Tilastotieteen perusteita lääketieteeseen suuntautuville opiskelijoille

Opettajan versio

Johdanto

Tilastotiede on *menetelmätiede*, jonka antamia työkaluja voidaan soveltaa hyvin erityyppisiin ongelmiin. Tilastolliset menetelmät auttavat tutkimusaineiston analysoinnissa ja järjestelyssä sekä mahdollistavat tilastollisen päättelyn, jonka tulokset voivat antaa hyödyllistä tietoa tutkittavan joukon (tai mahdollisesti suuremman joukon) ominaisuuksista.

Tässä esityksessä käsiteltävä tilastollinen tutkimus on luonteeltaan *kvantitatiivista* eli numeerista. Tästä huolimatta myös kvalitatiivisia (laadullisia) muuttujia voidaan analysoida, kunhan ne ovat "numerisoituvia" eli lasketaan tietyntyyppisten vastausten lukumäärät tai vastaavat yhteen ja tehdään niistä päätelmiä. *Muuttujat* esitellään luvussa 1 yhdessä mitta-asteikoiden ja mittausvirheiden analyysin kanssa. Luku 2 sisältää *tilastolliset tunnusluvut* (moodi, mediaani, keskiarvo) sekä *hajontaluvut* (keskihajonta, kvartiilit). Kolmannessa luvussa keskitytään tilastotieteen yleisimpien *jakaumien* (normaalijakauma, binomijakauma, Poissonin jakauma) ominaisuuksiin.

Kolmen ensimmäisen luvun jälkeen on hyvät lähtökohdat aloittaa varsinainen tilastollinen analyysi. Asiaan on paras tarttua kiinni esimerkkien kautta, joten luvussa 4 katsomme erilaisia *tilastollisen tutkimusdatan esitysmuotoja* (taulukot, graafit, pylväsdiagrammit ja histogrammit). Luvun lopussa kerromme, kuinka ei-lineaarisen tilastollisen aineiston voi matemaattisella muunnoksella muuttaa lineaariseksi[1]. Data-analyysin jälkeen pääsemme viimeiseen "teorialukuun", jossa keskitymme kahden muuttujan välisen riippuvuussuhteen kuvailuun. Kaksi tavallista ja toisiinsa voimakkaasti liittyvää menetelmää ovat *korrelaatiokertoimen laskeminen* ja *regressiosuoran* (PNS-suoran) sovittaminen dataan. Luvussa 6 odottavat laskuharjoitustehtävät ja liitteessä 8 on esitetty monisteen asioihin liittyviä oleellisia täydennyksiä.

Jokaisesta käsiteltävästä teoria-asiasta, johon liittyy merkittävästi matematiikkaa, on tässä monisteessa vähintään yksi esimerkki. Jos materiaali ei Galenoksen ohella vaikuta tarpeeksi kattavalta, niin voin suositella ensimmäiseksi lisämateriaaliksi teosta [1] ja varsinaisille "tilastohamstraajille" kirjaa [3]. Myös teoksessa [2] on hyviä käytännönläheisiä esimerkkejä ja ymmärrettävää teoriaa tilastotieteen perusasioista. Hieman haastavampi (mutta selkeä) esitys löytyy lähteestä [4] (erityisesti binomijakauma ja Poissonin jakauma).

> **Huomautus.** Tilastotieteen perusasiat ovat laskennallisesti melko yksinkertaisia. Tutkimuksen näkökulmasta huomattavasti laskuja oleellisempaa on ymmärtää analyysimenetelmien käyttömahdollisuudet ja muuttujien, mitta-asteikoiden, tunnuslukujen sekä jakaumien liittyminen toisiinsa. Laskennallisen puolen tulee tällöin olla hallussa ilman, että siihen tarvitsee juurikaan keskittyä. Tällaisen varmuuden saa ainoastaan laskemalla riittävän määrän tehtäviä eli ei kannata skipata laskuja, vaikka ne olisivat äärimmäisen helppoja.

[1] Tässä kohdassa oletetaan, että paraabelifunktioon ja logaritmifunktioon liittyvät matematiikka on hyvin hallussa.

Sisältö

1 Mittaaminen ja muuttujat

Tilastollisen tutkimuksen kohteena on *perusjoukko* (populaatio), josta valitaan *satunnaisotannalla* haluttu määrä *tilastoyksiköitä* (havaintoyksiköitä) analyysiin. Otantamenetelmiä on erilaisia ja niistä on lyhyt esitys luvussa 8.1. Tilastotutkimuksessa laskutoimitukset tehdään otannalle, jonka vuoksi aineistosta laskettuja tilastollisia suureita kutsutaan *otantasuureiksi* (esimerkiksi otoskeskiarvo ja otoskeskihajonta). Otantasuureista lasketuilla luvuilla pyritään kuvaamaan koko perusjoukon ominaisuuksia. Otannasta tehtäviä päätelmiä perusjoukolle ei käsitellä tässä monisteessa, vaan tyydytään yksittäiseen otantaan liittyvään analyysiin[2].

1.1 Mitta-asteikot

Tilastollisia muuttujia voivat olla periaatteessa mitkä tahansa. Muuttujat jaetaan kahteen pääluokkaan eli *epäjatkuviin* (diskreetteihin) ja *jatkuviin* muuttujiin. Diskreetti muuttuja voi saada vain toisistaan erillisiä arvoja (esimerkiksi kouluarvosana) ja jatkuva muuttuja teoriassa mitä tahansa arvoja (esim. 400 metrin juoksuaika). Jatkuvan muuttujan arvon tarkkuuden määrittää yleensä mittalaitteen tarkkuus. Muuttujan luonne määrää muuttujan mitta-asteikon:

Luokittelu- eli nominaaliasteikko on muuttujalla, jolla ei ole numeerista arvoa vaan pelkkä luokan tunnus (esimerkiksi sukupuoli). Tässä tilastoanalyysin näkökulmasta yksinkertaisimmassa asteikossa luokkien välille ei voida määritellä loogista järjestystä, vaan luokat ovat keskenään samanarvoisia. Mittalukuja ei voi laskea yhteen tai vähentää toisistaan, sillä niillä ei ole määrällistä tulkintaa (ts. ei ole olemassa lukua n siten, että se kattavasti sisältäisi kaiken saman informaation kuin esimerkiksi sana ”sukupuoli”).

Järjestys- eli ordinaaliasteikolla suoritettu mittaus kertoo mitatun arvon järjestyksen (suuruus, paremmuus) suhteessa toiseen mitattuun arvoon. Mittaluvuilla ei kuitenkaan ole määritelty mittayksikköä, eli ei voida tarkasti sanoa, kuinka paljon suurempi loogisessa järjestyksessä korkeammalla oleva arvo on kuin toinen vertailuarvo. Aritmeettisia laskutoimituksia luokkien välillä ei voida suorittaa. Esimerkkinä järjestysasteikkoisesta muuttujasta voidaan mainita mielipidekysely, jossa on 5-portainen asteikko: 1 = täysin samaa mieltä, 5 = täysin eri mieltä ja muut vaihtoehdot siinä välissä.

Välimatka- eli intervalliasteikolla mitattuja muuttujan arvoja voidaan verrata toisiinsa ja mittauksen tulos kertoo kohteen eron suuruuden toiseen kohteeseen. Mittayksikkö on usein olemassa ja yhteen- sekä vähennyslaskuja muuttujan arvojen välillä voidaan tehdä. Esimerkkejä välimatka-asteikkoisista muuttujista ovat ihmisen pituus, paino ja ikä.

[2] Otannan perusteella tehdyt päätelmät saadaan tilastollisella merkitsevyystestauksella, josta voi lueskella lisää teoksesta [1], luku 8.

Välimatka-asteikon erikoistapaus on **suhdeasteikko**. Suhdeasteikkoisella muuttujalla on erityinen ”absoluuttinen” nollapiste, jossa ominaisuus häviää. Esimerkkejä suhdeasteikkoisista muuttujista ovat kaikki edellä mainitut ihmisen muuttujat. Kannattaa kuitenkin huomata, että esimerkiksi lämpötilan mittaamisessa käytettävä celciusasteikko ei ole suhdeasteikko, sillä esimerkiksi 30 astetta ei ole ”kaksi kertaa niin lämmin” kuin 15 astetta.

Luokittelu- ja järjestysasteikolliset muuttujat ovat aina epäjatkuvia, kun taas välimatka-asteikollinen muuttuja voi olla jatkuva tai epäjatkuva. Joskus jatkuva muuttuja tulkitaan epäjatkuvana siitä syystä, että saadaan käytännöllisempiä tuloksia (esimerkiksi ihmisen ikä kokonaisina vuosina). Tälläinen muunnos voidaan tehdä sopivalla luokittelulla ja pyöristämisellä.

1.2 Mittaamisen tarkkuus

Mittaustulokseen liittyy aina mittausvirhe, joka voi olla seurausta mittalaitteen tarkkuudesta, käyttäjän osaamattomuudesta tai vaikkapa maanjäristyksestä, joka häiritse mittausta. Mittausvirheet jaetaan kahteen luokkaan:

Satunnaisvirhe on nimensä mukaan sattumaan liittyvä eli sen suuntaa ei voida ennustaa. Tähän kategoriaan kuuluvat mm. huolimattomuusvirheet ja mittalaitteen aiheuttamat satunnaisvirheet (esimerkiksi mittari heiluu molemmin puolin ”oikeaa” arvoa noudattamatta mitään sääntöä).

Systemaattinen virhe on aina johonkin suuntaan suhteessa oikeaan arvoon. Esimerkkejä ovat mm. aina liian paljon näyttävä lämpömittari ja systemaattinen mitta-asteikon lukeminen väärästä kohdasta (tai vaikkapa tuuma-asteikon lukeminen, vaikka tarkoituksena olisi lukea senttimetrejä).

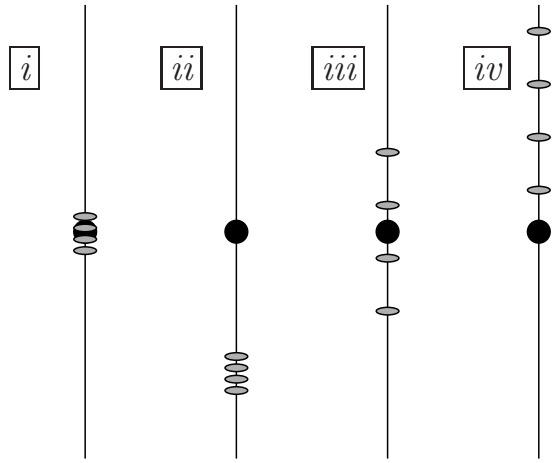
Mittaustulos ilmoitetaan tieteellisissä julkaisuissa usein muodossa

$$\boxed{\text{mittaustulos} = \text{havaittu arvo} \pm \text{systemaattinen virhe} \pm \text{satunnaisvirhe}}$$

Tarkkuuden yhteydessä puhutaan usein mittarin *reliabiliteetista* (sisäinen tarkkuus) ja *validiteetista* (ulkoinen tarkkuus). Reliabiliteetti tarkoittaa mittauksen toistettavuuden luotettavuutta satunnaisvirheen mielessä; hyvä reliabiliteetti on mittarilla, joka antaa useita kertoja samoissa olosuhteissa toistettaessa suunnilleen saman tuloksen (vaikka tulokset poikkeaisivatkin huomattavasti ”oikeasta” arvosta). Hyvä validiteetti on puolestaan mittarilla, joka antaa keskimäärin oikean tuloksen usean mittauksen sarjalle, vaikka yksittäiset mittausarvot poikkeaisivat paljonkin toisistaan (ja mahdollisesti myös oikeasta arvosta). Graafinen esitys reliabiliteetista ja validiteetista on kuvassa 1.

Esimerkki 1.1: Reliabiliteetti ja validiteetti
Kerro, liittyvätkö seuraavien mittaustilanteiden ongelmat reliabiliteettiin vai validiteettiin.

a) Lääketieteen pääsykokeen tarkastaja unohtaa antaa pisteet fysiikan tehtävän viimeisestä kohdasta kaikille kokeeseen vastanneille.



Kuva 1: Mittarin reliabiliteetti ja validiteetti graafisesti. Kohdassa $\boxed{i}$ on hyvä validiteetti ja reliabiliteetti, ts. mittaus on toistettavissa ja antaa lähes oikeita arvoja. Kohdassa $\boxed{ii}$ reliabiliteetti on edelleen hyvä, mutta validiteetti heikko. Kolmannessa kohdassa $\boxed{iii}$ validiteetti on hyvä (mittaus antaa keskimäärin oikean tuloksen) ja reliabiliteetti heikko. Viimeisessä kohdassa $\boxed{iv}$ sekä reliabiliteetti että validiteetti ovat heikkoja.

b) Sentrifugoinnin tuloksena saadaan koeputkien pohjalle samanlaisia soluelimiä, jotka eivät ole tutkimuksen kannalta merkittäviä.

c) Kokenut lääkäri diagnosoi potilaalle viruspohjaisen taudin, vaikka taudin aiheuttaja itse asiassa on bakteeri.

d) Röntgenkuvauksessa havaitaan 200:lle potilaalle syöpä, vaikka tietokonetomografiakuvan perusteella 95 % potilaista todetaan terveiksi eikä heillä ole syövän oireita.

