Stokastiikka ja tilastollinen ajattelu

Versio 0.990

Lasse Leskelä
Aalto-yliopisto

11. helmikuuta 2021

Sisältö

Alkusanat

Suurin osa meitä ympäröivistä asioista sisältää epävarmuutta. Tämä johtuu yleensä siitä, että tietomme asiaa kuvaavista muuttujista ja parametreista ovat puutteelliset tai siitä, ettemme voi varmuudella ennakoida luonnon ja muiden ihmisten käyttäytymistä. Monet tärkeät päätökset joudumme silti tekemään puutteellisen tai epävarman datan perusteella. Tällöin olemme pakotettuja tekemään arvauksia. Arvaamisen ei kuitenkaan tarvitse olla puhdasta hakuammuntaa, jos asiaan liittyvä epävarmuus on jollakin tapaa säännönmukaista. Esimerkiksi on luontevaa olettaa, että maailma huomenna näyttää jossain määrin samalta kuin tänäänkin. *Tilastotiede* on tieteenala, jonka tavoitteena on kehittää menetelmiä valistuneiden arvausten ja päätösten tekemiseen saatavilla olevan datan pohjalta. Tilastotieteessä epävarmuutta mitataan ja mallinnetaan todennäköisyyksillä. Sattuman ja todennäköisyyden lakeja käsittelevää matemaattista teoriaa kutsutaan *stokastiikaksi*. Siinä missä yksittäisen datajoukon ominaisuuksien tutkimiseen riittää työkaluiksi laskennan ja visualisoinnin tietokonealgoritmit, ovat stokastiikan matemaattiset mallit välttämättömiä silloin, kun havaitun datan pohjalta halutaan laatia ennusteita ja yleistyksiä laajempaan kontekstiin.

Tämän monisteen tavoitteena on tutustuttaa lukija tilastolliseen ajatteluta-paan sekä stokastiikan ja tilastotieteen tärkeimpiin periaatteisiin ja käsitteisiin. Alkuosassa tutustutaan todennäköisyyden laskusääntöihin ja opitaan mallintamaan satunnaisvaihtelua stokastisten mallien avulla. Monisteen toinen osa käsittelee tilastollisia menetelmiä, joiden avulla voi laatia estimaatteja ja ennusteita sekä analysoida tilastollista merkitsevyyttä havaitun datan ja prioritiedon valossa. Lisäksi tärkeänä tavoitteena antaa lukijalle mielikuva tilastollisten menetelmien mahdollisuuksista ja rajoituksista ja opettaa lukija kriittisesti arvioimaan tilastollisten menetelmien pohja-oletuksia.

Korjauksia ja parannusehdotuksia tekstiin ovat esittäneet Kalle Kytölä, Aki Vehtari, Pauliina Ilmonen, Alex Karrila, Jukka Kohonen, Georg Metsalo, Anssi Mirka, Joni Virta, Hoa Ngo ja Eric Hyyppä. Heille suuret kiitokset.

Luku 1

Todennäköisyyden käsite ja laskusäännöt

1.1 Todennäköisyyden käsite

Todennäköisyys on tapa kuvailla kvantitatiivisesti jonkin tapahtuman uskottavuutta, esimerkiksi:

- Kolikkoa heittämällä saadaan kruuna todennäköisyydellä $\frac{1}{2}$.
- Ensi maanantaina Otaniemessä sataa todennäköisyydellä 14% (Ilmatieteen laitos) tai todennäköisyydellä 19% (Foreca).

Ylläolevista ilmauksista ensimmäinen kiteyttää objektiivisen kokemuksemme kolikoista: pitkissä heittosarjoissa noin puolet heitoista tuottaa kruunan. Sen sijaan sateen todennäköisyyttä koskevat ilmaisut ovat subjektiivisia: sateen uskottavuus on Ilmatieteen laitoksen säämallien mukaan 14% ja Forecan säämallien mukaan 19%. Todennäköisyyden käsite esiintyy monenlaisissa arkielämän yhteyksissä[1] ja sen oikeaoppisesta tulkitsemisesta on ollut kiistaa eri koulukuntien kesken. Todennäköisyyden matemaattiset laskusäännöt ovat samat tulkinnasta riippumatta.

1.2 Satunnaisilmiön toteumat ja tapahtumat

Satunnaisilmiön stokastisen mallin pohjana on *perusjoukko* S, joka sisältää tarkasteltavan ilmiön mahdolliset toteumat. Satunnaisilmiön tapahtumia ovat toteumien joukot. Satunnaisilmiön *toteuma* on siis jokin perusjoukon alkio x ja *tapahtuma* jokin perusjoukon osajoukko A. Malli voidaan tulkita niin, että sattuma valitsee jonkin perusjoukon pisteen x, ja tapahtuma A toteutuu mikäli x kuuluu joukkoon A.

[1] David Aldous: Annotated list of contexts where we perceive chance
http://www.stat.berkeley.edu/~aldous/Real-World/100.html

Esimerkki 1.1 (Kolikko)**.** Kolikonheiton mahdolliset toteumat voidaan numeroida muodossa 0 = "klaava" ja 1 = "kruuna". Satunnaisilmiön perusjoukko on tällöin $S = \{0, 1\}$ ja sen tapahtumat on listattu allaolevassa taulukossa.

Tapahtuma	**Tulkinta**
$\{\}$	Mahdoton tapahtuma
$\{0\}$	Saadaan klaava
$\{1\}$	Saadaan kruuna
$\{0, 1\}$	Varma tapahtuma

■

Esimerkki 1.2 (Sademäärä)**.** Ennustettaessa ensi maanantain sademäärä (mm) Otaniemessä valitaan perusjoukoksi $S = [0, \infty)$. Satunnaisilmiön toteumia ovat ei-negatiiviset reaaliluvut ja esimerkkejä tapahtumista on taulukoitu alla.

Tapahtuma	**Tulkinta**
$(10, \infty)$	Otaniemessä sataa ensi ma yli 10 mm
$\{0\}$	Otaniemessä ei sada ensi ma

■

Arkikielessä monet tapahtumat ilmaistaan muiden tapahtumien loogisina yhdistelminä, esimerkiksi:

- "Ensimmäisellä nopalla saadaan vähintään 3 ja toisella vähintään 4."
- "Otaniemessä joko ei sada ollenkaan tai sataa vähintään 5 mm."

