \text{SE}(\hat{\delta}).$$

**Esim. 2** (jatkuu). Rottien pölyaltistuskokeessa keuhkojen painon ryhmäkohtaiset tunnusluvut (grammoina) olivat:

$$\begin{aligned} \bar{y}_1 &= 5.462, & s_1 &= 0.871, \\ \bar{y}_2 &= 4.658, & s_2 &= 0.856. \end{aligned}$$

Vertailuparametrin  $\delta$  piste-estimaatti on siten  $\hat{\delta} = 5.462 - 4.658 = 0.804$ . Yhteiseksi oletetun varianssin  $\sigma^2$  estimaatti on

$$s^2 = \frac{(13 - 1) \cdot 0.871^2 + (12 - 1) \cdot 0.856^2}{13 + 12 - 2} = 0.864^2,$$

jossa vapausasteluku on siis  $d = 13 + 12 - 2 = 23$ . Keskiarvojen erotuksen havaittu keskivirhe on  $\text{SE}(\hat{\delta}) = 0.864\sqrt{1/13 + 1/12} = 0.346$  grammaa.

Nollahypoteesia  $H_0 : \delta = 0$  eli “pölyaltistus ei vaikuta rottien keuhkojen kokoon” testattaessa saadaan  $T$ -suureen arvoksi

$$T_{\text{hav}} = \frac{0.804}{0.346} = 2.322,$$

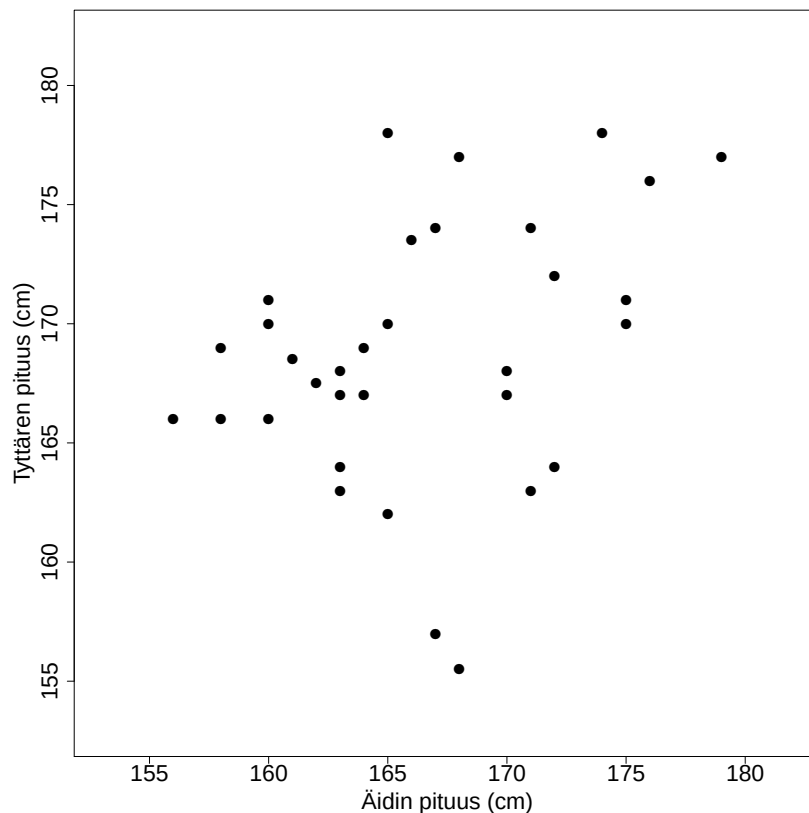
joka antaa P-arvoksi  $P = 2[1 - F(2.322; 23)] = 0.03$ . 95% luottamusväliksi  $\delta$ :lle tulee

$$0.804 \pm 2.069 \times 0.346 = [0.087, 1.519].$$

Koetulokset antavat jossain määrin näyttöä nollahypoteesia vastaan eli sen puolesta, että pölyaltistus todella lisää keuhkokudoksen painoa. Rottien välinen yksilöllinen satunnaisvaihtelu aiheuttaa kuitenkin sen, että näin pienillä koe- ja vertailuryhmillä vaikutusta kuvaava virhemarginaali jää varsin leveäksi.

## 5.6 Regressio ja korrelaatio

**Esim. 3.** Francis Galton tutki 1800-luvun puolivälissä erilaisten biologisten ominaisuuksien periytymistä. Galton oli kiinnostunut mm. siitä, kuinka vahvasti täysikasvuisen ihmisen pituus riippuu hänen vanhempiansa pituudesta. Oheisessa sirontakuviossa on kuvattu 33 naisopiskelijalta saadut havaintopisteet  $(x_i, y_i)$ , jossa  $x_i$  = opiskelijan  $i$  äidin pituus (cm) ja  $y_i$  on opiskelijan oma pituus (cm).