Vastaukset

a) Tarkastaja korjaa kaikki vastauspaperit samalla tavalla väärin, joten virhe on systemaattinen ja ongelma validiteetissa.

b) Sentrifugi toimii oikein siinä mielessä, että samanlaiset soluelimet saadaan kerättyä. Sentrifugin säätäminen on tehty väärin, koska tavoitteena oli kerätä eri tyypin soluelimiä. Sentrifugi kerää systemaattisesti vääriä havaintoyksiköitä, joten ongelma liittyy validiteettiin. Sentrifugin säätämisessä on tosin voinut tapahtua satunnaisvirhe.

c) Kyse on yksittäisestä potilaasta ja lääkärin tekemästä virhediagnoosista, joka on satunnaisvirhe. Reliabiliteettiongelma.

d) Röntgenkuvauslaitteisto näyttää tutkituille potilaille saman tuloksen potilaasta riippumatta, eli virhe on systemaattinen. Validiteettiongelma.

2 Tilastolliset tunnusluvut ja hajontaluvut

Tutkittavan aineiston ominaisuuksia voidaan kuvata typistetysti *tunnuslukujen* ja *hajontalukujen* avulla. Ne sisältävät jakauman sijaintia (keskikohtaa) ja muotoa (hajontaa) koskevaa informaatiota.

2.1 Tunnusluvut

Jakauman sijaintia kuvaavia tunnuslukuja ovat *moodi*, *mediaani* ja *keskiarvo* ja niillä on seuraavat ominaisuudet:

Moodi M_o eli tyyppiarvo kertoo, mitä muuttujan arvoa on havainnoissa eniten. Jos esimerkiksi lääketieteelliseen tutkimukseen osallistuu 10 miestä ja 8 naista, niin muuttujan "sukupuoli" moodi on "mies". Moodeja voi olla useita, jos useammalla luokalla on sama havaintoarvojen lukumäärä eli *frekvenssi*, joka samalla on maksimi. Moodia voi käyttää jo luokitteluasteikkoiselle muuttujalle ja sitä ei kannata käyttää jatkuvalle muuttujalle (muut keskiluvut sopivat paremmin).

Mediaani M_d on sen luokan muuttujan arvo, jota pienempiä ja suurempia havaintoarvoja on yhtä paljon (eli mediaaniluvun molemmilla puolilla on yhtä suuri määrä lukuja). Jos lukuja on parillinen määrä, niin mediaani on kahden keskimmäisen luvun keskiarvo, kun muuttuja on vähintään välimatka-asteikollinen. Mikäli muuttuja on järjestysasteikollinen, niin mediaaneja on kaksi (kaksi keskimmäistä lukua). Mediaania saa käyttää vähintään järjestysasteikolliselle muuttujalle.

Keskiarvo $\overline{x}$ kuvaa havaintojen keskimääräistä arvoa. Se vaatii tarkasteltavalta muuttujalta vähintään välimatka-asteikollisuutta. Keskiarvo saadaan laskemalla kaikki muuttujan havaintoarvot yhteen ja jakamalla tulos arvojen määrällä. Mikäli havaintoja on n kappaletta, saadaan

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Tunnuslukujen lisäksi määritellään tilastollisen muuttujan X odotusarvo μ yhtälöllä

$$\mu = \sum_{i=1}^{n} p_i x_i, \tag{2}$$

missä p_i on havaintoarvoa x_i vastaava todennäköisyys ($\sum_i pi = 1$). Jos jokaisella arvolla on sama todennäköisyys, niin odotusarvosta tulee keskiarvo (otoskoon n tapauksessa jokaisen arvon todennäköisyys on $1/n$; sijoittamalla tämä odotusarvon kaavaan saadaan keskiarvon kaava).

Esimerkki 2.1: Tunnuslukujen käyttöä
Radioaktiivisen merkkiaineen aktiivisuudelle tietyn ajan jälkeen on mitattu seuraavat arvot:

$$185 \text{ Bq},\ 162 \text{ Bq},\ 197 \text{ Bq},\ 143 \text{ Bq},\ 153 \text{ Bq} \quad \text{ja} \quad 183 \text{ Bq},$$

Määritä aineiston mediaani ja keskiarvo.

Vastaus Voidaan päätellä, että muuttuja on välimatka-asteikollinen, joten mediaani saadaan laskemalla kahden keskimmäisen mittaluvun keskiarvo (lukuja on parillinen määrä). Tulos on

$$M_d = \frac{162 \text{ Bq} + 183 \text{ Bq}}{2} = 172,5 \text{ Bq}$$

Keskiarvo $\overline{x}$ saadaan tavalliseen tapaan:

$$\overline{x} = \frac{1}{6} \cdot (185 + 162 + 197 + 143 + 153 + 183) \text{ Bq} = 170,5 \text{ Bq}$$

2.2 Hajontaluvut

Tunnuslukujen tarkoituksena on kuvata paikkaa, johon jakauman keskikohta on lokalisoitunut. Hajontaluku puolestaan mittaa jakauman muotoa eli sitä, kuinka laajalle alueelle keskikohdan ympärille havaintoarvot ovat jollakin todennäköisyydellä jakautuneet. Hajontalukuja ovat *vaihteluväli* (ja sen pituus), *varianssi*, *keskihajonta*, *kovarianssi* sekä *keskiarvon keskivirhe*. Seuraavassa lyhyesti hajontalukujen ominaisuuksista:

Vaihteluväli R tarkoittaa muuttujan pienimmän ja suurimman havaitun arvon väliä eli väliä $[min, max]$. Vaihteluväliä voidaan käyttää vähintään järjestysasteikolliselle muuttujalle, vaihteluvälin pituutta vähintään välimatka-asteikolliselle.

Varianssi s^2_{n-1} kannattaa ajatella seuraavasti: haluamme luvun, joka toimisi luonnollisena, "keskimääräisenä" hajontalukuna. Ongelmaksi muodostuu se, että laskemalla yhteen kaikki poikkeamat $x_i - \overline{x}$ (x_i on muuttujan X yksittäinen havaintoarvo ja $\overline{x}$ on muuttujan arvojen keskiarvo) saamme tulokseksi nollan. Tästä syystä lasketaan neliöllisten poikkeamien $(x_i - \overline{x})^2$ summa (sillä reaaliluvun neliö on aina ei-negatiivinen) ja neliöpoikkeamien keskiarvo pienellä muutoksella, eli jaetaan otoskoon n sijaan luvulla[3] $n-1$:

$$s^2_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{3}$$

Huomaa, että varianssilla on samat yksiköt kuin keskiarvon neliöllä. Varianssi (kuten myös seuraavaksi esiteltävä keskihajonta) sopivat vähintään välimatka-asteikolliselle muuttujalle. Mikäli halutaan korostaa, että varianssi liittyy nimenomaan tilastolliseen muuttujaan X, niin sitä voidaan merkitä s^2_x.

[3] Tämä on perusteltua, mutta perustelu vaatii huomattavasti syvempää perehtymistä tilastotieteen matemaattisiin oletuksiin kuin tällä kurssilla tehdään.

Keskihajonta s_{n-1} saadaan varianssin neliöjuurena:

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4}$$

Keskihajonta kuvaa jakauman leveyttä siten, että esimerkiksi normaalijakauman[4] tilanteessa yhden keskihajonnan päässä keskiarvosta (odotusarvosta) on yhteensä noin 68 % kaikista havaintoarvoista (kahden keskihajonnan päässä noin 95 %). Keskihajonnalla on samat yksiköt kuin keskiarvolla. Keskihajontaa voidaan merkitä myös s_x.

Kovarianssi s_{xy} kuvaa kahden tilastollisen muuttujan X ja Y "yhteisvaihtelua" ja se saadaan yhtälöstä

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{5}$$

Kovarianssia tarvitaan erityisesti korrelaatiokertoimen määrittelyssä luvussa 5.1.

Keskiarvon keskivirhe $s_{\overline{x}}$ on keskiarvon arvioitu poikkeama perusjoukon "todellisesta" odotusarvosta. Se saadaan yhtälöstä

$$s_{\overline{x}} = \frac{s_{n-1}}{\sqrt{n}} \tag{6}$$

Helpoiten kaavan muistaa todennäköisesti siitä, että keskiarvon varianssi eli keskimääräinen neliöpoikkema on s_{n-1}^2/n ja keskiarvon keskivirhe on itse asiassa keskiarvon keskihajonta eli tämän neliöjuuri (keskiarvon keskivirhe kuvaa perusjoukosta tehtyjen otosten keskiarvojen jakauman hajanaisuutta).

Esimerkki 2.2: Hajontalukujen käyttöä

Tarkastellaan taulukkoa 1, jossa on viiden henkilön kokonaispistemäärä lääketieteellisen tiedekunnan pääsykokeissa. Laske pistemäärien keskiarvo, keskiarvon keskivirhe ja keskihajonta. Kuinka moni taulukon henkilöistä pääsi sisään, jos tarvittava pistemäärä oli 120 % keskiarvosta?

Vastaus Aloitetaan laskemalla havaintojen keskiarvo:

$$\overline{x} = \frac{1}{5}\cdot(58 + 124 + 83 + 38 + 135) = 87,6$$

Mikäli sisäänpääsyraja oli 120% keskiarvosta, niin saamme rajaksi $R = 1,2 \cdot 87,6 = 105,12$. Näin ollen henkilöt 2 ja 5 (kaksi henkilöä) pääsivät sisään tiedekuntaan ja loput jäivät ulkopuolelle.

Keskihajontalaskuissa kannattaa laskea erikseen poikkeamat $x_i - \overline{x}$ sekä niiden neliöt ja taulukoida tulokset järkevästi. Tällä tavalla on helppo jatkaa analyysiä myöhemmin ja laskea esimerkiksi tarvittava korrelaatiokerroin tai regressiosuora, kuten luvussa 5 näemme. Laskeskellaan hetki laskimella ja kirjoitetaan taulukko 2.

[4] Ks. luku 3.4.

Taulukko 1: Lääketieteen pääsykokeisiin osallistuneiden henkilöiden pistemääriä valintakokeessa.

Henkilö	Pistemäärä
1	58
2	124
3	83
4	38
5	135

Taulukko 2: Lääketieteen pääsykokeen pistemäärien poikkeamat.

ID	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	58	-29,6	876,16
2	124	36,4	1324,96
3	83	-4,6	21,16
4	38	-49,6	2460,16
5	135	47,6	2265,76

Sijoittamalla poikkeamien neliöt yhtälöön (3) saamme varianssiksi

$$\begin{aligned} s_{n-1}^2 &= \frac{1}{4} \cdot (876,16 + 1324,96 + 21,16 + 2460,16 + 2265,76) \\ &= 1737,05 \end{aligned}$$

Keskihajonta saadaan varianssin neliöjuurena eli $s_{n-1} = \sqrt{1737,05} = 41,667\ldots \approx 41,7$. Keskiarvon keskivirhe lasketaan jakamalla keskihajonta otoskoon neliöjuurella:

$$s_{\overline{x}} = \frac{s_{n-1}}{\sqrt{n}} = \frac{\sqrt{1737,05}}{\sqrt{5}} = 18,638\ldots \approx 18,6$$

Lopputulokset on mielekästä ilmoittaa yhden desimaalin tarkkuudella:

$$\begin{cases} \overline{x} & = 87,6 \\ s_{n-1} & = 41,7 \\ s_{\overline{x}} & = 18,6 \end{cases}$$

3 Tilastolliset jakaumat

Havaintoaineiston jakauman muoto ja tyyppi määräävät muuttujien mitta-asteikoiden ohella hyvin pitkälti sen, millaisia tilastollisia analyysimenetelmiä aineistolle voidaan käyttää. Käydään lyhyesti läpi binomijakauma, Poissonin jakauma sekä normaalijakauma. Myös muita tilastollisia jakaumia on olemassa (mm. t-jakauma, F-jakauma ja χ^2-jakauma), mutta niiden käsittely sivuutetaan.