Koska stokastiikassa tapahtumat vastaavat joukkoja, tulee tapahtumien loogiset yhdistelmät ilmaista joukko-opin kielellä. Joukko-opissa merkitään $x \in A$ kun x kuuluu joukkoon A. Lisäksi merkitään $A \subset B$ kun A on B:n *osajoukko* eli jokainen A:n alkio kuuluu joukkoon B. Joukko-opin perusoperaatiot ja niitä havainnollistavat Venn-kaaviot ja tulkinnat on esitetty taulukossa 1.1. Lisäksi sanotaan, että

- tapahtumat $A_1, A_2, \ldots$ *poissulkevat toisensa*, jos vain yksi niistä voi toteutua, eli $A_i \cap A_j = \emptyset$ aina kun $i \neq j$,
- tapahtuman B *ositus* on kokoelma toisensa poissulkevia tapahtumia, joiden yhdiste on B.

Esimerkki 1.3 (Noppa)**.** Yhtä nopanheittoa mallintavan perusjoukon $S = \{1, 2, \ldots, 6\}$ tapahtumista

$$\begin{aligned} A &= \text{"Tulos on suurempi kuin 3"} = \{4, 5, 6\}, \\ B &= \text{"Tulos on parillinen"} = \{2, 4, 6\} \end{aligned}$$

Termi	Merkintä	Määritelmä	Venn-kaavio	Tulkinta
Perusjoukko	S	$\{x \in S : x \in S\}$		Varma tapahtuma
Osajoukko	A	$\{x \in S : x \in A\}$		A toteutuu
Osajoukko	B	$\{x \in S : x \in B\}$		B toteutuu
Leikkaus	$A \cap B$	$\{x \in S : x \in A \text{ ja } x \in B\}$		A ja B toteutuvat
Yhdiste	$A \cup B$	$\{x \in S : x \in A \text{ tai } x \in B\}$		A tai B toteutuu
Erotus	$A \setminus B$	$\{x \in S : x \in A \text{ ja } x \notin B\}$		A toteutuu mutta B ei
Erotus	$B \setminus A$	$\{x \in S : x \in B \text{ ja } x \notin A\}$		B toteutuu mutta A ei
Komplementti	A^c	$\{x \in S : x \notin A\}$		A ei toteudu
Komplementti	B^c	$\{x \in S : x \notin B\}$		B ei toteudu
Tyhjä joukko	$\emptyset$	$\{x \in S : x \notin S\}$		Mahdoton tapahtuma

Taulukko 1.1: Joukko-opin perusoperaatiot ja niiden stokastiikan tulkinnat.

muodostettuja yhdistelmiä ovat esimerkiksi

$$\begin{aligned} A \cap B &= \text{“Tulos on suurempi kuin 3 ja parillinen”} = \{4, 6\}, \\ A \cup B &= \text{“Tulos on suurempi kuin 3 tai parillinen”} = \{2, 4, 5, 6\}, \\ B \setminus A &= \text{“Tulos on parillinen ja enintään 3”} = \{2\}. \end{aligned}$$

Yksittäisiä tuloksia vastaavat tapahtumat $A_i =$ “Tulos on i” poissulkevat toisensa. Ne myös muodostavat perusjoukon S osituksen. ■

Stokastiikan mallit määritellään usein tulojoukkojen avulla. Joukkojen A ja B *tulojoukko* eli karteesinen tulo

$$A \times B = \{(x, y) : x \in A,\ y \in B\}$$

on joukko, jonka alkioita ovat joukon A ja B alkioista muodostetut järjestetyt parit. Vastaavasti määritellään joukkojen $A_1, \cdots, A_n$ tulojoukko

$$A_1 \times \cdots \times A_n = \{(x_1, \ldots, x_n) : x_1 \in A_1, \ldots, x_n \in A_n\},$$

jonka alkioita ovat joukkojen $A_1, \ldots, A_n$ alkioista muodostetut järjestetyt listat. Yhdestä joukosta A muodostettuja tulojoukkoja merkitään $A^2 = A \times A$, $A^3 = A \times A \times A$, ja niin edelleen.

Esimerkki 1.4 (Kaksi noppaa)**.** Kahden nopanheiton tulokset voidaan kirjata listaan (x, y), jossa x on ensimmäisen ja y toisen heiton tulos. Satunnaisilmiön perusjoukko on tällöin tulojoukko $S = \{1, \ldots, 6\}^2$, joka voidaan kirjoittaa muo-

dossa

$$\begin{aligned} S \;=\; \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}. \end{aligned}$$

Muutamia tapahtumia on listattu allaolevaan taulukkoon.

Tapahtuma	**Tulkinta**
$\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$	Ensimmäinen heitto $= 4$
$\{(5,1),(5,2),(5,3)\}$	Ensimmäinen heitto $= 5$ ja toinen ≤ 3
$\{(6,6)\}$	Molemmilla heitoilla saadaan 6

Kaikkien tapahtumien listaa ei ole tähän monisteeseen sisällytetty, sillä tapahtumia on $2^{36} \approx 69 \cdot 10^9$ kappaletta ja niiden taulukoiminen ylläolevalla esitystavalla vaatisi noin miljardi sivua. ■

1.3 Todennäköisyyden laskusäännöt

Perusjoukon S *todennäköisyysjakauma* eli *todennäköisyysmitta* on kuvaus, joka liittää jokaiseen tapahtumaan $A \subset S$ luvun $\mathbb{P}(A)$ ja toteuttaa ehdot:

(i) $0 \leq \mathbb{P}(A) \leq 1$.

(ii) $\mathbb{P}(S) = 1$.

(iii) Mille tahansa äärelliselle tai äärettömälle jonolle toisensa poissulkevia tapahtumia $A_1, A_2, \ldots$ pätee

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) \;=\; \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots$$

Ylläolevia ominaisuuksia kutsutaan todennäköisyyden *aksioomiksi*, koska niistä voidaan johtaa kaikki todennäköisyyden laskusäännöt. Tärkeimmät laskusäännöt on listattu alla.

Lause 1.5. *Jokainen todennäköisyysjakauma toteuttaa seuraavat laskusäännöt.*

- *Yleinen summasääntö:*

$$\mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{1.1}$$

- *Poissulkevien summasääntö:*

$$\mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B), \quad \textit{kun } A \cap B = \emptyset. \tag{1.2}$$

- *Erotuksen todennäköisyys:*

$$\mathbb{P}(B \setminus A) \;=\; \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{1.3}$$

- *Vastakohdan todennäköisyys:*

$$\mathbb{P}(A^c) \;=\; 1 - \mathbb{P}(A). \tag{1.4}$$

- *Monotonisuus:*

$$\mathbb{P}(A) \leq \mathbb{P}(B), \quad \textit{kun } A \subset B. \tag{1.5}$$

Todistus. Poissulkevien tapahtumien summasääntö (1.2) on erikoistapaus aksioomasta (iii).