Tyttären pituus  $Y$  on tässä tarkastelussa **vastemuuttuja** eli selitettävä tai ennustettava muuttuja, ja äidin pituus  $X$  on **selittävä muuttuja** tai **ennustetekijä**.  $X$ :n ja  $Y$ :n välisestä riippuvuudesta voidaan esittää mm. seuraavia kysymyksiä:

- minkä muotoinen ja kuinka vahva riippuvuus on,
- kuinka tarkasti voidaan ennustaa  $Y$ :n arvoa, jos tunnetaan  $X$ :n arvo?

Kun vastemuuttuja  $Y$  on jatkuva ja vähintään välimatka-asteikollinen muuttuja ja  $X$  on myös välimatka- tai suhdeasteikon muuttuja, niin muuttujien välistä yhteyttä on ensimmäisenä tapana mallittaa yksinkertaisen **lineaarisen regressiomallin** avulla. Siinä oletetaan, että kullakin havaintoyksiköllä  $i$  vaste  $Y_i$  noudattaa toisista havaintoyksiköistä riippumatta jakaumaa  $N(\mu_i, \sigma^2)$ , jossa vasteen odotusarvo  $\mu_i = \mathbb{E}(Y_i)$  riippuu havaintoyksikön  $i$  saamasta selittävän muuttujan  $X$  arvosta  $X_i$ :

$$\mu_i = \alpha + \beta X_i, \quad i = 1, \dots, n.$$

Tämä on **regressiosuoran** yhtälö, jota odotusarvot noudattavat. Parametrit  $\alpha$  ja  $\beta$  ovat mallin **regressiokertomia**. Niistä  $\beta$  on regressiosuoran kulma-kerroin, joka tulkitaan

$$\beta = \frac{\Delta\mu}{\Delta X},$$

eli vasteen  $Y$  keskimääräinen muutos selittävän muuttujan yksikkömuutosta kohti. Parametri  $\alpha$  on **vakiokerroin**, joka kuvaa  $Y$ :n odotusarvoa siinä usein hypoteettisessa tilanteessa, jossa  $X = 0$ .

Malli voidaan myös esittää muodossa

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

jossa  $\varepsilon_i$  on ns. **virhetermi**, jonka jakaumasta oletetaan:  $\varepsilon_i \sim N(0, \sigma^2)$ .

Tässä yksinkertaisessa mallissa vasteen vaihtelun odotusarvon ympärillä ei anneta riippua  $X$ :n arvosta, vaan **virhevarianssi**  $\sigma^2 = \text{var}(\varepsilon_i)$  oletetaan vakioksi.

Havaintoparien  $(x_i, y_i)$  määräämä uskottavuusfunktio on nyt seuraavanlainen

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \mu_i)^2 \right\} \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2 \right\}. \end{aligned}$$

Uskottavuusfunktion logaritmi on

$$l(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2.$$

Log-uskottavuuden maksimointitehtävä  $\alpha$ :n ja  $\beta$ :n suhteen on yhtäpitävä sen kanssa, että on minimoitava neliösumma

$$q(\alpha, \beta) = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2.$$

Tästä päädytään **pienimmän neliösumman menetelmään** (pns) parametrien  $\alpha$  ja  $\beta$  estimoinnissa.

Merkitään aluksi  $X$ :n ja  $Y$ :n otoskeskiarvoja ja variansseja

$$\begin{aligned} \bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i, & S_X^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \\ \bar{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i, & S_Y^2 &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

Määritellään nyt  $X$ :n ja  $Y$ :n yhteisvaihtelua ja lineaarisen riippuvuuden voimakkuutta kuvaavat tunnusluvut **kovarianssi**  $S_{XY}$  ja **korrelaatiokerroin**  $R$  seuraavasti

$$\begin{aligned} S_{XY} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})Y_i, \\ R &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{S_{XY}}{S_X S_Y}. \end{aligned}$$

Korrelaatiokerroin on laaduton luku, jonka arvot ovat välillä  $[-1, +1]$ , ja  $|R| = 1$  täsmälleen silloin, kun  $X$ :n ja  $Y$ :n välillä vallitsee täydellinen lineaarinen riippuvuus; ts.  $Y_i = \mathbb{E}(Y_i) = \alpha + \beta X_i$  todennäköisyydellä 1. Lineaarinen riippuvuus on positiivinen kun  $R > 0$ , se on negatiivinen kun  $R < 0$ , ja lineaarista riippuvuutta ei ole lainkaan kun  $R = 0$ .

Todettakoon, että  $R$  mittaa nimen omaan lineaarisen riippuvuuden voimakkuutta ja sitä ominaisuutta parhaimmillaan, kun muuttujaparin  $(X, Y)$  voidaan olettaa noudattavan kaksiulotteista normaalijakaumaa. Jos sen sijaan yhteys  $X$ :n ja  $Y$ :n välillä on epälineaarinen, tai esim.  $X$ :n jakauma on vahvasti vino, niin korrelaatiokerroin on huono riippuvuuden mitta. Lisäksi korrelaatiokerroin menettää tulkittavuutensa, jos koesuunnittelun tai otannan tehokkuusnäkökohtien johdosta  $X$ :n jakaumaa on manipuloitu niin, että sen hajonta on suurempi kuin jossain luonnollisessa populaatiossa.