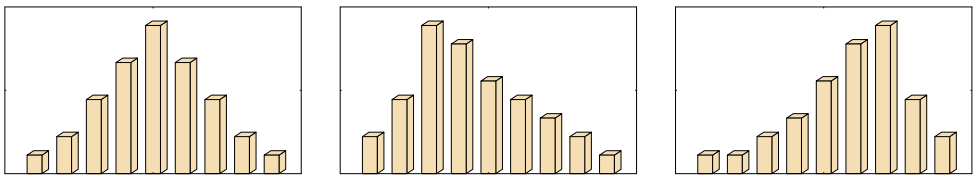
3.1 Jakauman muodosta

Jakauman muoto vaikuttaa siihen, kuinka aineistosta lasketut tunnusluvut suhtautuvat toisiinsa. Muodolle on olemassa kolme perustyyppiä (kuva 2):

Symmetrisellä jakaumalla on keskiarvon molemmin puolin suunnilleen yhtä paljon havaintoja. Aineistosta laskettu mediaani ja keskiarvo ovat lähes yhtä suuria.

Oikealle vinon jakauman häntä on oikealla puolella pidempi kuin vasemmalla puolella. Tästä syystä suurin osa arvoista on keskikohdan vasemmalla puolella, ja aineistosta laskettu mediaani on pienempi kuin vastaava keskiarvo, ts. $M_d < \overline{x}$.

Vasemmalle vinon jakauman vasen häntä on pidempi kuin oikea ja tästä syystä mediaani on suurempi kuin keskiarvo, $M_d > \overline{x}$.



Kuva 2: Symmetrinen, oikealle ja vasemmalle vino jakauma

3.2 Binomijakauma

Tarkastellaan binomijakaumaa esimerkin kautta.

Esimerkki 3.1: Johdanto binomijakaumaan
Terveyskeskuksessa on käynyt sikainfluenssarokotuksessa 3 henkilöä. Rokotuksen saaneen henkilön todennäköisyys sairastua sikainfluenssaan on arviolta 4 prosenttia. Laske todennäköisyys sille, että

a) kukaan ei sairastu

b) yksi rokotuksen saaneista sairastuu ?

c) Muodosta lopuksi satunnaismuuttujan $X =$ ”sairastuneiden lukumäärä” todennäköisyysjakauma (eli millä todennäköisyydellä 1 sairastuu sikainfluenssaan, 2 sairastuu jne.).

Vastaukset Kolmesta henkilöstä voi sairastua 0, 1, 2 tai 3, eli satunnaismuuttuja X voi saada juuri nämä arvot.

a) Ajatellaan ensin tilanne 0 eli että kukaan ei sairastu. Olkoon $p = 0{,}96$ todennäköisyys sille, että henkilö on terve, jolloin $q = 1 - p = 0,04$ on henkilön todennäköisyys sairastua. Kun kaikki ovat terveitä, niin todennäköisyys on yksittäisten todennäköisyyksien tulo eli

$$\mathcal{P}(0 \text{ sairasta}) = p^3 q^0 = 0,96^4 = 0,8847\ldots \tag{7}$$

b) Mietitään seuraavaksi tilannetta, jossa meillä on yksi sairastunut ja kolme tervettä. Nyt asiat muuttuvat siten, että yksi sairastuneista voi olla kuka tahansa kolmesta henkilöstä, ts. on kolme erilaista (toisensa poissulkevaa) mahdollisuutta sairastuneelle henkilölle. Mikäli ajatellaan, että ensimmäinen henkilö on sairas, saadaan tapausta vastaavaksi todennäköisyydeksi

$$\mathcal{P}(1. \text{ henkilö sairas, 2. ja 3. terveitä}) = q^1 p^2 = 0,04 \cdot 0,96^2 = 0,03686\ldots$$

Vastaavasti saadaan todennäköisyydet tilanteille, joissa henkilöt 2 ja 3 ovat sairaita:

$$\mathcal{P}(2. \text{ henkilö sairas, 1. ja 3. terveitä}) = pq^1 p = 0,96 \cdot 0,04 \cdot 0,96 = 0,03686\ldots$$
$$\mathcal{P}(3. \text{ henkilö sairas, 1. ja 2. terveitä}) = ppq = 0,96 \cdot 0,96 \cdot 0,04 = 0,03686\ldots$$

Havaitaan, että kaikista tulee todennäköisyydeksi sama luku. Kun laskemme satunnaismuuttujan X arvoa 1 vastaavaa todennäköisyyttä, Meidän tulee huomioida kaikki kolme mahdollisuutta, jotta saamme tuloksen oikein:

$$\mathcal{P}(1 \text{ sairas}) = 3 \cdot 0,03686\ldots = 0,110592 \tag{8}$$

c) Jos sairastuneita henkilöitä on kaksi, huomioidaan jälleen kaikki kolme mahdollisuutta kahdelle sairastuneelle henkilölle ja lasketaan todennäköisyys:

$$\mathcal{P}(2 \text{ sairasta}) = 3 \cdot q^2 p = 3 \cdot 0,04^2 \cdot 0,96 = 0,004608 \tag{9}$$

Jos kaikki kolme henkilöä ovat sairaana, saadaan todennäköisyydeksi

$$\mathcal{P}(3 \text{ sairasta}) = q^3 = 0,04^3 = 6 \cdot 10^{-5} \tag{10}$$

Taulukossa 3 on satunnaismuuttujan X arvot ja niitä vastaavat todennäköisyydet eli *todennäköisyysjakauma*.

Taulukko 3: Satunnaismuuttujan X = "sairastuneiden lukumäärä" todennäköisyysjakauma. Todennäköisyydet p_i on pyöristetty tuhannesosan tarkkuuteen.

x_i	p_i
0	0,884
1	0,111
2	0,005
3	0,000

3.2.1 Binomijakauma teoreettisesti

Binomijakauma liittyy satunnaiskokeeseen, jolta vaaditaan seuraavat ominaisuudet:

- Kokeen lopputulos voidaan määrittää siten, että sillä on olemassa kaksi toisensa poissulkevaa vaihtoehtoa: sairastuu/ei sairastu, onnistuu/ei onnistu tai jotakin vastaava. Kokeen onnistumisen todennäköisyys on p ja epäonnistumisen todennäköisyys $q = 1 - p$.
- Vaihtoehtojen todennäköisyydet pysyvät vakioina, kun koetta toistetaan[5]. Kokeen toistaminen tarkoittaa havaintoyksiköiden läpikäymistä (esimerkiksi ihmiset yksi kerrallaan) ja vastaavien todennäköisyyksien huomiointia.

Binomijakaumalla pyritään selvittämään todennäköisyys sille, että saadaan x kpl haluttuja tuloksia, kun tehdään n kpl satunnaiskokeita (esimerkiksi halutaan 0 kpl sairastuneita henkilöitä, kun 3 henkilöä testataan ja lasketaan tälle todennäköisyys). Kysymys kuuluu: kuinka monella eri tavalla voidaan n:n alkion joukko jakaa kahteen osajoukkoon siten, että toisessa joukossa on x kpl alkioita ja toisessa $n-x$, kun valintajärjestyksellä on väliä? Vastaus saadaan kombinatoriikasta, sen antaa binomikerroin $\binom{n}{x}$:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x(x-1)\cdots 1 \cdot (n-x)\cdots 1} \tag{11}$$

Esimerkissä 3.1 kerroin 3 yhden ja kahden sairaan henkilön todennäköisyyksien edessä saadaan juuri binomikertoimella, sillä

$$3 = \binom{3}{1} = \frac{3!}{1!2!} = \frac{6}{2} = 3 \tag{12}$$

Analogisesti löydetään kaikki kolme mahdollisuutta valita 2 sairasta henkilöä kolmen joukosta[6].

[5] Vastaesimerkiksi sopii vaikkapa korttipakka, josta vedettyjä kortteja ei palauteta takaisin: näin ollen pakassa jäljellä olevien korttien todennäköisyydet riippuvat aiemmista tapahtumista eikä tilannetta voida analysoida binomijakaumalla.

[6] Mikäli binomikertoimesta haluaa hieman kattavamman selvityksen, suosittelen erityisesti teosta [4] ja sen lukua kombinatoriikasta.

Nyt voidaan kirjoittaa binomijakaumalle yleinen muoto. Olkoon p todennäköisyys sille, että suotuisa tapahtuma tapahtuu ja vastaavasti $q = 1 - p$ on ei-suotuisan tapahtuman todennäköisyys. Kun satunnaiskoe toistetaan n kertaa, todennäköisyys saada x suotuisaa tulosta on

$$\begin{aligned} \mathcal{P}(x \text{ suotuisaa } n\text{:stä kokeesta}) &= \mathcal{B}(n, x, p) \\ &= \binom{n}{x} p^x q^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x q^{n-x} \end{aligned} \tag{13}$$

Binomijakaumafunktio $\mathcal{B}(n, x, p)$ on todennäköisyys, joten sille pätevät kaikki todennäköisyyden tavalliset ominaisuudet:

- $0 \leq \mathcal{B}(n, x, p) \leq 1$ kaikille x.
- $\sum_{x=0}^{n} \mathcal{B}(n, x, p) = 1$

Binomijakauma on *diskreetti* todennäköisyysjakauma eli eri todennäköisyyksiä vastaavat satunnaismuuttujan arvot ovat erillisiä. Tämä vihjaa siitä, että binomijakaumalla analysoidaan diskreettejä tilastollisia muuttujia.

Binomijakauman *odotusarvo* eli todennäköisyyksillä painotettu keskiarvo saadaan yhtälöstä

$$\overline{x} = \sum_{x=0}^{n} x\mathcal{B}(n, x, p) = np \tag{14}$$

Vastaavasti voidaan kirjoittaa yhtälö binomijakauman keskihajonnalle:

$$\sigma = \sqrt{np(1-p)} \tag{15}$$

Yhtälöiden (14) ja (15) perustelut vaativat differentiaalilaskentaa lukion ulkopuolelta ja ne löytyvät esimerkiksi teoksesta [5] (luku 10).

Esimerkki 3.2: Binomijakauman sovellus
Epävirallisten tietojen mukaan erään yrityksen lääketieteen valmennuskursseilta tiedekuntaan sisäänpääsyprosentti on 34 %. Valmennuskurssille osallistuu 23 henkeä.

a) Mikä on todennäköisyys, että seitsemän heistä saa opiskelupaikan lääkiksestä?

b) Millä todennäköisyydellä korkeintaan yksi pääsee lääkikseen?

c) Laske lääketieteelliseen pääsevien opiskelijoiden määrän odotusarvo sekä keskihajonta.

Vastaukset Ratkaistaan ongelma sujuvasti binomitodennäköisyyden yleisen yhtälön avulla.

a) Onnistumisten määrä on 7, jolloin epäonnistumisia on $23 - 7 = 16$. Sijoitetaan annetut luvut binomitodennäköisyyden kaavaan, jolloin saadaan

$$\begin{aligned}\mathcal{B}(23;7;0,34) &= \frac{23!}{7!(23-7)!} \cdot 0,34^7 \cdot 0,66^{23-7} \\ &= \frac{23 \cdot 22 \cdot 21 \cdot 20 \cdot 19 \cdot 18 \cdot 17}{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} \cdot \cdot 0,34^7 \cdot 0,66^{16} \\ &= 0,16691\ldots \\ &\approx 0,17\end{aligned} \tag{16}$$

b) Todennäköisyys sille, että korkeintaan yksi pääsee lääkikseen, saadaan todennäköisyyksien $\mathcal{P}$ (yksi pääsee lääkikseen) ja $\mathcal{P}$ (kukaan ei pääse lääkikseen) summana, sillä ks. tapahtumat ovat toisensa poissulkevia (jos toinen tapahtuu, niin toinen ei voi tapahtua). Todennäköisyyksiksi saamme

$$\begin{aligned}\mathcal{P}(\text{yksi pääsee lääkikseen}) = \mathcal{B}(23;1;0,34) &= \frac{23!}{1!(22)!} \cdot 0,34 \cdot 0,66^{22} \\ &= 23 \cdot 0,34 \cdot 0,66^{22} \\ &= 0,000837\ldots \\ \mathcal{P}(\text{kukaan ei pääse lääkikseen}) = \mathcal{B}(23;0;0,34) &= \frac{23!}{0!(23)!} \cdot 0,34^0 \cdot 0,66^{23} \\ &= 0,0000707\ldots\end{aligned} \tag{17}$$

Laskemalla todennäköisyydet yhteen saamme noin $0,001 = 0,1$ %.

c) Odotusarvo saadaan kertomalla otoskoko sisäänpääsytodennäköisyydellä eli

$$\mu = 23 \cdot 0,34 = 7,82 \approx 8 \text{ henkilöä}$$

Keskihajonta saadaan yhtälöstä (15) ja vastaukseksi tulee

$$\sigma = \sqrt{23 \cdot 0,34 \cdot 0,66} = 2,2718\ldots \approx 2 \text{ henkilöä}$$

Lopputulos voidaan kirjoittaa muodossa 8 ± 2 henkilöä.

3.3 Poissonin jakauma

Poissonin jakaumaa käytetään kuvaamaan sellaisten kokeiden tuloksia, joissa tapahtumia (esimerkiksi radioaktiivisten ydinten hajoamisia) tapahtuu satunnaisin aikavälein, mutta kuitenkin tietyllä nopeudella (eli on olemassa jokin arvio esimerkiksi sille, kuinka monta ydintä hajoaa aikayksikössä, ts. puoliintumisaika).