Erotuksen laskusäännön todistamiseksi kirjoitetaan tapahtuma B kahden toisensa poissulkevan tapahtuman yhdisteenä $B = (A \cap B) \cup (B \setminus A)$. Tällöin kaavasta (1.2) seuraa $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A)$. Tästä kun ratkaistaan $\mathbb{P}(B \setminus A)$, saadaan kaava (1.3).

Vastakohdan laskusääntö (1.4) seuraa erotuksen laskusäännön (1.3) avulla aksioomasta (i), sillä $\mathbb{P}(A^c) = \mathbb{P}(S \setminus A) = \mathbb{P}(S) - \mathbb{P}(A \cap S) = 1 - \mathbb{P}(A)$.

Monotonisuuden (1.5) todistamiseksi todetaan ensiksi, että $A \cap B = A$ kun $A \subset B$. Tällöin erotuksen laskusäännön (1.3) avulla havaitaan, että $\mathbb{P}(A) = \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(B \setminus A) \leq \mathbb{P}(B)$.

Todistetaan viimeiseksi yleinen summasääntö. Kirjoitetaan tapahtuma $A \cup B$ yhdisteenä muodossa $A \cup B = A \cup (B \setminus A)$. Tällöin soveltamalla laskusääntöjä (1.2) ja (1.3) havaitaan, että

$$\begin{aligned} \mathbb{P}(A \cup B) \;&=\; \mathbb{P}(A) + \mathbb{P}(B \setminus A) \\ &=\; \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \end{aligned}$$

□

Esimerkki 1.6. Paneelitutkimuksen mukaan erään kaupungin aikuisväestöstä 18% seuraa Salattuja elämiä ja 11% seuraa Emmerdalea. Lisäksi todettiin, että 5% aikuisista seuraa molempia tv-sarjoja. Mikä osuus kaupungin aikuisväestöstä ei seuraa kumpaakaan tv-sarjaa?

Määritellään tapahtumat

$$\begin{aligned} A \;&=\; \text{“satunnaisesti valittu aikuinen seuraa Salattuja elämiä”}, \\ B \;&=\; \text{“satunnaisesti valittu aikuinen seuraa Emmerdalea”}. \end{aligned}$$

Tällöin $\mathbb{P}(A) = 0.18$, $\mathbb{P}(B) = 0.11$ ja $\mathbb{P}(A \cap B) = 0.05$. Yleisen summasäännön (1.1) mukaan

$$\mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \;=\; 0.24,$$

joten vastakohdan laskusäännön (1.4) mukaan

$$\mathbb{P}((A \cup B)^c) \;=\; 1 - \mathbb{P}(A \cup B) \;=\; 0.76.$$

Näin ollen 76% kaupungin aikuisväestöstä ei seuraa kumpaakaan tv-sarjaa. ■

1.4 Ehdollinen todennäköisyys

Ehdollinen todennäköisyys kertoo, miten tapahtuman todennäköisyys muuttuu, kun satunnaisilmiöstä saadaan lisätietoa.

Esimerkki 1.7 (3 kolikonheittoa)**.** Tavallista kolikkoa heitetään kolme kertaa peräkkäin. Mikä on todennäköisyys saada kolme kruunaa, kun ensimmäisen heiton tuloksen on havaittu olevan kruuna?

Kun 0 = klaava ja 1 = kruuna, voidaan kolmen heiton tulossarjoja vastaava perusjoukko kirjoittaa muodossa $S = \{000, 001, 010, 011, 100, 101, 110, 111\}$. Merkitään

$$\begin{aligned} A &= \text{“saadaan kolme kruunaa”}, \\ B &= \text{“ensimmäisellä heitolla saadaan kruuna”}. \end{aligned}$$

Ilman mitään taustatietoa kolikonheitoista ovat kaikki tulossarjat yhtä todennäköisiä, joten lähtökohtaisesti tapahtuman A todennäköisyys on $\frac{1}{8}$. Satunnaisilmiön luonne muuttuu, jos ensimmäisen heiton tiedetään olevan kruuna. Tällöin mahdolliset toteumat rajoittuvat joukon $B = \{100, 101, 110, 111\}$ alkioihin. Koska kaikki B:n toteumat ovat yhtä todennäköisiä, on tapahtuman A todennäköisyys tapahtuman B toteutuessa näin ollen $\frac{1}{4}$. ■

Tapahtuman A *ehdollinen todennäköisyys* tapahtuman B toteutuessa määritellään kaavalla

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \text{kun } \mathbb{P}(B) \neq 0. \tag{1.6}$$

Mikäli $\mathbb{P}(B) = 0$, jätetään $\mathbb{P}(A \,|\, B)$ määrittelemättä.

Esimerkki 1.8 (3 kolikonheittoa)**.** Lasketaan esimerkin 1.7 ehdollinen todennäköisyys ylläolevan yleisen määritelmän avulla. Koska tapahtuma A sisältyy tapahtumaan B, pätee $A \cap B = A$. Näin ollen

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{1/8}{1/2} = \frac{1}{4}.$$

■

Ehdollisen todennäköisyyden määritelmästä (1.6) seuraa suoraan allaoleva laskusääntö.

Lause 1.9 (Tulosääntö)**.** *Aina kun* $\mathbb{P}(A) > 0$*, pätee*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \,|\, A).$$

Esimerkki 1.10 (2 korttia)**.** Hyvin sekoitetusta korttipakasta[2] nostetaan palauttamatta kaksi korttia. Millä todennäköisyydellä molemmat ovat patoja?

[2]Tavallinen länsimainen 52 kortin pakka, jossa on 13 numeroitua korttia kutakin maata (pata ♠, risti ♣, hertta ♡, ruutu ♢), jotka on numeroitu luvuin 1,2,...,13.