Palataan regressiomallin parametrien estimointiin. Neliösumman  $q(\alpha, \beta)$  minimointi tuottaa regressiokertoimien pns-estimaattorien lausekkeiksi

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2} = R \frac{S_Y}{S_X},$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

Kullakin havaintoyksiköllä estimoidut regressiokertoimet määräävät vasteen **sovitteen**  $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$ , jotka myös noudattavat suoran yhtälöä. Tässä tapauksessa kyse on havaintojen perusteella estimoidusta tai sovitetusta regressiosuorasta.

**Esim. 3.** Äitien ja tyttärien pituuksien (cm) keskiarvot ja otosvarianssit ovat

$$\bar{X} = 166.4, \quad S_X^2 = 5.8^2, \quad \bar{Y} = 168.8, \quad S_Y^2 = 5.5^2.$$

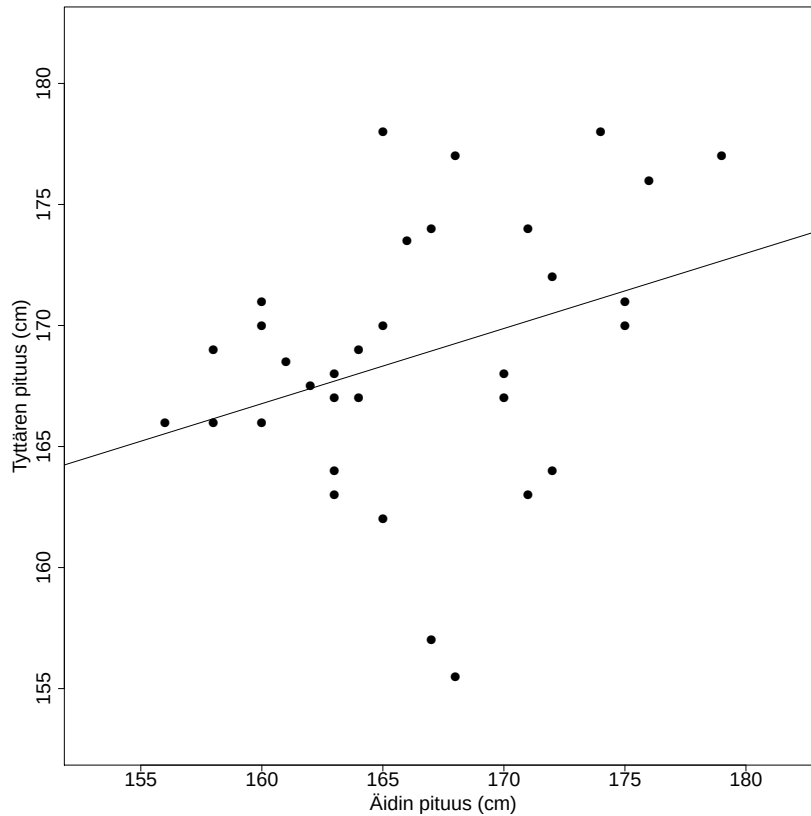
Kovarianssin ja korrelaatiokertoimen arvoiksi tulee

$$S_{XY} = 10.5 \text{ cm}^2, \quad R = 0.328,$$

ja regressiokertoimien estimaateiksi

$$\hat{\beta} = \frac{10.5}{5.8^2} = 0.31, \quad \hat{\alpha} = 168.8 - 0.31 \times 166.4 = 117.1.$$

Sovitettu regressiosuora on seuraavassa piirretty havaintopisteiden muodostamaan sirontakuvioon.



Jos esimerkiksi äidin pituus on ollut 160 cm, niin tyttären pituuden sovite eli estimoitu odotusarvo on  $117.1 \text{ cm} + 0.31 \times 160 \text{ cm} = 166.7 \text{ cm}$ . Äidin pituuden ollessa 175 cm tyttären pituuden sovite on puolestaan  $117.1 \text{ cm} + 0.31 \times 175 \text{ cm} = 171.4 \text{ cm}$ .

Huomaa, kuinka tässäkin aineistossa toteutuu jo Galtonin havaitsema ilmiö nimeltä *regressio keskiarvoa kohti*. Keskimääräistä pitempien äitien tyttärien odotuspituudet ovat myös ikäluokkansa naisten keskiarvoa suurempia, mutta eivät niin paljon oman sukupolvensa keskiarvosta poikkeavia kuin äitinsä pituudet. Sama pätee keskimääräistä lyhyempien äitien tyttärille.

Regressiosuoran kulmakerroin ja korrelaatiokerroin kuvaavat eri asioita. Regressiokulmakerroin kertoo, kuinka paljon  $Y$ :n odotusarvo muuttuu, kun  $X$ :n arvo lisääntyy yhdellä mittayksiköllä. Korrelaatiokerroin puolestaan mittaa lineaarisen riippuvuuden voimakkuutta  $X$ :n ja  $Y$ :n välillä, eli sitä kuinka vahvasti sirontakuviassa havaintopisteet ovat keskittyneet regressiosuoran ympärille.