Radioaktiivisen esimerkin avulla voidaan myös ajatella, että periaatteessa Poissonin jakauma on binomijakauma. Mikäli ytimiä on alussa n kappaletta, x niistä hajoaa annetussa ajassa ja todennäköisyys hajoamiselle on p, niin tilannetta kuvaa binomijakauma

$\mathcal{B}(n, x, p)$. Matemaatikot ovat osoittaneet, että mikäli n on hyvin suuri (atomiytimien tapauksessa voi olla luokkaa 10^{20}) ja todennäköisyys p yhden ytimen hajoamiselle on "pieni", niin binomijakaumafunktiota $\mathcal{B}(n, x, p)$ approksimoi erittäin suurella tarkkuudella funktio

$$P_\mu(x) = e^{-\mu}\frac{\mu^x}{x!}, \tag{18}$$

jota kutsutaan *Poissonin* jakaumaksi. Yhtälössä (18) parametri $\mu = \overline{x}$ on jakauman odotusarvo[7].

Poissonin jakaumalle pätevät kaikki todennäköisyyden yleiset ominaisuudet ja se on binomijakauman tapaan diskreetti jakauma. Jakauman keskihajonnalle on voimassa yhtälö

$$\sigma_x = \sqrt{\mu}, \tag{19}$$

eli jakauman levinneisyyttä odotusarvon ympärillä kuvaa odotusarvon neliöjuuri.

Esimerkki 3.3: Poissonin jakauman käyttöä
Alkuaineen radioaktiivinen isotooppi emittoi alfahiukkasia keskimäärin 1200 kpl viidessä minuutissa. Määritä keskimääräinen emittoitumistaajuus ja sen keskihajonta SI-yksiköissä.

Vastaus Viisi minuuttia vastaa SI-yksiköissä aikaa 300 sekuntia, joten alfahiukkasia emittoituu sekunnissa $1200/300 = 4$ kpl. Keskihajonta σ_x saadaan ottamalla neliöjuuri emittoitumistaajuudesta eli $\sigma_x = \sqrt{4} = 2$. Lopputulos ilmaistaa tavallisesti muodossa

$$\text{Emittoitumistaajuus } = R = 4 \pm 2 \text{ hiukkasta/s}$$

Kirjain R tulee englannin kielen sanasta *Rate*.

Esimerkki 3.4: Poissonin jakauman sovellus
Radioaktiivisen merkkiaineen puoliintumisaika on 7 kuukautta (210 päivää). Tarkastellaan ensemblea, jossa on yhteensä $1,8144 \cdot 10^8$ tämän aineen atomeja.

a) Määritä keskimääräinen hajoamistaajuus SI-yksiköissä.

b) Tarkastellaan ensimmäistä sekuntia edellisestä ajanjaksosta. Laske todennäköisyys sille, että juuri hajoamistaajuuden ilmoittama määrä ytimiä hajoaa ks. sekunnin aikana.

c) Laske todennäköisyys, että yksikään ydin ei hajoa ensimmäisen sekunnin aikana.

Vastaukset

a) Aloitetaan laskemalla aikayksikössä hajonneiden atomiytimien odotusarvo. Puoliintumisajan kuluessa keskimäärin puolet atomiytimistä hajoaa (näytteen aktiivisuus putoaa puoleen). Jaetaan arvioitu hajonneiden ytimien määrä puoliintumisajalla, jolloin saadaan

$$R = \frac{0,5 \cdot 1,8144 \cdot 10^8}{210 \cdot 24 \cdot 3600 \text{ s}} = 5 \ \frac{\text{hiukkasta}}{s}$$

[7] Perustelu esim. [5], luku 11.

b) Todennäköisyys $P_5(5)$ saadaan Poissonin jakauman yhtälöstä (18) arvoilla $x = 5$ ja $\mu = 5$:

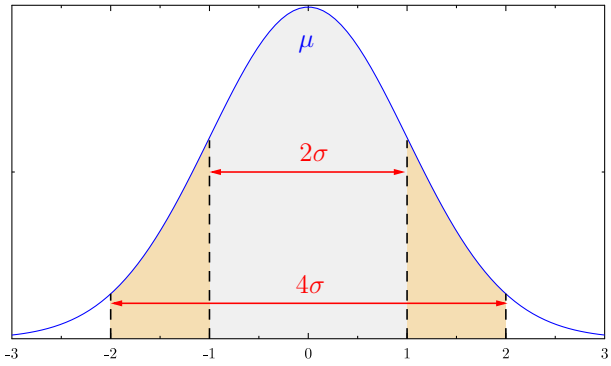
$$P_5(5) = e^{-5} \cdot \frac{5^5}{5!} = 0,17546\ldots \approx 0,18 \tag{20}$$

c) Lasketaan kuten edellisessä kohdassa, mutta korvataan havaittujen hajoamisten määrä arvolla $x = 0$:

$$P_5(5) = e^{-5} \cdot \frac{5^0}{0!} = e^{-5} = 0,00673\ldots \approx 0,01 \tag{21}$$

3.4 Normaalijakauma

Normaalijakauma on erittäin yleinen jatkuva tilastollinen jakauma, jonka kuvaaja on tuttu ”kellokäyrä” tai ”Gaussin käyrä”, ks. kuva 3. Normaalijakauman sijainnin ja muodon määräävät yksikäsitteisesti parametrit μ (odotusarvo) ja σ (keskihajonta) ja usein sanotaan, että satunnaismuuttuja X noudattaa normaalijakaumaa parametreilla μ ja σ, merkitään $X \sim N(\mu, \sigma)$.



Kuva 3: Periaatekuva normaalijakaumasta. Jakauma on symmetrinen odotusarvon μ suhteen. Yksinkertaisuuden vuoksi kuvaan on valittu $\sigma = 1$ ja $\mu = 0$; näillä parametreilla varustettua normaalijakaumaa kutsutaan *standardoiduksi normaalijakaumaksi*.

Normaalijakauma eroaa binomijakaumasta ja Poissonin jakaumasta oleellisesti siinä, että se on *jatkuva*. Näin ollen yksittäisten havaintoarvojen todennäköisyydet normaalijakaumalla laskettaessa ovat nollia, ja niiden sijaan lasketaan todennäköisyyksiä halutun kokoisille väleille. Todennäköisyyden kertoo normaalijakauman käyrän alle jäävä pinta-ala tarkasteluvälillä, kunhan ks. jakauma on ensin muunnettu *standardimuotoon*. Katsotaan tätä hieman tarkemmin.

Kellokäyrän funktionaalinen muoto on

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{22}$$

jota kuitenkin harvoin käytetään "oikeasti" sen matemaattisen monimutkaisuuden vuoksi[8]. Yleensä tehdään muuttujanvaihto

$$Z = \frac{X - \mu}{\sigma} \Leftrightarrow X = Z\sigma + \mu, \tag{23}$$

joka sijoitettuna yhtälöön (22) antaa tulokseksi funktion

$$f(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}z^2} \tag{24}$$

Funktion $f(z)$ odotusarvo on 0 (funktio saavuttaa maksiminsa nollassa) ja keskihajonta 1 (eksponentissa x:n jakajaksi tulee 1). Vastaavaa normaalijakaumaa merkitään $N(0,1)$ ja sitä kutsutaan *standardoiduksi normaalijakaumaksi*. Tätä normaalijakauman erikoistapausta vastaavat todennäköisyydet (käyrän alle jääneet pinta-alat) ovat taulukkokirjakamaa.

Normaalijakauman todennäköisyydet lasketaan tavallisesti siten, että integroidaan (lasketaan pinta-alan suuruutta) kellokäyrän alle jäävää pinta-alaa $-\infty$:stä arvoon z, joka on tarkastelun kannalta mielekäs. Integroinnin tulos on aina välillä $[0,1]$, sillä koko käyrän alle jäävä ala on 1 (vastaa todennäköisyyksien summaa diskreetille todennäköisyysjakaumalle). Tällä ajatusmallilla määritellään normaalijakauman *kertymäfunktio*

$$\Phi(Z \leq z) = \int_{-\infty}^{z} f(z)\, dz = \mathbb{P}(Z \leq z) \tag{25}$$

Yhtälössä (25) termi $\mathbb{P}(Z \leq z)$ kuvaa todennäköisyyttä, jolla satunnaismuuttujan Z arvo on pienempää kuin kiinnitetty z, ts. kellokäyrän alle jäävä pinta-ala $-\infty$:stä arvoon z.

Esimerkki 3.5: todennäköisyyden laskeminen normaalijakauman avulla
Tässä esimerkissä oletetaan, että sinulla on käsissäsi normaalijakaumataulukko (löytyy esimerkiksi MAOLin taulukkokirjasta). Huomaa, että taulukkoa luettaessa puhutaan aina standardoidusta normaalijakaumasta.

a) Laske todennäköisyys $\mathbb{P}(Z \leq 1.1)$.

b) Laske todennäköisyys $\mathbb{P}(Z \leq -0.8)$.

c) Laske todennäköisyys $\mathbb{P}(-0.5 \leq Z \leq 0.4)$.

Vastaukset

a) Luetaan normaalijakaumataulukkoa (pystyakselilta kokonaiset ja ensimmäinen desimaali, vaaka-akselilta toinen desimaali) ja saadaan tulos $\mathbb{P}(Z \leq 1.1) = 0,8643$.

b) Nyt kiinnitetty arvo z on negatiivinen, joten taulukko ei suoraan anna tulosta. Tulee miettiä hieman kääntäen: kokonaispinta-ala on 1, joten halutun pinta-alan suuruus on itse asiassa $1 - \mathbb{P}(Z \leq 0.8) = 1 - 0,7881 = 0,2119$. Tämä tulee suoraan, kun miettii graafisesti pinta-alan kautta.

[8] Funktion integraalia eli käyrän ja x-akselin väliin jäävää pinta-alaa ei ole helppoa laskea suoraan.

c) Viimeisessä kohdassa on äärellinen väli. Mietitään jälleen graafisesti, jolloin havaitaan, että pinta-alan suuruus on itse asiassa välin päätepisteisiin laskettujen todennäköisyyksien erotus:

$$\begin{aligned}
\mathbb{P}(-0.5 \leq Z \leq 0.4) &= \mathbb{P}(Z \leq 0.4) - \mathbb{P}(Z \leq -0.5) \\
&= \mathbb{P}(Z \leq 0.4) - (1 - \mathbb{P}(Z \leq 0.5)) \\
&= 0,6554 - (1 - 0,6915) \\
&= 0,3469
\end{aligned}$$

4 Tilastollisen datan esitysmuodoista

Ennen kuin sukellamme tilastollisen riippuvuuden testaamiseen, käydään lyhyesti läpi erilaisia tapoja esittää dataa. Tarkastellaan myös matemaattisesti sitä, kuinka ei-lineaarinen käyrä saadaan muutettua suoraksi käyttäen esimerkkeinä toisen asteen käyrää (paraabelifunktiota) sekä eksponenttifunktiota. Linearisointia tarvitaan regressiosuoran määrittämisen yhteydessä, jos annetut muuttujat eivät ole suoraan lineaarisesti riippuvaisia toisitaan.

4.1 Taulukot, kuvaajat, pylväsdiagrammit ja histogrammit

Sama tieto voidaan esittää useassa eri muodossa. Oleellista on se, että lukija pystyy tulkitsemaan kirjoittajan esittämän tiedon oikein riippumatta tiedon esitystavasta. Seuraavassa listataan tavallisimman tiedon esitysmuodot sekä niiden erityispiirteet. Esimerkit alla olevista tavoista esittää tietoa löytyvät liitteen luvusta 8.2.

Taulukko sisältää usein hyvin paljon tietoa pienessä tilassa. Tieto on tarkkoja lukuarvoja, joista voidaan tehdä nopeasti eksakteja yksittäisiä päätelmiä. Taulukon huono puoli on siinä, että yleisten suuntaviivojen ja ”trendien” löytäminen vaatii jonkin verran laskemista eikä useinkaan ole selvästi esillä.

Tilastotieteessä puhutaan kahdenlaisista taulukoista. *Havaintomatriisi* on esitys, jossa on luetteloitu kaikki havaitut arvot tapaus kerrallaan havainnointijärjestyksessä (ks. taulukko 15). Havaintomatriisista voidaan edelleen sopivaa luokittelua käyttämällä muodostaa muuttujien *frekvenssitaulu* jonka sarakkeisiin merkitään näkyviin havaittujen arvojen lukumäärät (taulukko 16).

Kuvaaja tarkoittaa tilastollisessa analyysissä yleensä *sirontakuvaajaa*, johon on vaaka-akselille merkitty muuttuja 1 ja pystyakselille muuttuja 2. Vaaka-akselilla oleva muuttuja on tavallisesti *selittäjä*, eli se pyrkii selittämään muuttujan 2 käyttäytymistä. Sirontakuvio esittää näin ollen kahden muuttujan *yhteisjakaumaa*. Esimerkki sirontakuviosta on kuvassa 8.