Tarkasteltava tapahtuma voidaan kirjoittaa muodossa $A = A_1 \cap A_2$, jossa $A_i =$ "i:s kortti on pata". Ensimmäistä korttia nostettaessa pakan 52 kortista 13 on patoja, joten $\mathbb{P}(A_1) = \frac{13}{52}$. Kun tiedetään tapahtuman A_1 toteutuneen, on toista korttia nostettaessa pakassa jäljellä olevista 51 kortista 12 patoja. Näin ollen $\mathbb{P}(A_2 \,|\, A_1) = \frac{12}{51}$. Tulosäännön mukaan molemmat kortit ovat patoja todennäköisyydellä

$$\mathbb{P}(A) \;=\; \mathbb{P}(A_1)\mathbb{P}(A_2 \,|\, A_1) \;=\; \frac{13}{52} \cdot \frac{12}{51} \;=\; \frac{1}{17}.$$

■

1.5 Tapahtumien riippuvuus ja riippumattomuus

Kaksi satunnaisilmiöön liittyvää tapahtumaa ovat riippumattomat, jos tieto toisen toteutumisesta ei vaikuta toisen todennäköisyyteen. Matemaattisesti ilmaistuna tapahtumat A ja B ovat *riippumattomat*, jos

$$\mathbb{P}(A \cap B) \;=\; \mathbb{P}(A)\mathbb{P}(B).$$

Silloin kun A:n ja B:n todennäköisyydet ovat nollasta poikkeavia, on ylläoleva ehto yhtäpitävä yhtälöiden

$$\begin{aligned} \mathbb{P}(A \,|\, B) &= \mathbb{P}(A), \\ \mathbb{P}(B \,|\, A) &= \mathbb{P}(B), \end{aligned}$$

kanssa. Nämä voidaan tulkita niin, että tapahtuman B toteutumisesta saadusta informaatiosta ei ole hyötyä tapahtuman A ennustamiseen eikä päinvastoin. Useamman tapahtuman kokoelma on riippumaton, jos mille tahansa siitä valituille tapahtumille $A_1, \ldots, A_k$ pätee

$$\mathbb{P}(A_1 \cap \cdots \cap A_k) \;=\; \mathbb{P}(A_1) \cdots \mathbb{P}(A_k). \tag{1.7}$$

Esimerkki 1.11 (1 kortti)**.** Sekoitetusta korttipakasta nostetaan yksi kortti. Ovatko tapahtumat

$$\begin{aligned} A &= \text{“kortti on pata”} \\ B &= \text{“kortti on ässä”} \end{aligned}$$

toisistaan riippuvat vai riippumattomat?

Yksi tapa ratkaista tehtävä on tutkia laskemalla, päteekö $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Koska pakassa on täsmälleen yksi pataässä,

$$\mathbb{P}(A \cap B) \;=\; \mathbb{P}(\text{“kortti on pataässä”}) \;=\; \frac{1}{52}.$$

Koska pakassa on yhteensä 13 pataa ja 4 ässää, havaitaan että $\mathbb{P}(A) = \frac{13}{52}$ ja $\mathbb{P}(B) = \frac{4}{52}$. Näin ollen $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, joten tapahtumat A ja B ovat toisistaan riippumattomat. ■

Esimerkki 1.12 (Palvelin)**.** Palvelin on varmennettu kolmella rinnakkaisella komponentilla niin, että palvelin toimii mikäli vähintään yksi komponenteista toimii. Komponentti i toimii muista komponenteista riippumattomasti todennäköisyydellä p_i, missä $p_1 = 0.999$, $p_2 = 0.99$ ja $p_3 = 0.99$. Määritä todennäköisyys p, jolla palvelin toimii?

Tapahtuma $A =$ "palvelin toimii" voidaan esittää yhdisteenä $A = A_1 \cup A_2 \cup A_3$, jossa $A_i =$ "komponentti i toimii". Yhdistetapahtuman sijaan on helpompaa laskea sen vastakohdan todennäköisyys, sillä

$$\begin{aligned} A^c &= \text{“palvelin ei toimi”} \\ &= \text{“komponentti 1 ei toimi, komponentti 2 ei toimi, komponentti 3 ei toimi”} \\ &= A_1^c \cap A_2^c \cap A_3^c, \end{aligned}$$

ja tapahtumat A_1^c, A_2^c, A_3^c ovat toisistaan riippumattomat. Riippumattomien tapahtumien tulokaavan (1.7) mukaan

$$\mathbb{P}(A^c) \;=\; \mathbb{P}(A_1^c)\mathbb{P}(A_2^c)\mathbb{P}(A_3^c) \;=\; (1-p_1)(1-p_2)(1-p_3),$$

joten kysytty todennäköisyys on

$$p \;=\; 1-(1-p_1)(1-p_2)(1-p_3) \;=\; 0.9999999.$$

■

1.6 Osituskaava

Hyödyllinen tapa laskea tapahtumien todennäköisyyksiä on pilkkoa perusjoukko osatapahtumiin, joiden toteutuessa satunnaisilmiötä on helpompi analysoida. Perusjoukon ositus on kokoelma toisensa poissulkevia tapahtumia $A_1, \ldots, A_n$, jotka kattavat perusjoukon kaikki toteumat eli $A_1 \cup \cdots \cup A_n = S$.

Lause 1.13 (Osituskaava)**.** *Jos tapahtumat $A_1, \ldots, A_n$ muodostavat perusjoukon osituksen ja $\mathbb{P}(A_i) > 0$ kaikilla i, niin*

$$\mathbb{P}(B) \;=\; \sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B \mid A_i).$$

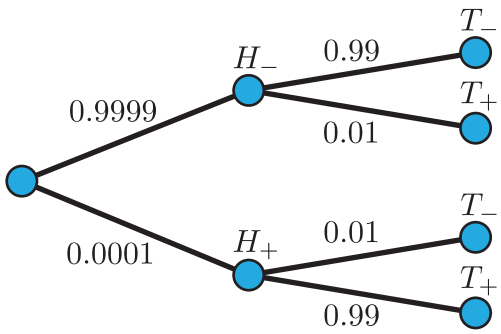
Todistus. Tapahtuman $C_i = A_i \cap B$ todennäköisyys on yleisen tulosäännön mukaan

$$\mathbb{P}(C_i) \;=\; \mathbb{P}(A_i)\mathbb{P}(B \mid A_i).$$

Lisäksi tapahtumat $C_1, \ldots, C_n$ poissulkevat toisensa ja niiden yhdiste on B. Poissulkevien tapahtumien summasäännöstä seuraa näin ollen

$$\mathbb{P}(B) \;=\; \mathbb{P}\left(C_1 \cup \cdots \cup C_n\right) \;=\; \sum_{i=1}^{n} \mathbb{P}(C_i) \;=\; \sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B \mid A_i).$$

□



Kuva 1.1: Tapahtuman T_+ todennäköisyys $\mathbb{P}(T_+) = 0.9999 \cdot 0.01 + 0.0001 \cdot 0.99$ voidaan määrittää summana T_+:lla merkittyihin solmuihin johtavien polkujen todennäköisyyksien tuloista.

Esimerkki 1.14 (Harvinainen tauti)**.** Erästä tautia esiintyy yhdellä kymmenestuhannesosalla väestöstä. Taudin toteamiseen on kehitetty kohtuullisen luotettava testi, joka tuottaa vääriä positiivisia[3] ja vääriä negatiivisia[4] todennäköisyydellä 1%. Millä todennäköisyydellä satunnaisesti valitun henkilön testitulos on positiivinen?