Jäännösvarianssin  $\sigma^2$  harhaton estimaattori on mallin **jäännöstermien** eli **residuaalien**  $E_i = Y_i - \hat{Y}_i$  keskineliösumma

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

eli jäännösneliösumma jaettuna jäännösvapausasteiden lukumäärällä  $n-2$ , jossa jälkimmäinen määräytyy siitä, että regressiokertoimien estimoinnin jälkeen on olemassa  $n-2$  vapaata havaintoa.

Regressiokertoimien keskivirheet ovat

$$\text{SE}(\hat{\beta}) = \frac{S}{\sqrt{(n-1)S_X^2}}, \quad \text{SE}(\hat{\alpha}) = S \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}},$$

jossa  $S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2$  on muuttujan  $X$  otosvarianssi.

Olkoon  $H_0 : \beta = \beta_0$  kulmakerrointa koskeva nollahypoteesi. Jos havainnoille tehty mallioletus  $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$  on pätevä ja nollahypoteesi on tosi, niin testisuure

$$T = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}$$

noudattaa Studentin jakaumaa vapausastein  $n-2$ . Tältä pohjalta luottamustavalli  $\beta$ :lle luottamustasolla  $100(1-\gamma) \%$  on

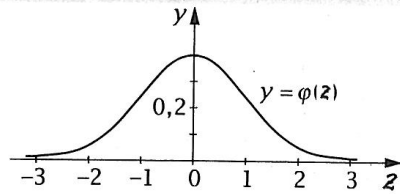
$$\hat{\beta} \pm t_{1-\gamma/2}(n-2) \times \text{SE}(\hat{\beta})$$

**Esim. 3.** Tyttöjen pituuksien jäännösvarianssi oli  $S^2 = 5.28^2 \text{ cm}^2$ . Estimoidun kulmakertoimen  $\hat{\beta}$  keskivirhe oli  $\text{SE}(\hat{\beta}) = 5.28 / \sqrt{(33-1)5.28^2} = 0.16 \text{ cm}$ .

Jos nollahypoteesina pidetään  $H_0 : \beta = 0$ , eli “äidin pituudella ja tyttären pituudella ei ole mitään yhteyttä”, niin  $T$ -suureen havaittu arvo on  $T_{\text{hav}} = 0.31/0.16 = 1.93$ , jota vastaava  $P$ -arvo on saadaan Stud(31)-jakaumasta:  $P = 0.063$ . Luottamusväli  $\beta$ :lle luottamustasolla 95% on  $0.31 \pm 2.04 \times 0.16 = [-0.02, 0.64]$ . Tässä aineistossa näyttö äidin ja tyttären pituuksien välisestä riippuvuudesta jää varsin heikoksi, mikä on selitettävissä ainakin aineiston pienuudella. Havaitun yhteyden heikkouteen ovat myös saattaneet vaikuttaa mittausvirheet sekä äidin että tyttärien pituuslukemissa.

## Normaalijakauman tiheysfunktio

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

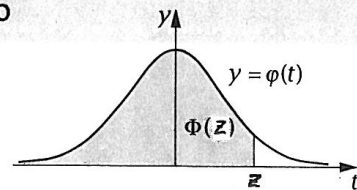


| z  | 0     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|----|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0, | 0,399 | 397 | 391 | 381 | 368 | 352 | 333 | 312 | 290 | 266 |
| 1, | 242   | 218 | 194 | 171 | 150 | 130 | 111 | 094 | 079 | 066 |
| 2, | 054   | 044 | 035 | 028 | 022 | 018 | 014 | 010 | 008 | 006 |
| 3, | 004   | 003 | 002 | 002 | 001 | 001 | 001 | 000 | 000 | 000 |