Pylväsdiagrammi on yksi tilastollisen datan esittämisen peruspilareista. Pylväät voivat olla pystysuorassa tai vaakasuorassa. Mikäli tarkasteltava muuttuja on vähintään järjestysasteikollinen, esitetään pylväät loogisessa suuruus/paremmuusjärjestyksessä. Luokitteluasteikkoiselle muuttujalle voidaan pylväät laittaa mihin tahansa järjestykseen. Pylväiden väliin jätetään yleensä pieni tyhjä tila. Esimerkki on kuvassa 9.

Histogrammi on pylväsdiagrammi tapauksessa, jossa tarkasteltava muuttuja on luonteeltaan jatkuva ja havaintoarvot on jaettu sopiviin luokkiin. Pylväät piirretään kiinni toisiinsa ikään kuin kuvastamaan tilastollisen muuttujan jatkuvuutta (muuten esitystapa on täysin sama kuin pylväsdiagrammille).

4.2 Käyrän linearisointi matemaattisesti

Käyrän (tai paremmin funktion) *linearisointi* tarkoittaa matemaattisesti sitä, että yhtälössä $y = f(x)$ halutaan jollakin muuttujanvaihdolla $x \to z$ saada lauseke $f(x)$ lineaariseen

muotoon $Az + B$, missä kertoimet A (kulmakerroin) ja B (vakiokerroin) ovat numeroita (reaalilukuja). Muuttujanvaihdon tyyppi riippuu täysin siitä, millaista linearisointia vaaditaan, ts. mikä on tarkasteltavan funktion $f(x)$ muoto. Seuraavassa käydään läpi toisen asteen käyrän eli paraabelin ja eksponenttifunktion linearisoinnit. Linearisointia tarvitaan luvussa 5.2, kun tarkastellaan kahden muuttujan välistä tilastollista riippuvuutta regressiosuoran avulla.

Esimerkki 4.1: Toisen asteen käyrän linearisointi
Tarkastellaan kuvan 4 a-kohdan mukaista tilannetta, jossa on esitetty funktio $y = 2x^2 - 2$ positiivisille muuttujan x arvoille. Linearisoidaan käyrä muunnoksella $z = x^2$ ja saadaan yhtälö $y = 2z - 2$. Havaitaan, että kyseessä on suoran yhtälö, missä suoran kulmakerroin on 2 ja vakiotermi on -2. Kuvan 4 a-kohdassa on alkuperäinen käyrä ja b-kohdassa linearisoitu muoto.

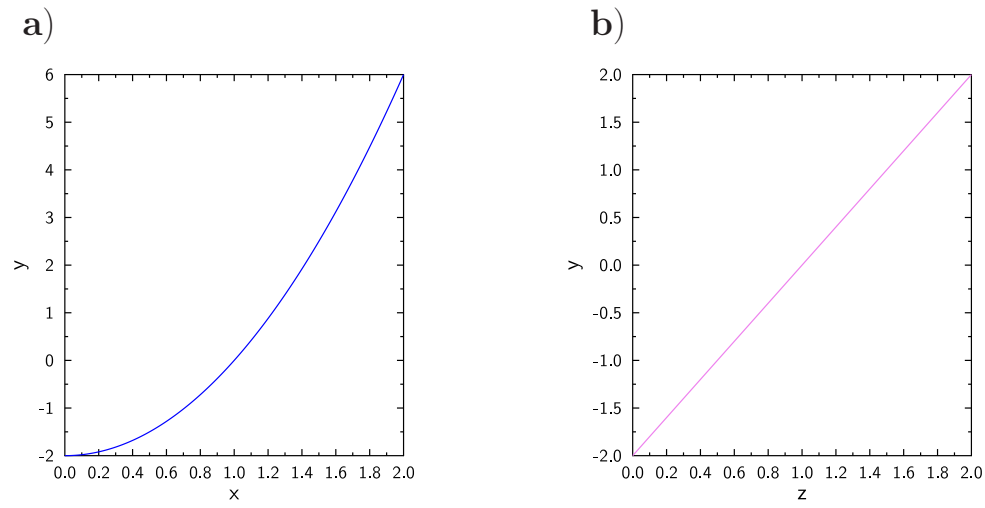
Esimerkki 4.2: Eksponentiaalisen käyrän linearisointi
Tarkastellaan eksponenttifunktiota

$$y = 3e^{-\frac{x}{2}}$$

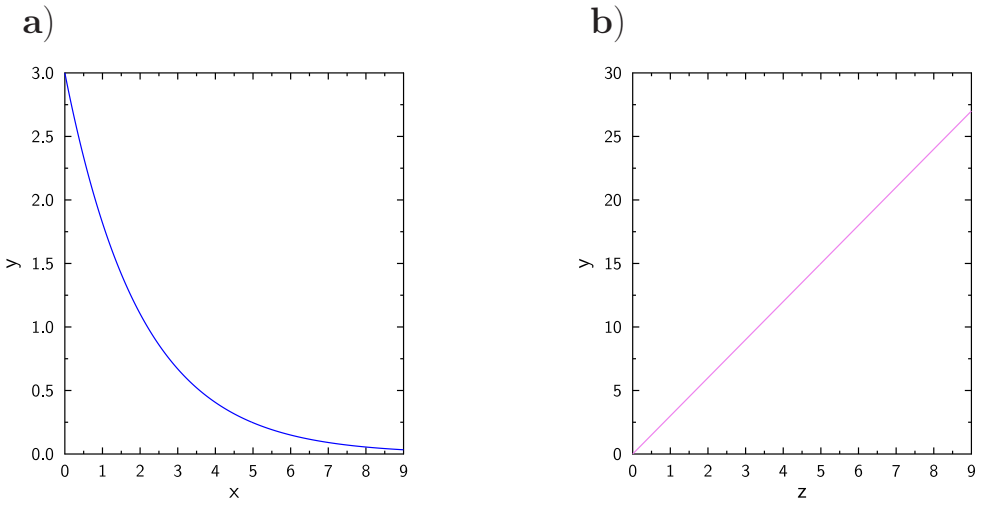
Tehdään muuttujanvaihto $x = -2\ln(z)$ ja kirjoitetaan lauseke hieman eri muodossa, jolloin saadaan

$$\begin{aligned} 3e^{-\frac{x}{2}} &= 3e^{-\frac{-2\ln(z)}{2}} \\ &= 3e^{\ln(z)} \\ &= 3z \end{aligned}$$

Funktio $g(z) = 3z$ on lineaarinen, kulmakerroin $B = 3$ ja vakiotermi $A = 0$. Kuvasta 5 voi tarkastella alkuperäistä funktiota sekä linearisoitua käyrää.



Kuva 4: Esimerkin 4.1 paraabeli (a) ja käyrän linearisoitu muoto (b).



Kuva 5: Esimerkin 4.2 eksponenttifunktio (a) ja funktion linearisoitu muoto (b). Koska teimme ”negatiivisen” muuttujanvaihdon, niin käyrä on periaatteessa kääntynyt ympäri. Tästä ei kuitenkaan ole yleensä haittaa, sillä tilastollisessa analyysissä johtopäätökset tehdään aina alkuperäisille arvoille tekemällä takaisinsijoitus (eli käännetään muuttujanvaihto lopuksi toisin päin).

5 Korrelaatiokerroin ja regressiosuora

Tilastollisen analyysin lähtökohtana on lähes aina muuttujien välisten riippuvuussuhteiden spesifioiminen. Tässä luvussa keskitymme kahteen tapaan ilmaista *tilastollista lineaarista riippuvuutta* eli *korrelaatiokertoimeen* ja *regressiosuoraan*. Menetelmät on tarkoitettu kahden tilastollisen muuttujan välisen riippuvuuden kuvailuun[9].

5.1 Korrelaatiokerroin

Kahdelle tilastolliselle muuttujalle X ja Y määritellään (Pearsonin tulomomentti)korrelaatiokerroin r_{xy} muodossa

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{26}$$

Korrelaatiokertoimen muistaa parhaiten siitä, että se suhteuttaa muuttujien X ja Y yhteisvaihtelun eli kovarianssin s_{xy} muuttujien yksittäisiin tai "sisäisiin" vaihteluihin eli keskihajontoihin s_x ja s_y. Filosofisessa mielessä korrelaatio on kahden muuttujan yhdenmukaisen käytöksen mittari, ts. se kertoo, miten muuttujien arvojen muuttuessa vaihtelut vastaavat toisiaan.

Kertoimella r_{xy} on seuraavat ominaisuudet:

- $r_{xy} \approx 1 \Rightarrow$ erittäin voimakas positiivinen korrelaatio eli muuttuja X kasvaa, kun muuttuja Y kasvaa. Hetkisen kuluttua huomataan, että tämäntyyppistä korrelaatiota vastaa nouseva regressiosuora.
- $r_{xy} \approx -1 \Rightarrow$ erittäin voimakas negatiivinen korrelaatio eli muuttuja X pienenee muuttujan Y kasvaessa ja päinvastoin. Tilannetta vastaa laskeva regressiosuora.
- $r_{xy} \approx 0 \Rightarrow$ ei korrelaatiota muuttujien X ja Y välillä. Sirontakuviossa tämä tarkoittaa tilannetta, jossa havainnot ovat kuin "haulikolla ammuttuja" eli sijoittuvat minne sattuu tai vaihtoehtoisesti noudattavat ei-lineaarista käyrää, jolloin on kenties mahdollista linearisoida ja jatkaa analyysiä tätä kautta.

Esimerkki 5.1: Korrelaatiokerroin
Taulukossa 4 on esitetty erään yrityksen työntekijöiden viikkopalkat ja viikottaisten työtuntien määrät. Saavatko kovimmat ahertajat vaivoistaan suurimman korvauksen?

Vastaus Korrelaatiota varten tarvitsemme muuttujien X = h/vko ja Y = €/vko keskihajonnat ja kovarianssin. Aloitetaan laskemalla keskiarvot:

$$\overline{x} = \frac{36 + 60 + 19 + 45}{4} \text{ h/vko} = 40 \text{ h/vko}$$
$$\overline{y} = \frac{800 + 500 + 1400 + 740}{4} \text{ h/vko} = 860 \text{ €/vko}$$

[9] Useamman muuttujan menetelmät (varianssianalyysi) sivuutetaan.

Taulukko 4: Työntekijöiden viikkopalkat ja viikottaiset tuntimäärät.

Henkilö	h/vko	€/vko
1	36	800
2	60	500
3	19	1400
4	45	740

Taulukoidaan poikkeamat keskiarvosta ja niiden neliöt (taulukko 5). Muuttujan X varianssille saadaan

$$s_x^2 = \frac{1}{4-1}\{16 + 400 + 441 + 25\} = 294,$$

josta edelleen keskihajonta on $s_x = \sqrt{294} = 17,1464\ldots \approx 17$. Täysin analogisella laskulla saadaan muuttujan Y keskihajonnaksi $s_y = 382,6225\ldots \approx 383$. Kovarianssi on

$$\begin{aligned} s_{xy} &= \frac{1}{3}\{(-4)\cdot(-60) + 20\cdot(-360) + (-21)\cdot 540 + 5\cdot(-120)\} \\ &= -6300 \end{aligned}$$

Taulukko 5: Työntekijöiden palkkojen ja tuntimäärien poikkeamat.

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	y_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$
36	-4	16	800	-60	3600
60	20	400	500	-360	129600
19	-21	441	1400	540	291600
45	5	25	740	-120	14400

Sijoitetaan lasketut tulokset korrelaatiokertoimen yhtälöön (26), jolloin saadaan

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-6300}{17,1464 \cdot 381,6193} = -0,96027\ldots \approx -0,96$$

Havaittiin erittäin voimakas negatiivinen korrelaatio. Tulos tulkitaan siten, että mitä enemmän työtunteja firmassa tekee, sitä vähemmän saa palkkaa.

Yleensä tilastollisia menetelmiä voidaan soveltaa vain, mikäli aineisto on tarpeeksi suuri. Tässä esimerkissä haluttiin vain antaa periaatteellinen kuva laskuista, jotka käytännössä aina tehdään tietokoneella tilasto-ohjelmalla (kukapa haluaisi laskea käsin esimerkiksi 200:n muuttujan keskihajontaa...). Liian pienen otannan tapauksessa otantavirheen suuruus voi olla hyvin merkittävä; edellä tuskin kukaan haluaisi työskennellä firmassa, jossa palkka laskee lineaarisesti työpanoksen mukaan (voidaan olettaa, että saatu tulos oli seurausta juuri suuresta otantavirheestä).