Merkitään

$$\begin{array}{lll} H_- & = & \text{“henkilö ei sairasta tautia”}, \\ H_+ & = & \text{“henkilö sairastaa tautia”}, \end{array} \qquad \begin{array}{lll} T_- & = & \text{“testitulos on negatiivinen”}, \\ T_+ & = & \text{“testitulos on positiivinen”}. \end{array}$$

Tällöin $\mathbb{P}(H_+) = 0.0001$, $\mathbb{P}(T_+ \,|\, H_-) = 0.01$ ja $\mathbb{P}(T_- \,|\, H_+) = 0.01$. Toistensa vastakohtina tapahtumat H_- and H_+ muodostavat perusjoukon osituksen, joten osituskaavan avulla

$$\begin{aligned} \mathbb{P}(T_+) &= \mathbb{P}(H_-)\mathbb{P}(T_+ \,|\, H_-) + \mathbb{P}(H_+)\mathbb{P}(T_+ \,|\, H_+) \\ &= 0.9999 \cdot 0.01 + 0.0001 \cdot 0.99 \\ &= 0.010098. \end{aligned}$$

Osituskaavan käyttöä voidaan havainnollistaa allaolevan kuvan 1.1 puuverkolla, jossa juurisolmusta lähteviin linkkeihin on merkitty osittavien tapahtumien H_- ja H_+ todennäköisyydet ja toisen vaiheen linkkeihin testitulosten T_- ja T_+ ehdolliset todennäköisyydet osittavien tapahtumien toteutuessa. ■

1.7 Bayesin kaava

Monissa tilanteissa tunnetaan $\mathbb{P}(A \,|\, B)$ ja halutaan määrittää käänteinen ehdollinen todennäköisyys $\mathbb{P}(B \,|\, A)$. Englannissa 1700-luvulla vaikuttaneen Thomas Bayesin nimeä kantava kuuluisa kaava soveltuu tähän.

[3] indikoi terveen ihmisen tautia sairastavaksi
[4] indikoi tautia sairastavan ihmisen terveeksi

Lause 1.15 (Bayesin kaava)**.** *Aina kun* $\mathbb{P}(A) > 0$ *ja* $\mathbb{P}(B) > 0$*, pätee*

$$\mathbb{P}(B \mid A) \;=\; \frac{\mathbb{P}(B)\mathbb{P}(A \mid B)}{\mathbb{P}(A)}.$$

Todistus. Ehdollisen todennäköisyyden määritelmästä

$$\mathbb{P}(B|A) \;=\; \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \;=\; \frac{\mathbb{P}(B)}{\mathbb{P}(A)}\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \;=\; \frac{\mathbb{P}(B)}{\mathbb{P}(A)}\mathbb{P}(A \mid B).$$

□

Esimerkki 1.16 (Harvinainen tauti)**.** Erästä tautia esiintyy yhdellä kymmenestuhannesosalla väestöstä. Taudin toteamiseen on kehitetty kohtuullisen luotettava testi, joka tuottaa vääriä positiivisia ja vääriä negatiivisia todennäköisyydellä 1%. Millä todennäköisyydellä positiivisen testituloksen saanut henkilö sairastaa tautia?

Käytetään samoja merkintöjä kuin esimerkissä 1.14, jossa positiivisen testituloksen todennäköisyydeksi saatiin $\mathbb{P}(T_+) = 0.010098$. Bayesin kaavan mukaan positiivisen testituloksen saanut henkilö sairastaa tautia todennäköisyydellä

$$\mathbb{P}(H_+ \mid T_+) \;=\; \frac{\mathbb{P}(H_+)\mathbb{P}(T_+ \mid H_+)}{\mathbb{P}(T_+)} \;=\; \frac{0.0001 \cdot 0.99}{0.010098} \;\approx\; 0.0098.$$

Näin pieni todennäköisyys vaikuttaa paradoksaaliselta, koska 99% testituloksista tiedetään olevan oikeita. Tämä on esimerkki *esiintyvyysharhasta*: vaikka kaikista testituloksista 99% on oikeita, on positiivisista testituloksista yli 99% vääriä. ■

Esimerkki 1.17 (Laadunvalvonta)**.** Samaa tuotetta valmistetaan tehtaassa kolmella eri tuotantolinjalla. Valmiit tuotteet sekoitetaan ja pakataan laatikoihin. Tuotantolinjojen suorituskykyä kuvaa allaoleva taulukko.

Linja	Tuotantomäärä	Viallisten osuus
1	3.1/min	2%
2	5.0/min	9%
3	4.5/min	8%

Satunnaisesti valitusta laatikosta poimitaan tuote tarkastettavaksi. Millä todennäköisyydellä vialliseksi havaittu tuote on linjalta 1?

Merkitään

$$\begin{aligned} L_i &= \text{“tarkastettava tuote on linjalta } i\text{”}, \\ V &= \text{“tarkastettava tuote on viallinen”}. \end{aligned}$$

Linjalta 1 peräisin oleva tuote on viallinen todennäköisyydellä $\mathbb{P}(V \mid L_1) = 0.02$. Käänteisen todennäköisyyden $\mathbb{P}(L_1 \mid V)$ määrittämiseksi Bayesin kaavalla tulee ensin laskea todennäköisyydet $\mathbb{P}(L_1)$ ja $\mathbb{P}(V)$. Ensimmäinen näistä (kolmen numeron tarkkuudella) saadaan normittamalla tuotantomäärät:

$$\mathbb{P}(L_1) \;=\; \frac{3.1}{3.1 + 5.0 + 4.5} \;=\; 0.246.$$

Vastaavasti voidaan laskea $\mathbb{P}(L_2) = 0.397$ ja $\mathbb{P}(L_3) = 0.357$. Tapahtuman V todennäköisyyden laskemiseksi sovelletaan osituskaavaa tapahtumien L_1, L_2, L_3 muodostamaan ositukseen, jolloin

$$\begin{aligned}\mathbb{P}(V) &= \mathbb{P}(L_1)\mathbb{P}(V \mid L_1) + \mathbb{P}(L_2)\mathbb{P}(V \mid L_2) + \mathbb{P}(L_3)\mathbb{P}(V \mid L_3) \\ &= 0.246 \cdot 0.02 + 0.397 \cdot 0.09 + 0.357 \cdot 0.08 \\ &= 0.0692.\end{aligned}$$

Bayesin kaavan mukaan vialliseksi havaittu tuote on peräisin linjalta 1 todennäköisyydellä

$$\mathbb{P}(L_1 \mid V) = \frac{\mathbb{P}(L_1)\mathbb{P}(V \mid L_1)}{\mathbb{P}(V)} = \frac{0.246 \cdot 0.02}{0.0692} = 0.0711.$$

■

1.8 Todennäköisyys ja kombinatoriikka

Jos äärellisen perusjoukon S jokainen toteuma on yhtä todennäköinen, saadaan tapahtuman $A \subset S$ todennäköisyys kaavasta

$$\mathbb{P}(A) = \frac{\#A}{\#S} = \frac{\text{tapahtuman } A \text{ toteumien lkm}}{\text{kaikkien toteumien lkm}}.$$

Tällöin siis todennäköisyyksien laskeminen palautuu joukkojen kokojen laskemiseksi. Suuressa perusjoukossa voi lukumäärien $\#A$ ja $\#S$ laskeminen kuitenkin olla vaikeaa, ellei jopa mahdotonta. Kombinatoriikka on tämäntyyppisiin ongelmiin keskittynyt matematiikan osa-alue.