## Normaalijakauman kertymäfunktio

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{t^2}{2}} dt$$

$$\Phi(a) = P(X \leq a), \quad \Phi(-a) = 1 - \Phi(a)$$



| z   | 0      | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    |
|-----|--------|------|------|------|------|------|------|------|------|------|
| 0,0 | 0,5000 | 5040 | 5080 | 5120 | 5160 | 5199 | 5239 | 5279 | 5319 | 5359 |
| 0,1 | 5398   | 5438 | 5478 | 5517 | 5557 | 5596 | 5636 | 5675 | 5714 | 5753 |
| 0,2 | 5793   | 5832 | 5871 | 5910 | 5948 | 5987 | 6026 | 6064 | 6103 | 6141 |
| 0,3 | 6179   | 6217 | 6255 | 6293 | 6331 | 6368 | 6406 | 6443 | 6480 | 6517 |
| 0,4 | 6554   | 6591 | 6628 | 6664 | 6700 | 6736 | 6772 | 6808 | 6844 | 6879 |
| 0,5 | 6915   | 6950 | 6985 | 7019 | 7054 | 7088 | 7123 | 7157 | 7190 | 7224 |
| 0,6 | 7257   | 7291 | 7324 | 7357 | 7389 | 7422 | 7454 | 7486 | 7517 | 7549 |
| 0,7 | 7580   | 7611 | 7642 | 7673 | 7704 | 7734 | 7764 | 7794 | 7823 | 7852 |
| 0,8 | 7881   | 7910 | 7939 | 7967 | 7995 | 8023 | 8051 | 8078 | 8106 | 8133 |
| 0,9 | 8159   | 8186 | 8212 | 8238 | 8264 | 8289 | 8315 | 8340 | 8365 | 8389 |
| 1,0 | 8413   | 8438 | 8461 | 8485 | 8508 | 8531 | 8554 | 8577 | 8599 | 8621 |
| 1,1 | 8643   | 8665 | 8686 | 8708 | 8729 | 8749 | 8770 | 8790 | 8810 | 8830 |
| 1,2 | 8849   | 8869 | 8888 | 8907 | 8925 | 8944 | 8962 | 8980 | 8997 | 9015 |
| 1,3 | 9032   | 9049 | 9066 | 9082 | 9099 | 9115 | 9131 | 9147 | 9162 | 9177 |
| 1,4 | 9192   | 9207 | 9222 | 9236 | 9251 | 9265 | 9279 | 9292 | 9306 | 9319 |
| 1,5 | 9332   | 9345 | 9357 | 9370 | 9382 | 9394 | 9406 | 9418 | 9429 | 9441 |
| 1,6 | 9452   | 9463 | 9474 | 9484 | 9495 | 9505 | 9515 | 9525 | 9535 | 9545 |
| 1,7 | 9554   | 9564 | 9573 | 9582 | 9591 | 9599 | 9608 | 9616 | 9625 | 9633 |
| 1,8 | 9641   | 9649 | 9656 | 9664 | 9671 | 9678 | 9686 | 9693 | 9699 | 9706 |
| 1,9 | 9713   | 9719 | 9726 | 9732 | 9738 | 9744 | 9750 | 9756 | 9761 | 9767 |
| 2,0 | 9772   | 9778 | 9783 | 9788 | 9793 | 9798 | 9803 | 9808 | 9812 | 9817 |
| 2,1 | 9821   | 9826 | 9830 | 9834 | 9838 | 9842 | 9846 | 9850 | 9854 | 9857 |
| 2,2 | 9861   | 9864 | 9868 | 9871 | 9875 | 9878 | 9881 | 9884 | 9887 | 9890 |
| 2,3 | 9893   | 9896 | 9898 | 9901 | 9904 | 9906 | 9909 | 9911 | 9913 | 9916 |
| 2,4 | 9918   | 9920 | 9922 | 9925 | 9927 | 9929 | 9931 | 9932 | 9934 | 9936 |
| 2,5 | 9938   | 9940 | 9941 | 9943 | 9945 | 9946 | 9948 | 9949 | 9951 | 9952 |
| 2,6 | 9953   | 9955 | 9956 | 9957 | 9959 | 9960 | 9961 | 9962 | 9963 | 9964 |
| 2,7 | 9965   | 9966 | 9967 | 9968 | 9969 | 9970 | 9971 | 9972 | 9973 | 9974 |
| 2,8 | 9974   | 9975 | 9976 | 9977 | 9977 | 9978 | 9979 | 9979 | 9980 | 9981 |
| 2,9 | 9981   | 9982 | 9983 | 9983 | 9984 | 9984 | 9985 | 9985 | 9986 | 9986 |
| 3,0 | 9987   | 9987 | 9987 | 9988 | 9988 | 9989 | 9989 | 9989 | 9990 | 9990 |
| 3,1 | 9990   | 9991 | 9991 | 9991 | 9992 | 9992 | 9992 | 9992 | 9993 | 9993 |
| 3,2 | 9993   | 9993 | 9994 | 9994 | 9994 | 9994 | 9994 | 9995 | 9995 | 9995 |
| 3,3 | 9995   | 9995 | 9995 | 9996 | 9996 | 9996 | 9996 | 9996 | 9996 | 9997 |
| 3,4 | 9997   | 9997 | 9997 | 9997 | 9997 | 9997 | 9997 | 9997 | 9997 | 9998 |

| Φ(z) | 0,90   | 0,95   | 0,975  | 0,990  | 0,995  | 0,999  | 0,9995 | 0,9999 | 0,99995 |
|------|--------|--------|--------|--------|--------|--------|--------|--------|---------|
| z    | 1,2816 | 1,6449 | 1,9600 | 2,3264 | 2,5758 | 3,0902 | 3,2905 | 3,7190 | 3,8906  |

Studentin jakauma eli t-jakauma

Studentin jakauman fraktilleja  $t_u(d)$  eri vapausasteluvuilla  $d$ , kun  $0.5 < u < 0.9996$

Esim. Kun  $d = 5$  ja  $u = 0.975$  on  $t_{0.975}(5) = 2.57$ , jolloin pätee, että