Huomatus. Korrelaatiokerroin r_{xy} toimii vain muuttujille, jotka ovat vähintään välimatka-asteikollisia. Mikäli muuttuja on järjestysasteikollinen, käytetään yleensä (Spearmanin) *järjestyskorrelaatiokerrointa*, jonka voi halutessaan lueskella vaikkapa teoksesta [1].

5.2 Regressiosuora

Tarkastellaan kahta tilastollista muuttujaa X ja Y, joita vastaavat havaintoarvot ovat $x_1, \ldots, x_n$ ja $y_1, \ldots, y_n$ ja joiden välillä on *lineaarinen riippuvuussuhde*[10]. Kun haluamme esittää tämän riippuvuuden matemaattisesti, käytämme suoran yhtälöä

$$y = ax + b, \tag{27}$$

missä kertoimet a ja b ovat vakioita: a on suoran kulmakerroin ja b suoran ja y-akselin leikkauskohta. Tavallisesti kertoimet määritetään *PNS-menetelmällä* (pienimmän neliösumman menetelmä), jolla minimoidaan muuttujien poikkeamien neliöiden summaa. Matemaattisesti voidaan osoittaa, että pistejoukkoon $\{(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)\}$ sovitetun PNS-suoran kertoimet a ja b määräytyvät yhtälöistä

$$a = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} = \frac{s_{xy}}{s_x^2} \tag{28}$$

$$b = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} = \overline{y} - a\overline{x} \tag{29}$$

Yllä kaikki summat ovat yli indeksijoukon $i \in \{1, \ldots, n\}$. Huomaa, että molempien kertoimien nimittäjissä on sama termi. Koska mittauspisteisiin (x_i, y_i) liittyy jokaiseen tietyn suuruinen mittausvirhe[11], niin kertoimien a ja b virheille voidaan johtaa yhtälöt[12]

$$\Delta a = \frac{1}{\sqrt{n-2}} \cdot \frac{\sqrt{n\sum_i y_i^2 - \left(\sum_i y_i\right)^2}}{\sqrt{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}} \tag{30}$$

$$\Delta b = \Delta a \cdot \sqrt{\frac{\sum_i x_i^2}{n}} \tag{31}$$

Regressiosuora liittyy usein kiinteästi yhteen korrelaatiokertoimen kanssa. Näin ollen joskus voi olla mielekästä laskea parametrien arvot kovarianssin ja varianssin yhtälöiden kautta. Kaavat on myös helppo muistaa, kun ajattelee, että esimerkiksi parametri a (suoran kulmakerroin) saadaan laskettua muuttujien X ja Y "yhteisvaihtelutekijän" eli kovarianssin ja muuttujan X varianssin (vaihtelun) osamääränä. Kertoimien virhekaavat ovat puolestaan sen verran persoonallisia, että ne havaitsee kokeen yhteydessä jaettavasta mahdollisesta kaavakokoelmasta väkisin (näitä kaavoja ei ole tarkoitus opetella ulkoa).

[10] Huomaa, että kyseessä voi olla *positiivinen*/*negatiivinen* lineaarinen riippuvuus, joka vastaa nousevaa/laskevaa suoraa.

[11] Tarkalleen ottaen tässä oletetaan, että ainoastaan koordinaattiin y liittyy virhe; ks. [5], luku 8.2.

[12] PNS-suoran ymmärtäminen vaatii jonkin verran matemaattista osaamista, katso esimerkiksi [5], luvut 8.2.–8.4.

Regressiosuoraa laskettaessa vaaka-akselille laitetaan yleensä muuttuja, jonka tehtävänä on selittää toisessa muuttujassa tapahtuvia vaihteluita (esimerkiksi matematiikan opiskelun määrä voi selittää matemaattista ongelmanratkaisukykyä, jota mitataan matematiikan kokeen arvosanalla tms.). Vaaka-akselille laitettavaa muuttujaa kutsutaan *selittajäksi* (faktoriksi) ja pystyakselille laitettavaa muuttujaa *selitettäväksi muuttujaksi* (vastemuuttujaksi), ks. luku 4.

Esimerkki 5.2: Regressiosuoran laskeminen
Taulukossa 6 on viiden urheilijan mitatut hemoglobiiniarvot sekä 10 km:n testijuoksun ajat. Sovita tutkimusdataan PNS-menetelmällä regressiosuora ja piirrä tilanteesta mallikuva. Voidaan olettaa, että suoran parametrien a ja b virheet ovat merkityksettömän pieniä.

Taulukko 6: Urheilijoiden hemoglobiiniarvot ja 10 km:n juoksuajat.

Urheilija	Hemoglobiini (g/l)	10 km:n aika (min.ss)
1	163	40.35
2	175	36.26
3	159	41.55
4	184	33.41

Vastaus Lasketaan regressiosuoran parametrien arvot pitkän kaavan mukaan eli ei käytetä varianssia ja kovarianssia (näin tehdään, koska korrelaatiokertoimesta ei olla kiinnostuneita). Muunnetaan ajat sekunneiksi (päässälasku) ja lasketaan aputuloksena ensin muutama regressiokertoimien a ja b yhtälöissä esiintyvä summatermi:

$$\sum_i x_i y_i = (163 \cdot 2435 + 175 \cdot 2186 + 159 \cdot 2515 + 184 \cdot 2021)\ \text{gs/l} = 1551204\ \text{gs/l}$$

$$\sum_i x_i = (163 + 175 + 159 + 184)\ \text{g/l} = 681\ \text{g/l}$$

$$\sum_i y_i = (2435 + 2186 + 2515 + 2021)\ \text{s} = 9157\ \text{s}$$

$$\sum_i x_i^2 = (163^2 + 175^2 + 159^2 + 184^2)\ (\text{g/l})^2 = 116331\ (\text{g/l})^2$$

$$\sum_i y_i^2 = (2435^2 + 2186^2 + 2515^2 + 2021^2)\ \text{s}^2 = 21117487\ \text{s}^2$$

Edellä muuttujalle X yksiköt ovat g/l (hemoglobiinin määrä veressä) ja muuttujalle Y s

(sekunti). Sijoitetaan lasketut termit ensin kertoimen a yhtälöön:

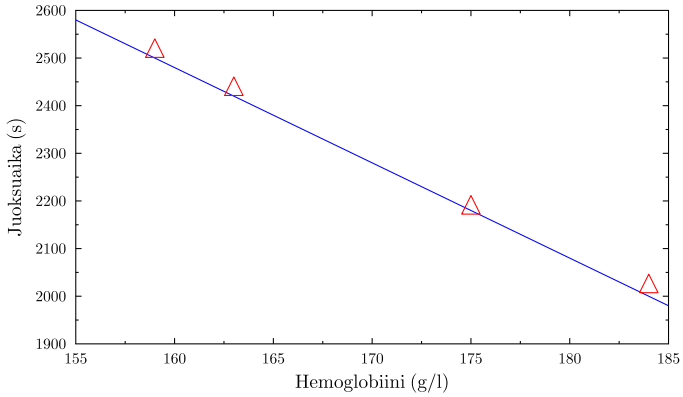
$$\begin{aligned} a &= \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \\ &= \frac{4 \cdot 1551204 \text{ gs/l} - 681 \text{ g/l} \cdot 9157 \text{ s}}{4 \cdot 116331 \text{ (g/l)}^2 - (681 \text{ g/l})^2} \\ &= -19,89827\ldots \frac{\text{s}}{\text{g/l}} \end{aligned}$$

Vastaavasti kertoimelle b saamme

$$\begin{aligned} b &= \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \\ &= \frac{116331 \text{ (g/l)}^2 \cdot 9157 \text{ s} - 681 \text{ g/l} \cdot 1551204 \text{ gs/l}}{4 \cdot 116331 \text{ (g/l)}^2 - (681 \text{ g/l})^2} \\ &= 5676,930902\ldots \text{ s} \end{aligned}$$

Huomaa, että kerroin b olisi saatu melko helposti myös laskemalla keskiarvot molemmille muuttujille ja sijoittamalla ne yhtälöön $b = \overline{y} - a\overline{x}$. Nyt voidaan kirjoitta regressiosuora muodossa

$$y = -20 \cdot x + 5680 \tag{32}$$



Kuva 6: Urheilijoidet henoglobiiniarvot ja juoksuajat sekä regressiosuora. Havaitsemme, että suora sopii melko hyvin pistejoukkoon.

Esimerkki 5.3: Regressioparametrien virheet
Jatketaan edellistä esimerkkiä 5.2 laskemalla regressioparametreille myös virheet. Tulokset saadaan yhtälöstä (30):

$$\begin{aligned}\Delta a &= \frac{1}{\sqrt{4-2}} \cdot \frac{\sqrt{4 \cdot 21117487 - 9157^2}}{\sqrt{4 \cdot 116331 - 681^2}} \ \frac{\text{s}}{\text{g/l}} \\ &= 14,075519\ldots \\ \Delta b &= 14,075519 \cdot \sqrt{\frac{116331}{4}} \ \text{s} \\ &= 2400,39189\ldots\end{aligned}$$

Lopullinen regressiosuora voidaan nyt kirjoittaa muodossa

$$y = (-20 \pm 14)x + (5700 \pm 2400)$$

Havaitaan, että virhe-estimaatit ovat melko suuria verrattuna tuloksiin (samaa suuruusluokkaa). Tämä johtuu suurelta osin siitä, että otoskoko n on erittäin pieni ajatellen tilastollista analyysiä (jakajassa $\sqrt{n}$ ei paljoa vaikuta).

Esimerkki 5.4: Datan linearisointi ja regressiosuoran sovitus
Ultraäänen intensiteettiä ihmiskehossa syvyydellä x voidaan arvioida eksponentiaalisella vaimenemislailla

$$I(x) = I_0 e^{-ax},$$

missä I_0 on ultraäänen intensiteetti kehon ja ilman rajapinnassa ennen allon saapumista kehoon ja a on vaimenemiskerroin, jonka yksiköt ovat dB/cm. Ihmisen pehmytkudoksia tutkittaessa on saatu taulukon 7 mukaista dataa ultraäänen intensiteetille ja etäisyydelle. Vaimenemiskertoimen suuruus on $a = 0,3$ dB/cm ja intensiteetti kehon pinnalla on $I_0 = 2 \cdot 10^{-4}$ W/m^2.

a) Linearisoi annettu data siten, että intensiteetin ja uuden linearisoinnissa määriteltävän muuttujan z välillä on lineaarinen riippuvuus. Käytä apunasi ultraäänen intensiteetin eksponentiaalista yhtälöä.

b) Määritä intensiteetin $I(x)$ ja muuttujan z välinen regressiosuora (virheineen) siten, että faktorimuuttujana on z ja vastemuuttujana intensiteetti. Piirrä mallikuva.

Vastaus

a) Eksponentiaalinen yhtälö antaa vihjeen sille, millaisen muuttujanvaihdon teemme. Kirjoitetaan yhtälö hieman eri muodossa:

$$I(x) = I_0 e^{-ax} = I_0 z,$$

missä $z = e^{-ax}$ on uusi muuttuja eli teemme muuttujanvaihdon $x = -\ln(z)/a$. Laskemalla datapisteiden arvot muuttujalle z saamme taulukon 8.

Taulukko 7: Ultraäänen mitatut intensiteetit ja niitä vastaavat etäisyydet.

$I(x)(\cdot 10^{-5})$ W/m^2	x (m)
4,46	0,05
2,45	0,07
0,10	0,10
0,05	0,12

Taulukko 8: Ultraäänen mitatut intensiteetit ja uuden muuttujan z arvot.

$I(z)(\cdot 10^{-5})$ W/m^2	z
4,46	0,22313
2,45	0,12246
0,10	0,04979
0,05	0,02732

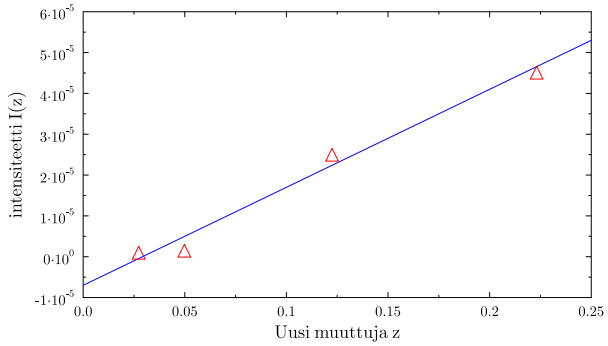
b) Faktorimuuttuja tarkoittaa selittävää muuttujaa, joka laitetaan x-akselille. Vastemuuttuja laitetaan y-akselille. Regressiosuoran parametrit virheineen lasketaan kuten esimerkeissä 5.2 ja 5.3; tulokseksi saadaan

$$a = (2,3799 \pm 1,6971) \cdot 10^{-4} \quad ; \quad b = (-0,7500 \pm 2,2130) \cdot 10^{-5},$$

eli regressiosuora voidaan lyhyesti kirjoittaa muodossa

$$I(z) = (2,4 \pm 1,7) \cdot 10^{-4} z - (0,7 \pm 2,2) \cdot 10^{-5}$$

Regressiosuora on piirretty kuvaan 7.