Toimiva tapa joukon alkioiden lukumäärän laskemiseksi on laatia kuvaelma, jonka avulla joukon alkiot voidaan listata vaihe vaiheelta, ja tämän jälkeen laskea mahdollisten tapojen lukumäärä kunkin vaiheen toteuttamiseksi. Tärkeimmät kombinatoriikan perustehtävät ovat laskea:

(i) Kuinka monta tietyn pituista järjestettyä listaa voidaan valituista alkioista muodostaa, mikäli (a) toistot ovat sallittuja ja (b) toistot ovat kiellettyjä?

(ii) Kuinka monta järjestämätöntä osajoukkoa voidaan valituista alkioista muodostaa?

Näitä kysymyksiä tarkastellaan ensiksi muutamien konkreettisten esimerkkien valossa ja sen jälkeen johdetaan yleiset ratkaisukaavat.

Esimerkki 1.18 (PIN-koodit)**.** Montako nelinumeroista eri PIN-koodia voidaan muodostaa numeroista $\{0, 1, 2, \ldots, 9\}$?

Kaikkien PIN-koodien lista

0000, 0001, 0002, 0003, 0004, 0005, 0006, 0007, 0008, 0009, 0010, 0011, 0012, 0013, 0014,
0015, 0016, 0017, 0018, 0019, 0020, 0021, 0022, 0023, 0024, 0025, 0026, 0027, 0028, 0029,
0030, 0031, 0032, 0033, 0034, 0035, 0036, 0037, 0038, 0039, 0040, . . . , 9997, 9998, 9999

on liian pitkä käsin kirjoitettavaksi. Alla on mahdollinen tapa tuottaa PIN-koodi neljässä vaiheessa on.

1. Valitaan PIN-koodin ensimmäinen numero
2. Valitaan PIN-koodin seuraava numero
3. Valitaan PIN-koodin seuraava numero
4. Valitaan PIN-koodin seuraava numero

Koska jokaisen vaiheen suorittamiseen on 10 mahdollista tapaa ja jokainen vaihe voidaan suorittaa muista vaiheista riippumattomasti, on mahdollisia tapoja PIN-koodin tuottamiseksi yhteensä $10 \times 10 \times 10 \times 10 = 10\,000$. ■

Esimerkki 1.19 (Mitalisijat)**.** Monellako tapaa on mahdollista jakaa mitalisijat jääkiekon SM-liigassa pelaavien 15 joukkueen HPK, IFK, ILV, JUK, JYP, KAL, KÄR, KOO, LUK, PEL, SAI, SPO, TAP, TPS ja ÄSS kesken?

Kaikkien mitalisijakombinaatioiden lista

(HPK,IFK,ILV),	(HPK,IFK,JUK),	(HPK,IFK,JYP),	(HPK,IFK,KAL),	(HPK,IFK,KÄR),	(HPK,IFK,KOO),
(HPK,IFK,LUK),	(HPK,IFK,PEL),	(HPK,IFK,SAI),	(HPK,IFK,SPO),	(HPK,IFK,TAP),	(HPK,IFK,TPS),
(HPK,IFK,ÄSS),	(HPK,ILV,IFK),	(HPK,ILV,JUK),	(HPK,ILV,JYP),	(HPK,ILV,KAL),	(HPK,ILV,KÄR),
(HPK,ILV,KOO),	(HPK,ILV,LUK),	(HPK,ILV,PEL),	(HPK,ILV,SAI),	(HPK,ILV,SPO),	(HPK,ILV,TAP),
(HPK,ILV,TPS),	(HPK,ILV,ÄSS),	...	(ÄSS,TPS,SAI),	(ÄSS,TPS,SPO),	(ÄSS,TPS,TAP)

on selvästi liian pitkä käsin kirjoitettavaksi. Muodostetaan kaikki mitalisijakombinaatiot kolmessa vaiheessa:

1. Valitaan sijalle 1 jokin joukkue
2. Valitaan sijalle 2 jokin vielä sijoittamaton joukkue
3. Valitaan sijalle 3 jokin vielä sijoittamaton joukkue

Toisin kuin esimerkissä 1.18, mitalisijoja jaettaessa vaiheet riippuvat toisistaan niin, että sama joukkue voi sijoittua korkeintaan yhdelle mitalisijalle. Vaiheessa 1 voidaan kultamitalin saava joukkue valita 15 eri tavalla. Tämän jälkeen vaiheessa 2 voidaan hopeamitalin saajaksi valita jokin *vielä sijoittamaton* joukkue 14 eri tavalla. Vastaavasti vaiheessa 3 on jäljellä 13 eri tapaa valita pronssijoukkue. Näin ollen tapoja valita 3 joukkuetta mitalisijoille on yhteensä $15 \times 14 \times 13 = 2730$ kappaletta. ■

Seuraava tulos kiteyttää esimerkeissä 1.18 ja 1.19 tehdyt laskelmat yleiseen muotoon.

Lause 1.20 (Listojen lukumäärä)**.** *Järjestettyjä k:n alkion listoja voidaan n:n alkion joukosta muodostaa:*

- *toistojen kanssa n^k kappaletta,*
- *ilman toistoja $n(n-1)\cdots(n-k+1)$ kappaletta.*

Positiivisen kokonaisluvun *kertoma* määritellään kaavalla

$$n! \;=\; n(n-1)\cdots 2\cdot 1.$$

Sijoittamalla lauseen 1.20 jälkimmäiseen kaavaan $k = n$ havaitaan, että n:n alkion joukon kaikki alkiot voidaan järjestää listaan $n!$ tavalla.

Esimerkki 1.21 (Pelaajaviisikot)**.** Kuinka monta eri viisikkoa voidaan jääkiekkojoukkueen 20 kenttäpelaajan joukosta muodostaa?