(a)  $P(T < 2.57) = 0.975$ , (b)  $P(T > 2.57) = 0.025$ , (c)  $P(|T| > 2.57) = 0.05$ ,

(d)  $P(-2.57 < T < 2.57) = 0.95$

| d      | u    |      |      |      |      |      |       |      |       |       |       |        |       |        |        |        |
|--------|------|------|------|------|------|------|-------|------|-------|-------|-------|--------|-------|--------|--------|--------|
|        | 0.6  | 0.7  | 0.75 | 0.8  | 0.85 | 0.9  | 0.925 | 0.95 | 0.975 | 0.985 | 0.99  | 0.9925 | 0.995 | 0.9975 | 0.999  | 0.9995 |
| 1      | 0.32 | 0.73 | 1.00 | 1.38 | 1.96 | 3.08 | 4.17  | 6.31 | 12.71 | 21.20 | 31.82 | 42.43  | 63.66 | 127.32 | 318.31 | 636.62 |
| 2      | 0.29 | 0.62 | 0.82 | 1.06 | 1.39 | 1.89 | 2.28  | 2.92 | 4.30  | 5.64  | 6.96  | 8.07   | 9.92  | 14.09  | 22.33  | 31.60  |
| 3      | 0.28 | 0.58 | 0.76 | 0.98 | 1.25 | 1.64 | 1.92  | 2.35 | 3.18  | 3.90  | 4.54  | 5.05   | 5.84  | 7.45   | 10.21  | 12.92  |
| 4      | 0.27 | 0.57 | 0.74 | 0.94 | 1.19 | 1.53 | 1.78  | 2.13 | 2.78  | 3.30  | 3.75  | 4.09   | 4.60  | 5.60   | 7.17   | 8.61   |
| 5      | 0.27 | 0.56 | 0.73 | 0.92 | 1.16 | 1.48 | 1.70  | 2.02 | 2.57  | 3.00  | 3.36  | 3.63   | 4.03  | 4.77   | 5.89   | 6.87   |
| 6      | 0.26 | 0.55 | 0.72 | 0.91 | 1.13 | 1.44 | 1.65  | 1.94 | 2.45  | 2.83  | 3.14  | 3.37   | 3.71  | 4.32   | 5.21   | 5.96   |
| 7      | 0.26 | 0.55 | 0.71 | 0.90 | 1.12 | 1.41 | 1.62  | 1.89 | 2.36  | 2.71  | 3.00  | 3.20   | 3.50  | 4.03   | 4.79   | 5.41   |
| 8      | 0.26 | 0.55 | 0.71 | 0.89 | 1.11 | 1.40 | 1.59  | 1.86 | 2.31  | 2.63  | 2.90  | 3.09   | 3.36  | 3.83   | 4.50   | 5.04   |
| 9      | 0.26 | 0.54 | 0.70 | 0.88 | 1.10 | 1.38 | 1.57  | 1.83 | 2.26  | 2.57  | 2.82  | 3.00   | 3.25  | 3.69   | 4.30   | 4.78   |
| 10     | 0.26 | 0.54 | 0.70 | 0.88 | 1.09 | 1.37 | 1.56  | 1.81 | 2.23  | 2.53  | 2.76  | 2.93   | 3.17  | 3.58   | 4.14   | 4.59   |
| 11     | 0.26 | 0.54 | 0.70 | 0.88 | 1.09 | 1.36 | 1.55  | 1.80 | 2.20  | 2.49  | 2.72  | 2.88   | 3.11  | 3.50   | 4.02   | 4.44   |
| 12     | 0.26 | 0.54 | 0.70 | 0.87 | 1.08 | 1.36 | 1.54  | 1.78 | 2.18  | 2.46  | 2.68  | 2.84   | 3.05  | 3.43   | 3.93   | 4.32   |
| 13     | 0.26 | 0.54 | 0.69 | 0.87 | 1.08 | 1.35 | 1.53  | 1.77 | 2.16  | 2.44  | 2.65  | 2.80   | 3.01  | 3.37   | 3.85   | 4.22   |
| 14     | 0.26 | 0.54 | 0.69 | 0.87 | 1.08 | 1.35 | 1.52  | 1.76 | 2.14  | 2.41  | 2.62  | 2.77   | 2.98  | 3.33   | 3.79   | 4.14   |
| 15     | 0.26 | 0.54 | 0.69 | 0.87 | 1.07 | 1.34 | 1.52  | 1.75 | 2.13  | 2.40  | 2.60  | 2.75   | 2.95  | 3.29   | 3.73   | 4.07   |
| 16     | 0.26 | 0.54 | 0.69 | 0.86 | 1.07 | 1.34 | 1.51  | 1.75 | 2.12  | 2.38  | 2.58  | 2.72   | 2.92  | 3.25   | 3.69   | 4.01   |
| 17     | 0.26 | 0.53 | 0.69 | 0.86 | 1.07 | 1.33 | 1.51  | 1.74 | 2.11  | 2.37  | 2.57  | 2.71   | 2.90  | 3.22   | 3.65   | 3.97   |
| 18     | 0.26 | 0.53 | 0.69 | 0.86 | 1.07 | 1.33 | 1.50  | 1.73 | 2.10  | 2.36  | 2.55  | 2.69   | 2.88  | 3.20   | 3.61   | 3.92   |
| 19     | 0.26 | 0.53 | 0.69 | 0.86 | 1.07 | 1.33 | 1.50  | 1.73 | 2.09  | 2.35  | 2.54  | 2.67   | 2.86  | 3.17   | 3.58   | 3.88   |
| 20     | 0.26 | 0.53 | 0.69 | 0.86 | 1.06 | 1.33 | 1.50  | 1.72 | 2.09  | 2.34  | 2.53  | 2.66   | 2.85  | 3.15   | 3.55   | 3.85   |
| 21     | 0.26 | 0.53 | 0.69 | 0.86 | 1.06 | 1.32 | 1.49  | 1.72 | 2.08  | 2.33  | 2.52  | 2.65   | 2.83  | 3.14   | 3.53   | 3.82   |