Kuva 7: Intensiteetin $I(z)$ ja muuttujan z datapisteet ja regressiosuora.

6 Harjoitustehtäviä

Laskuharjoitustehtävät ovat samassa loogisessa järjestyksessä kuin monisteen teoriaosuus. Tehtävät on jaettu kolmeen pääkategoriaan:

- Tunnusluvut ja hajontaluvut
- Tilastolliset jakaumat ja tilastollisen datan esittäminen
- Korrelaatio ja regressio.

Ajatuksena on, että tehtävissä edetään siinä järjestyksessä, kun ne on tässä esitetty. Myöhemmissä tehtävissä kerrataan myös aiemmissa laskuissa esiintyneitä asioita, jotta kokonaisuudesta on helpompi muodostaa järkevä kuva. Tehtävien vastaukset ovat tehtäviä seuraavassa luvussa 7.

Tunnusluvut ja hajontaluvut

1. Jatkoa esimerkkiin 5.2. Määritä urheilijoiden hemoglobiiniarvojen ja juoksuaikojen mediaanit, keskiarvot ja keskihajonnat.

2. Apteekissa myydään erilaisia voimakkaita särkylääkkeitä, joiden pakettihinnat ovat seuraavat:

$$4\ €,\ 6\ €,\ 3\ €,\ 8\ €\ \text{ja}\ x\ €$$

Kuinka suuri on hinta x, jos hintojen keskiarvoksi on saatu 7,5 €? Laske myös hintojen keskiarvon keskivirhe.

Tilastolliset jakaumat ja tilastollisen datan esittäminen

1. Tarkastellaan binomijakaumaa.

a) Kuinka monella eri tavalla voidaan viiden laskuharjoitustehtävän joukosta valita kolme, jos järjestyksellä on merkitystä?

b) Laske todennäköisyydet $\mathcal{B}(11; 4; 0,32)$ ja $\mathcal{B}(5; 2; 0,04)$.

2. Viidenkymmenen opiskelijan ryhmä valmistautuu lääketieteen pääsykokeisiin. Koetta edeltävänä yönä useat heistä ovat lukeneet pääsykoekirjaa erittäin myöhään eivätkä ole nukkuneet läheskään tarpeeksi. Todennäköisyys sille, että opiskelija nukkuu pääsykoeaamuna pommiin, on 0,25. Laske

a) todennäköisyys, jolla 10 henkilöä nukkuu pommiin. **[0,1]**

b) jakauman odotusarvo ja keskihajonta pommiin nukkuneiden määrälle. **[13 ± 3]**

c) todennäköisyys, että vähintään 2 nukkuu pommiin. **[1]**

3. Katsotaan Galenos-teoksesta [6] esimerkkinä luonnossa esiintyvän ionisoivan taustasäteilyn mittaamista. Puolijohdearkkitehtuurilla varustettu mittauslaitteisto ilmoittaa säteilyn aiheuttamien pulssien lukumäärän sekuntia kohden. Mittauksia on tehty yhteensä 1000 ja niiden tulokset ovat taulukossa 9 (frekvenssitaulu).

Taulukko 9: Mitattujen pulssien frekvenssijakauma.

x_i (pulssia/s)	f_i (lukumäärä)
0	47
1	152
2	221
3	233
4	160
5	105
6	51
7	17
8	10
9	3
10	1
11	0
12	0

a) Laske jakauman odotusarvo ja keskihajonta. Käytä keskihajonnan laskemisessa Poissonin statistiikkaa.

b) Määritä havaintojen perusteella klassinen todennäköisyys sille, että havaitaan sekunnin aikana kaksi pulssia.

c) Laske Poissonin statistiikan mukainen todennäköisyys havaita ainakin kaksi pulssia sekunnissa.

4. Sairaalassa kirjataan ylös syntyneiden lasten lukumäärä $n = 20$ kahden päivän aikana. Käyttäen Poissonin jakaumaa, laske

a) Keskimääräinen syntyneiden lasten lukumäärä tuntia kohden.

b) Todennäköisyys sille, että kaksi lasta syntyy tunnissa.

5. Poliisi on saanut kiinni kymmenen henkilöä, joita epäillään esiintymisestä valelääkäreinä. Aiemman tilastollisen analyysin perusteella voidaan arvioida, että pidätetyistä 7 % on oikeasti syyllisiä. Laske

a) valelääkärien määrän odotusarvo ja keskihajonta.

b) todennäköisyys, että kaksi henkilöä on valelääkäreitä.

6. Määritä seuraavat normaalijakauman todennäköisyydet:

a) $\mathbb{P}(Z \geq 1.1)$

b) $\mathbb{P}(Z \leq -0.5)$

c) $\mathbb{P}(-0.7 \leq Z \leq 1.7)$

7. Tarkastellaan taulukon 10 dataa, jossa on analysoitu koehenkilöiden työtyytyväisyyttä asteikolla 1 = ”vihaan työtäni”, 2 = ”en pidä työstäni”, 3 = ”en osaa sanoa,” 4 = ”pidän työstäni” ja 5 = ”rakastan työtäni”.

a) Muodosta annetusta havaintomatriisista frekvenssitaulu. Luokittele kyselyyn vastaajat ikäryhmittäin 18–30, 31–45 ja 46+.

b) Piirrä samalla luokittelulla histogrammi (pylväät pystyssä), jossa pystyakselille tulee kunkin luokan vastausten mediaani.

Taulukko 10: Mielipiteet työtyytyväisyydestä.

Henkilö	Ikä	Työtyytyväisyys
1	23	4
2	34	5
3	18	1
4	56	3
5	45	2
6	76	5
7	27	4
8	45	2
9	76	5
10	21	4
11	53	1
12	71	5
13	22	2
14	37	1
15	38	5

8. Linearisoi seuraavat funktiot sopivilla muunnoksilla:

a) $y = (x-1)^2 + 2x - 3$

b) $y = 8x^2 - 7$

c) $y = e^{-\frac{1}{x+4}}$

d) $y = \frac{2}{e^{x^2-3}}$

Korrelaatio ja regressio

1. Jatkoa esimerkkiin 5.2. Laske korrelaatiokerroin ja vertaa sitä esimerkin 5.2 tulokseen.

2. Laajennetaan esimerkkiä 2.2 siten, että kirjoitetaan taulukkoon 1 myös opiskelijoiden pääsykokeeseen valmistautumisajat. Tiedot löytyvät uudesta taulukosta 11.

Taulukko 11: Lääketieteen pääsykokeisiin osallistuneiden henkilöiden pistemääriä valintakokeessa.

Henkilö	Pistemäärä	Valm. aika (h)
1	58	100
2	124	500
3	83	340
4	38	220
5	135	480

a) Koepistemäärien keskiarvo ja keskihajonta on jo laskettu aiemmin. Laske vastaavat tilastolliset tunnusluvut valmistautumisajalle sekä muuttujien kovarianssi.

b) Laske muuttujien välinen korrelaatiokerroin ja tulkitse tulos sanallisesti.

c) Määritä regressiosuoran yhtälö (virheineen) ja piirrä tilanteesta mallikuva.

3. Uusi kauppoihin tullut flunssalääke mainostaa, että syömällä enemmän lääkettä flunssa paranee nopeammin:

> "... Tuotteemme toimii sitä paremmin, mitä enemmän syöt lääkettämme. Taulukosta 12 voit lukea annostukset sen mukaan, kuinka nopeasti haluat parantua flunssasta. Sivuvaikutuksena liiallisesta lääkkeen käytöstä voi esiintyä kuumetta, nuhaa ja yskää. Keskustele lääkärisi kanssa ennen lääkkeen käytön aloittamista."

Taulukko 12: Flunssasta parantuminen lisää lääkkeitä syömällä.

Parantumisaika (tunteja)	Lääkemäärä (kapseleita)
102	1
86	2
54	4
24	8
13	17

a) Piirrä datasta mallikuva ja linearisoi data sopivalla muuttujanvaihdolla.

b) Laske regressiosuoran parametrit virheineen.

4. Katsotaan taulukon 13 mukaista aineistoa, jossa on tutkittu työntekijän viikottaisen liikunnan tuntimäärää LII (h/vko) ja vuosittaisten sairaslomapäivien SAI (pvä/vuosi) lukumäärää.

a) Laske tilastollisten muuttujien LII ja SAI keskiarvot, keskihajonnat ja määrää muuttujien tyypit.

Taulukko 13: Työntekijän viikottaisen liikunnan määrä ja vuosittaiset sairaslomapäivät.

ID	LII (h/vko)	SAI (pvä/v)
1	3	10
2	12	5
3	7	13
4	1	2

b) Laske muuttujien LII ja SAI välinen korrelaatiokerroin ja tulkitse tulos sanallisesti.

7 Vastaukset harjoitustehtäviin

Seuraavassa pelkät numeroarvolliset vastaukset tehtäviin. Myöhemmin saattaa ilmestyä myös tehtävien malliratkaisut.

Tunnusluvut ja hajontaluvut

1. Hemoglobiinille X saamme $M_{d_x} = 167$ g/l (välimatka-asteikollinen muuttuja ja parillinen määrä havaintoja), $\overline{x} = 170,25$ g/l ja $s_x = 11,4127\ldots \approx 11,4$. Juoksuajalle Y saadaan vastaavasti luvut $M_{d_y} = 2310,5$ s, $\overline{y} = 2289,25$ s ja $s_y = 227,1744\ldots \approx 228$.

2. Hinta $x = 16,5$ €, keskiarvon keskivirhe $s_{\overline{x}} = 2,408\ldots \approx 2,4$.

Tilastolliset jakaumat ja tilastollisen datan esittäminen

1. Binomijakauman perusominaisuuksia...

a) 10 mahdollisuutta (binomikaavalla).

b) $\mathcal{B}(11;4;0,32) = 0,2326\ldots \approx 0,23$, $\mathcal{B}(5;2;0,04) = 0,0141\ldots \approx 0,01$

2. Pommiin nukkuvat opiskelijat...

a) $\mathcal{P}(10 \text{ myöhästyy}) = 0,0985\ldots \approx 0,1$

b) $\mu = 12,5 \approx 13$, $\sigma = 3,06186\ldots \approx 3$

c) $\mathcal{P}(\text{väh. } 2 \text{ myöhästyy}) = 0,999\ldots \approx 1$

3. Galenoksen ionisoivan taustasäteilyn mittaus...

a) Odotusarvo $\mu = 3$, keskihajonta $\sigma = \sqrt{3}$.

b) Klassinen todennäköisyys = havaintojen lkm/kaikki havainnot eli 0,16.

c) $\mathcal{P}(\text{ainakin 2 pulssia /s}) = 0,8008\ldots \approx 0,8$.

4. Sairaalassa syntyneet lapset...

a) Lapsia syntyy keskimäärin 5/12 tuntia kohden.

b) $P_{5/12}(2) = 0,05722\ldots \approx 0,06$.

5. Valelääkärit... käytetään binomijakaumaa, toisensa poissulkevat tapaukset ovat ”on valelääkäri”, jota vastaa todennäköisyys $p = 0,07$ ja ”ei ole valelääkäri,” jota vastaa todennäköisyys $q = 1 - p = 0,93$.

a) Binomijakauman odotusarvo on $\mu = np = 10 \cdot 0,07 = 0,7 \approx 1$ eli odotusarvona yksi henkilöistä on valelääkäri. Keskihajonnaksi saamme

$$\sigma = \sqrt{10 \cdot 0,07 \cdot 0,93} = 0,806845\ldots \approx 0,8.$$

b) Laitetaan binomijakauman parametreiksi $n = 10,\ x = 2$ ja todennäköisyydet $p = 0,07;\ q = 0,93$ ja kirjoitetaan todennäköisyys:

$$\begin{aligned}\mathcal{B}(10;2;0,07) &= \binom{10}{2} \cdot 0,07^2 \cdot 0,93^8 \\ &= \frac{10!}{2!8!} \cdot 0,07^2 \cdot 0,93^8 \\ &= 0,12338\ldots \\ &\approx 0,12\end{aligned}$$

Todennäköisyys sille, että kaksi pidätettyä todella on valelääkäreitä, on noin 12 prosenttia.