Lauseen 1.20 mukaan $n = 20$ kenttäpelaajan joukosta voidaan muodostaa $k = 5$ eri pelaajan järjestettyjä listoja $20 \times 19 \times 18 \times 17 \times 16 = 1\ 860\ 480$ kappaletta. Tämä luku yliarvioi viisikkojen lukumäärän, sillä pelaajaviisikko on sama huolimatta siitä, missä järjestyksessä sen pelaajat listataan. Koska jokainen viisikko voidaan listata $5! = 120$ eri tavalla, on kysytty viisikkojen lukumäärä

$$\frac{20 \times 19 \times 18 \times 17 \times 16}{5!} = \frac{1\ 860\ 480}{120} = 15\ 504.$$

■

Ylläolevan esimerkin laskelma yleistyy seuraavaan muotoon.

Lause 1.22 (Osajoukkojen lukumäärä)**.** *Järjestämättömiä k:n alkion joukkoja voidaan n:n alkion joukosta muodostaa binomikertoimen*

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$$

ilmaisema lukumäärä.

Esimerkki 1.23 (Lotto)**.** Mikä on todennäköisyys saada yhdellä lottorivillä 7 oikein Veikkaus Oy:n lottoarvonnassa?

Lottoarvonnan perusjoukko on

$$S = \text{“7:n alkion osajoukot joukosta } \{1, \ldots, 40\}\text{”}$$

ja sen koko on lauseen 1.22 mukaan $\#S = \binom{40}{7}$. Tapahtuma

$$A = \text{“valitulla lottorivillä 7 oikein”}$$

sisältää täsmälleen yhden toteuman, joten $\#A = 1$. Symmetrian perusteella lottoarvonnan jokainen toteuma on yhtä todennäköinen, joten

$$\mathbb{P}(A) = \frac{\#A}{\#S} = \frac{1}{\binom{40}{7}} = \frac{1}{18\ 643\ 560}.$$

■

Esimerkki 1.24 (Johtoryhmä)**.** Yrityksen uuteen viiden hengen johtoryhmään oli hakijoina 6 miestä ja 10 naista. Jos johtoryhmä jäsenet valittaisiin arpomalla, niin millä todennäköisyydellä johtoryhmään tulisi valituksi 3 miestä ja 2 naista?

Kun arvonta tehdään täysin satunnaisesti, on jokainen arvonnan tulos yhtä todennäköinen. Perusjoukko S sisältää kaikki 5 henkilön osajoukot 16 henkilön hakijajoukosta, joten sen koko on $\#S = \binom{16}{5}$. Tapahtumaa

$$A = \text{“valitaan 3 miestä ja 2 naista”}$$

vastaavat henkilökombinaatiot voidaan muodostaa seuraavasti: valitaan ensin 3 miestä 6 miehen joukosta ja sen jälkeen 2 naista 10 naisen joukosta. Näin ollen $\#A = \binom{6}{3}\binom{10}{2}$ ja kysytty todennäköisyys on

$$\mathbb{P}(A) \;=\; \frac{\binom{6}{3}\binom{10}{2}}{\binom{16}{5}} \;=\; \frac{900}{4368} \;\approx\; 20.6\%.$$

■

Esimerkki 1.25 (Pokeri)**.** Viiden kortin vetopokerissa pelaaja saa käteensä 5 korttia sekoitetusta 52 kortin pakasta. Laske todennäköisyys saada "kolmoset" eli kolme samanarvoista korttia, esim. 4♡, 7♢, 4♣, 4♠, A♠.

Matemaattisesti "kolmoset" = viiden kortin joukko, jossa esiintyy kolme eri arvoa niin, että yksi arvo esiintyy kolmesti ja muut arvot kerran. Tällaisten joukkojen lukumäärä voidaan laskea monella eri tapaa. Yksi niistä on seuraava:

1. Valitaan kolmen arvon joukko, joita pokerikädessä esiintyy: $\binom{13}{3} = 286$ tapaa.
2. Valitaan kolmesta arvosta yksi, joka esiintyy kolmesti: $\binom{3}{1} = 3$ tapaa.
   (Muut edellisessä kohtaa valituista arvoista esiintyvät kerran.)
3. Valitaan kolmesti esiintyvälle arvolle kolmen maan joukko: $\binom{4}{3} = 4$ tapaa.
4. Valitaan pienemmälle kerran esiintyvälle arvolle maa: $\binom{4}{1} = 4$ tapaa.
5. Valitaan suuremmalle kerran esiintyvälle arvolle maa: $\binom{4}{1} = 4$ tapaa.

Kertomalla eri vaiheiden vaihtoehtojen lukumäärät saadaan "kolmosten" lukumääräksi

$$\binom{13}{3}\binom{3}{1}\binom{4}{1}\binom{4}{1}\binom{4}{1} \;=\; 54\,912.$$

Koska viiden kortin joukkoja voidaan 52 kortin pakasta poimia $\binom{52}{5} = 2\,598\,960$ eri tavalla, on kysytty todennäköisyys

$$\frac{\binom{13}{3}\binom{3}{1}\binom{4}{1}\binom{4}{1}\binom{4}{1}}{\binom{52}{5}} \;=\; \frac{54\,912}{2\,598\,960} \;\approx\; 2.11\%.$$

Samaan tapaan voidaan laskea kaikkien pokerikäsien todennäköisyydet, ks. esim. http://en.wikipedia.org/wiki/Poker_probability. ■

1.9 Kommentteja

Ylinumeroituvasti äärettömän perusjoukon kohdalla tapahtumien kokoelmaa pitää rajoittaa tiettyjen paradoksien poissulkemiseksi. Toimiva valinta on olettaa, että satunnaisilmiöön liittyvien tapahtumien kokoelma muodostaa sigma-algebran. Perusjoukon osajoukkojen kokoelma on *sigma-algebra*, jos se on numeroituvien yhdisteiden ja leikkausten sekä komplementin suhteen suljettu.

Sigma-algebran alkioita kutsutaan *mitallisiksi joukoiksi*. Syy rajoittua sigma-algebroihin on se, että näillä määritellyille todennäköisyysmitoille on mahdollista rakentaa toimiva integroinnin ja stokastisen analyysin teoria, jonka esitteli venäläismatemaatikko Andrei Kolmogorov vuonna 1933. Tästä syystä todennäköisyyden aksioomia kutsutaankin usein *Kolmogorovin aksioomiksi*. Stokastiikan yleisestä teoriasta kiinnostuneille lisätietoa löytyy oppikirjoista [Wil91, JP04] ja lähes kaikenkattavasta yleisteoksesta [Kal02].