| 22     | 0.26 | 0.53 | 0.69 | 0.86 | 1.06 | 1.32 | 1.49  | 1.72 | 2.07  | 2.32  | 2.51  | 2.64   | 2.82  | 3.12   | 3.50   | 3.79   |
| 23     | 0.26 | 0.53 | 0.69 | 0.86 | 1.06 | 1.32 | 1.49  | 1.71 | 2.07  | 2.31  | 2.50  | 2.63   | 2.81  | 3.10   | 3.48   | 3.77   |
| 24     | 0.26 | 0.53 | 0.68 | 0.86 | 1.06 | 1.32 | 1.49  | 1.71 | 2.06  | 2.31  | 2.49  | 2.62   | 2.80  | 3.09   | 3.47   | 3.75   |
| 25     | 0.26 | 0.53 | 0.68 | 0.86 | 1.06 | 1.32 | 1.49  | 1.71 | 2.06  | 2.30  | 2.49  | 2.61   | 2.79  | 3.08   | 3.45   | 3.73   |
| 26     | 0.26 | 0.53 | 0.68 | 0.86 | 1.06 | 1.31 | 1.48  | 1.71 | 2.06  | 2.30  | 2.48  | 2.60   | 2.78  | 3.07   | 3.43   | 3.71   |
| 27     | 0.26 | 0.53 | 0.68 | 0.86 | 1.06 | 1.31 | 1.48  | 1.70 | 2.05  | 2.29  | 2.47  | 2.60   | 2.77  | 3.06   | 3.42   | 3.69   |
| 28     | 0.26 | 0.53 | 0.68 | 0.85 | 1.06 | 1.31 | 1.48  | 1.70 | 2.05  | 2.29  | 2.47  | 2.59   | 2.76  | 3.05   | 3.41   | 3.67   |
| 29     | 0.26 | 0.53 | 0.68 | 0.85 | 1.06 | 1.31 | 1.48  | 1.70 | 2.05  | 2.28  | 2.46  | 2.59   | 2.76  | 3.04   | 3.40   | 3.66   |
| 30     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.48  | 1.70 | 2.04  | 2.28  | 2.46  | 2.58   | 2.75  | 3.03   | 3.39   | 3.65   |
| 31     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.48  | 1.70 | 2.04  | 2.27  | 2.45  | 2.58   | 2.74  | 3.02   | 3.37   | 3.63   |
| 32     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.47  | 1.69 | 2.04  | 2.27  | 2.45  | 2.57   | 2.74  | 3.01   | 3.37   | 3.62   |
| 33     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.47  | 1.69 | 2.03  | 2.27  | 2.44  | 2.57   | 2.73  | 3.01   | 3.36   | 3.61   |
| 34     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.47  | 1.69 | 2.03  | 2.27  | 2.44  | 2.56   | 2.73  | 3.00   | 3.35   | 3.60   |
| 35     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.47  | 1.69 | 2.03  | 2.26  | 2.44  | 2.56   | 2.72  | 3.00   | 3.34   | 3.59   |
| 36     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.31 | 1.47  | 1.69 | 2.03  | 2.26  | 2.43  | 2.55   | 2.72  | 2.99   | 3.33   | 3.58   |
| 37     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.47  | 1.69 | 2.03  | 2.26  | 2.43  | 2.55   | 2.72  | 2.99   | 3.33   | 3.57   |
| 38     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.47  | 1.69 | 2.02  | 2.25  | 2.43  | 2.55   | 2.71  | 2.98   | 3.32   | 3.57   |
| 39     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.47  | 1.68 | 2.02  | 2.25  | 2.43  | 2.54   | 2.71  | 2.98   | 3.31   | 3.56   |
| 40     | 0.26 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.47  | 1.68 | 2.02  | 2.25  | 2.42  | 2.54   | 2.70  | 2.97   | 3.31   | 3.55   |
| 45     | 0.25 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.46  | 1.68 | 2.01  | 2.24  | 2.41  | 2.53   | 2.69  | 2.95   | 3.28   | 3.52   |
| 50     | 0.25 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.46  | 1.68 | 2.01  | 2.23  | 2.40  | 2.52   | 2.68  | 2.94   | 3.26   | 3.50   |
| 55     | 0.25 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.46  | 1.67 | 2.00  | 2.23  | 2.40  | 2.51   | 2.67  | 2.92   | 3.25   | 3.48   |
| 60     | 0.25 | 0.53 | 0.68 | 0.85 | 1.05 | 1.30 | 1.46  | 1.67 | 2.00  | 2.22  | 2.39  | 2.50   | 2.66  | 2.91   | 3.23   | 3.46   |