6. Normaalijakauman todennäköisyydet...

a)
$$\mathbb{P}(Z \geq 1.1) = 1 - \mathbb{P}(Z \leq 1.1) = 1 - 0,8643 = 0,1357.$$

b)
$$\mathbb{P}(Z \leq -0.5) = 1 - \mathbb{P}(Z \leq 0.5) = 1 - 0,6915 = 0,3085.$$

c)
$$\begin{aligned}\mathbb{P}(-0.7 \leq Z \leq 1.7) &= \mathbb{P}(Z \leq 1.7) - \mathbb{P}(Z \leq -0.7) \\ &= \mathbb{P}(Z \leq 1.7) - [1 - \mathbb{P}(Z \leq 0.7)] \\ &= 0,9554 - (1 - 0,7580) \\ &= 0,7134\end{aligned}$$

7. Taulukot ja kuvat ilmestyvät tänne myöhemmin. Mediaanit ovat 4, 2 ja 5 vastaavissa luokissa 18–30, 31–45 ja 46+.

8. Linearisoinnit...

a) $y = z - 2$, muuttujanvaihto $z = x^2$

b) $y = 8z - 7$, muuttujanvaihto $z = x^2$

c) $y = z$, muuttujanvaihto $\ln(z) = -\frac{1}{x+4} \Leftrightarrow z = e^{-\frac{1}{x+4}}$.

d) $y = 2z$, muuttujanvaihto $x^2 - 3 = z \Leftrightarrow x = \pm\sqrt{3 - \ln(z)}$.

Korrelaatio ja regressio

1. Esimerkin 5.2 jatkoa. Korrelaatiokertoimeksi saadaan $r_{xy} = -0,999\ldots \approx -1$ eli suuri hemoglobiiniarvo lyhentää juoksuaikaa. Tämä on sopusoinnussa aiemmin tehdyn regressioanalyysin kanssa.

2. Esimerkin 2.2 jatkoa.

a) Vamistumisajan X keskiarvo, varianssi ja keskihajonta ovat $\overline{x} = 328$ h, $s_x^2 = 29120$ ja $s_x \approx 171$. Muuttujien X ja $Y =$ koepistemäärä kovarianssiksi saadaan $s_{xy} = 6386,6$.

b) Korrelaatiokerroin $r_{xy} = 0,89798\ldots \approx 0,90$ eli muuttujien välillä on voimakas positiivinen korrelaatio.

c) Regressiosuoran parametreille ja parametrien virheille saadaan arvot

$$\begin{aligned} a &= 0,21932\ldots \\ b &= 15,6696\ldots \\ \Delta a &= 0,140817\ldots \\ \Delta b &= 50,94386\ldots \end{aligned}$$

Regressiosuora voidaan kirjoittaa muodossa

$$y = (0,22 \pm 0,14)x + (15,67 \pm 50,94)$$

Parametrien a ja b erittäin suuret virheet ovat seurausta hyvin pienestä otoskoosta (virhetermien jakajassa n on melko pieni; otoskoon kasvaessa virhe pienenee).

3. Uusi flunssalääke...

a) Mallikuva tulee myöhemmin. Kuvasta havaitaan, että parantumisajan Y ja lääkemäärän X välillä on riippuvuus $Y \propto \frac{1}{\sqrt{x}}$ ja tämän mukaan tehdään muuttujanvaihto $z = \frac{1}{\sqrt{x}}$. Linearisoitu data on esitetty taulukossa 14.

Taulukko 14: Flunssasta parantuminen lisää lääkkeitä syömällä, linearisoitu data muuttujanvaihdolla $z_i = \frac{1}{\sqrt{x_i}}$.

y_i (h)	z_i
102	1
86	$1/\sqrt{2}$
54	$1/2$
24	$1/\sqrt{8}$
13	$1/\sqrt{17}$

b) Regressiosuoran parametreille ja niiden virheille saadaan arvot

$$\begin{aligned} a &= 32,79317\ldots \\ b &= -3,665388\ldots \\ \Delta a &= 37,82128\ldots \\ \Delta b &= 23,5211\ldots \end{aligned}$$

Lopullinen regressiosuora voidaan kirjoittaa muodossa

$$y = (32 \pm 38)z + (-4 \pm 24))$$

Parametrien virheet ovat jälleen aivan järkyttävän suuret johtuen liian pienestä otoskoosta. Mikäli otoskoko olisi[13] esimerkiksi 100, niin virhe pienenisi noin kymmesosaan nykyisestä.

4. Merkitään X =LII ja Y =SAI, ts. liikunnan määrän oletetaan selittävän sairaslomapäivien määrää. Molemmat muuttujat ovat välimatka-asteikollisia ja diskreettejä: liikunnan määrä on pyöristetty tunteihin ja sairasloman pituus kokonaisiin päiviin. Keskiarvoille saadaan $\overline{x} = 5,75$ ja $\overline{y} = 7,5$. Vastaavasti keskihajonnoille $s_x = 4,8562\ldots$ ja $s_y = 4,9328\ldots$.

5. Muuttujien välinen kovarianssi on $s_{xy} = 3,5$ ja korrelaatiokertoimen arvoksi saadaan $r_{xy0} = 0,1555\ldots \approx 0,16$ eli muuttujien välillä on hyvin pieni (käytännössä merkityksetön) positiivinen korrelaatio.

[13]Tässä analyysissähän periaatteessa on oletettu, että on mitattu vain viisi erillistä parantumisaikaa ja niiden perusteella tehdään analyysi. Käytännössä parantumisajoille annetut arvot ovat keskiarvoja tms. useista havainnoista (voi olla vaikkapa tuhat mittausta yhtä arvoa kohti), joten oikeasti virheelle saataisiin huomattavasti pienempi arvo.

8 Liite

8.1 Otantamenetelmät lyhyesti

Tilastollinen tutkimus perustuu siihen, että havaintoyksiköt (perusjoukon alkiot) on valittu satunnaisesti perusjoukosta. Satunnaistaminen voidaan toteuttaa usealla eri tavalla, seuraavassa muutama esimerkki.

Yksinkertainen satunnaisotanta tarkoittaa alkioiden valitsemista kaikki alkiot käsittävästä luettelosta esimerkiksi arpomalla (annetaan jokaiselle alkiolle indeksinumero ja arvotaan haluttu määrä numeroita otokseen).

Systemaattinen otanta Voidaan toteuttaa esimerkiksi siten, että arvotaan yksi alkio ja valitaan siitä eteenpäin joka k. alkio kaikkien alkioiden luettelosta. Jos alkioita on kaikkiaan N kappaletta ja otoskoko on n, niin poimintaväli on $k = N/n$. Systemaattista otantaa käyttäessä tulee olla tarkkana sen suhteen, ettei tutkittava ilmiö riipu jaksollisuudesta. Jos esimerkiksi tehdään haastattelupohjaista seurantatutkimusta työpaikan henkilökunnan mielialoista, ei ole järkevää toteuttaa haastatteluja järjestelmällisesti maanantai-aamuisin klo 8 (mieliala riippuu ajankohdasta ja päivästä).

Ositettu otanta tulee kysymykseen silloin, kun perusjoukko on mielekästä jakaa ryhmiin eli *ositteisiin* jonkin kiinnostavan taustatekijän suhteen. Ositettu otanta perustuu siihen, että taustatekijä on yhteydessä tutkittavaan ominaisuuteen. Esimerkiksi käy vaikkapa tutkimus, jossa halutaan tutkia lääketieteen opiskelijoiden ajatuksia omasta tieteenalasta ja osittavana tekijänä käytetään alan opiskeluvuosien lukumäärää. Osittaminen voidaan tehdä *tasaisella kiintiöinnillä*, jossa jokaisesta ositteesta poimitaan yhtä monta alkiota ositteiden koosta riippumatta. *Suhteellisessa kiintiöinnissä* jokaisesta ositteesta poimitaan suhteellisesti se määrä alkioita kuin mitä ositteen prosenttiosuus on koko perusjoukosta.

Ryväsotantaa käytetään laajahkoissa tutkimuksissa. Jos vaikkapa halutaan tutkia kaikkien nuorten suomalaisten (alle 25 v) mielipiteitä elokuvien vuokraamisesta, voidaan otannan ensimmäisessä vaiheessa jakaa perusjoukko (suomalaiset) rypäiksi vaikkapa paikkakunnan suhteen, jolloin otantayksiköksi tulee ryväs. Valitaan tietty määrä rypäitä satunnaisotannalla jatkotutkimuksta varten ja näistä edelleen toisessa otannan vaiheessa haluttu määrä ihmisiä varsinaiseen lopulliseen otantaan. Ryväsotanta helpottaa aineiston keräämistä huomattavasti, koska ei tarvitse indeksoida kaikkia perusjoukon alkioita.

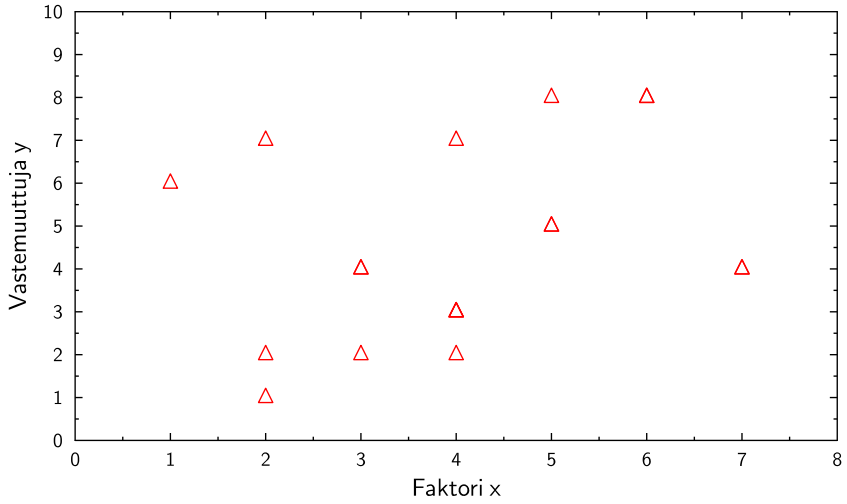
8.2 Esimerkkejä tilastollisen datan esitysmuodoista

Taulukko 15: Havaintomatriisi. Erilliset havainnot on kirjattu siihen järjestykseen, missä havainnot on tehty.

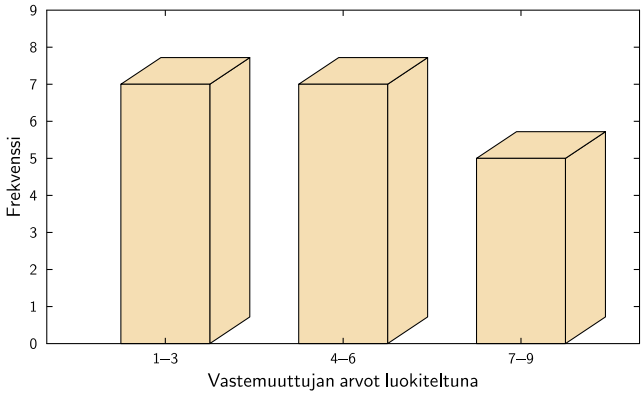
ID	x	y
1	1	6
2	2	7
3	3	4
4	4	3
5	2	2
6	3	4
7	5	5
8	6	8
9	7	4
10	4	3
11	3	2
12	2	1
13	4	2
14	5	8
15	5	5
16	6	8
17	7	4
18	4	7
19	4	3

Taulukko 16: Esimerkki frekvenssijakaumataulukosta. Muuttujien x ja y luokittelu riippuu siitä, kuinka suuri havaintoaineisto on ja minkä kokoiset luokat ovat jatkoanalyysin kannalta mielekkäitä. Usein luokkia yhdistellään tutkimuksen lopussa sopivan kokoisiksi.

Faktori x					
		1–3	4–6	7–9	Yht.
Vastemuuttuja y	1–3	3	4	0	7
	4–6	3	2	2	7
	7–10	1	4	0	5
Yht.		7	10	2	19



Kuva 8: Sirontakuvio. Vaaka-akselilla on selittävä muuttuja eli faktori x ja pystyakselilla selitettävä muuttuja eli vastemuuttuja y. Selittävän ja selitettävän muuttujan valinta riippuu aina asiayhteydestä ja selviää usein käyttämällä ”maalaisjärkeä”.



Kuva 9: Pylväsdiagrammi. Vaaka-akselilla on vastemuuttujan y arvot kolmeen luokkaan luokiteltuna ja pystyakselilla luokkia vastaavat frekvenssit (havaintoarvojen lukumäärät).

Viitteet

[1] Kärkkäinen, S. ja Högmander, H. : *Tilastomenetelmien peruskurssin luentomoniste*, 5. painos, Jyväskylän yliopistopaino 2008

[2] Valli, R. : *Johdatus tilastolliseen tutkimukseen*, PS-kustannus 2001

[3] D. S. Moore, G. P. McCabe & B. Craig: *Introduction to the practice of statistics*, 5th edition, Freeman, W. H. & Company Publishing 2007

[4] R. E. Walpole et al. : *Probability & Statistics for Engineers & Scientists*, 8th edition, Pearson Education 2007

[5] John R. Taylor: *An Introduction to Error Analysis*, 2nd edition, University Science Books 1997

[6] Hiltunen et al. (toim.): *Galenos – Johdanto lääketieteen opintoihin*, 1. painos, WSOYpro OY 2010