| 65     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.46  | 1.67 | 2.00  | 2.22  | 2.39  | 2.50   | 2.65  | 2.91   | 3.22   | 3.45   |
| 70     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.46  | 1.67 | 1.99  | 2.22  | 2.38  | 2.49   | 2.65  | 2.90   | 3.21   | 3.44   |
| 75     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.45  | 1.67 | 1.99  | 2.21  | 2.38  | 2.49   | 2.64  | 2.89   | 3.20   | 3.43   |
| 80     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.45  | 1.66 | 1.99  | 2.21  | 2.37  | 2.49   | 2.64  | 2.89   | 3.20   | 3.42   |
| 85     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.45  | 1.66 | 1.99  | 2.21  | 2.37  | 2.48   | 2.63  | 2.88   | 3.19   | 3.41   |
| 90     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.45  | 1.66 | 1.99  | 2.21  | 2.37  | 2.48   | 2.63  | 2.88   | 3.18   | 3.40   |
| 95     | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.45  | 1.66 | 1.99  | 2.20  | 2.37  | 2.48   | 2.63  | 2.87   | 3.18   | 3.40   |
| 100    | 0.25 | 0.53 | 0.68 | 0.85 | 1.04 | 1.29 | 1.45  | 1.66 | 1.98  | 2.20  | 2.36  | 2.48   | 2.63  | 2.87   | 3.17   | 3.39   |
| 110    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.66 | 1.98  | 2.20  | 2.36  | 2.47   | 2.62  | 2.86   | 3.17   | 3.38   |
| 120    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.66 | 1.98  | 2.20  | 2.36  | 2.47   | 2.62  | 2.86   | 3.16   | 3.37   |
| 130    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.66 | 1.98  | 2.19  | 2.36  | 2.47   | 2.61  | 2.86   | 3.15   | 3.37   |
| 140    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.66 | 1.98  | 2.19  | 2.35  | 2.46   | 2.61  | 2.85   | 3.15   | 3.36   |
| 150    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.66 | 1.98  | 2.19  | 2.35  | 2.46   | 2.61  | 2.85   | 3.15   | 3.36   |
| 160    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.65 | 1.97  | 2.19  | 2.35  | 2.46   | 2.61  | 2.85   | 3.14   | 3.35   |
| 170    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.65 | 1.97  | 2.19  | 2.35  | 2.46   | 2.61  | 2.84   | 3.14   | 3.35   |
| 180    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.65 | 1.97  | 2.19  | 2.35  | 2.46   | 2.60  | 2.84   | 3.14   | 3.35   |
| 190    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.65 | 1.97  | 2.19  | 2.35  | 2.45   | 2.60  | 2.84   | 3.13   | 3.34   |
| 200    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.29 | 1.45  | 1.65 | 1.97  | 2.19  | 2.35  | 2.45   | 2.60  | 2.84   | 3.13   | 3.34   |
| 250    | 0.25 | 0.53 | 0.68 | 0.84 | 1.04 | 1.28 | 1.44  | 1.65 | 1.97  | 2.18  | 2.34  | 2.45   | 2.60  | 2.83   | 3.12   | 3.33   |
| 300    | 0.25 | 0.52 | 0.68 | 0.84 | 1.04 | 1.28 | 1.44  | 1.65 | 1.97  | 2.18  | 2.34  | 2.45   | 2.59  | 2.83   | 3.12   | 3.32   |
| 350    | 0.25 | 0.52 | 0.68 | 0.84 | 1.04 | 1.28 | 1.44  | 1.65 | 1.97  | 2.18  | 2.34  | 2.44   | 2.59  | 2.82   | 3.11   | 3.32   |
| 400    | 0.25 | 0.52 | 0.68 | 0.84 | 1.04 | 1.28 | 1.44  | 1.65 | 1.97  | 2.18  | 2.34  | 2.44   | 2.59  | 2.82   | 3.11   | 3.32   |
| 500    | 0.25 | 0.52 | 0.67 | 0.84 | 1.04 | 1.28 | 1.44  | 1.65 | 1.96  | 2.18  | 2.33  | 2.44   | 2.59  | 2.82   | 3.11   | 3.31   |
| N(0,1) | 0.25 | 0.52 | 0.67 | 0.84 | 1.04 | 1.28 | 1.44  | 1.65 | 1.96  | 2.17  | 2.33  | 2.43   | 2.58  | 2.81   | 3.09   | 3.29   |