Luku 2

Satunnaismuuttujat ja jakaumat

2.1 Satunnaismuuttujan käsite

Käytännön tilanteissa ei yleensä olla kiinnostuneita satunnaisilmiön kaikista yksityiskohdista, vaan ainostaan tietyn ilmiöön liittyvän suureen arvosta. Esimerkiksi kaupan varastonhallinnassa riittää yksittäisten myyntitapahtumien sijaan yleensä tietää päiväkohtaiset myyntimäärät. *Satunnaismuuttuja* X on suure, jonka arvo määräytyy satunnaisilmiön toteumasta. Sattuma siis määrää satunnaisilmiön toteuman $s \in S$ ja toteuma satunnaismuuttujan arvon $X(s)$. Tapahtuma "X saa arvon a" sisältää ne toteumat s, joille $X(s) = a$. Sitä merkitään

$$\{X = a\} \;=\; \{s \in S : X(s) = a\}.$$

Esimerkki 2.1 (Kaksi nopanheittoa)**.** Kahta nopanheittoa mallintavan satunnaisilmiön toteumia ovat lukuparit $s = (s_1, s_2)$, jossa s_i on heiton i tulos. Satunnaisilmiöön liittyviä satunnaismuuttujia ovat esimerkiksi

- heittotulosten summa $N(s) = s_1 + s_2$,
- heittotulosten maksimi $M(s) = \max\{s_1, s_2\}$.

■

Matemaattisesti satunnaismuuttuja on mitallinen[1] funktio $X : S \to S'$ perusjoukosta S arvojoukkoon S'. Tässä monisteessa käsitellään pääasiassa lukuarvoisia satunnaismuuttujia. Yleisemmistä satunnaismuuttujista saatetaan arvojoukon tyypin mukaan käyttää allaolevia nimityksiä:

Nimitys	**Arvojoukko**	
Satunnaisluku	$S' \subset \mathbb{R}$	
Satunnaisvektori	$S' \subset \mathbb{R}^n$	
Satunnaismatriisi	$S' \subset \mathbb{R}^{m \times n}$	
Stokastinen prosessi	$S' \subset \mathbb{R}^T$	(aikavälin T funktiot)
Satunnaiskenttä	$S' \subset \mathbb{R}^U$	(alueen U funktiot)
Satunnaisverkko	$S' \subset \{0,1\}^{V \times V}$	(solmujoukon V verkot)

[1] Mitallisuus on funktion tekninen ehto, joka sulkee pois tietyt ylinumeroituvien joukkojen väliset patologiset erikoistapaukset, ks. luku 2.8.

2.2 Jakauma ja kertymäfunktio

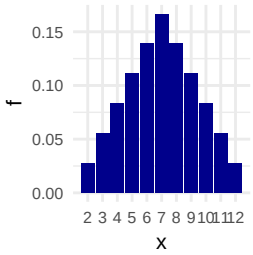
Satunnaismuuttujan X *jakauma* on taulukko tai funktio, josta voidaan määrittää X:n mahdolliset arvot ja niiden todennäköisyydet.

Esimerkki 2.2 (Kaksi nopanheittoa)**.** Kahta nopanheittoa mallinnetaan perusjoukolla $S = \{1, \ldots, 6\}^2$, jonka alkioita ovat tulosparit $s = (s_1, s_2)$. Satunnaismuuttujan $N(s) = s_1 + s_2$ arvojoukko on $\{2, \ldots, 12\}$. Tapahtumaa "N saa arvon 3" vastaa joukko

$$\{N = 3\} \;=\; \{(1,2), (2,1)\}.$$

Koska jokainen tulospari on yhtä todennäköinen, on $\mathbb{P}(N = 3) = \frac{2}{36}$. Samalla tapaa voidaan määrittää muidenkin arvojen todennäköisyydet ja satunnaismuuttujan N jakauma voidaan esittää alla olevana taulukkona.

x	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(N = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

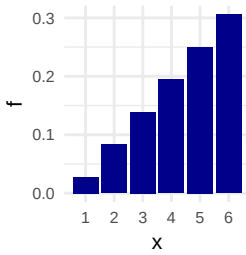


Heittotulosten maksimi on satunnaismuuttuja $M(s) = \max\{s_1, s_2\}$, jonka arvojoukko on $\{1, \ldots, 6\}$. Tapahtuma "M saa arvon 3" on joukko

$$\{M = 3\} \;=\; \{(1,3), (2,3), (3,3), (3,2), (3,1)\}.$$

Koska jokainen tulospari on yhtä todennäköinen, on $\mathbb{P}(M = 3) = \frac{5}{36}$. Vastaavaan tapaan voidaan määrittää muidenkin arvojen todennäköisyydet ja satunnaismuuttujan M jakauma voidaan esittää alla olevana taulukkona.

x	1	2	3	4	5	6
$\mathbb{P}(M = x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$



Kaikkien satunnaismuuttujien jakaumia ei voi esittää taulukkona. Tarkastellaan seuraavaa esimerkkiä.

Esimerkki 2.3 (Metron odotusaika)**.** Asemalle saapuu metroja 10 minuutin väliajoin. Asemalle saapuu matkustaja tasaisen satunnaisella ajanhetkellä. Millä todennäköisyydellä seuraavan metron odotusaika on 3 minuuttia?

Satunnaismuuttujan X mahdollisia arvoja ovat kaikki reaaliluvut jatkuvalta väliltä $[0, 10]$, kun aikayksikkönä on minuutti. Intuitiivisesti on selvää, että X:n

todennäköisyys osua lukuvälille $[a, b] \subset [0, 10]$ on kyseisen välin pituus $b - a$ jaettuna koko aikavälin pituudella 10. Näin ollen esimerkiksi

$$\mathbb{P}(2.9 \leq X \leq 3) \;=\; \frac{0.1}{10} \;=\; \frac{1}{100}.$$

Vastaavasti päätelleen havaitaan, että

$$\begin{aligned} \mathbb{P}(2.99 \leq X \leq 3) &= 0.001, \\ \mathbb{P}(2.999 \leq X \leq 3) &= 0.0001, \\ \mathbb{P}(2.9999 \leq X \leq 3) &= 0.00001. \end{aligned}$$

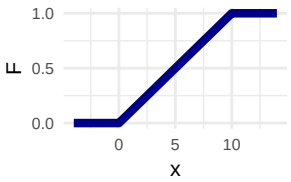
Koska tapahtuma $X = 3$ sisältyy jokaiseen ylläolevaa muotoa olevaan tapahtumaan, seuraa todennäköisyyden monotonisuuden (1.5) perusteella

$$\mathbb{P}(X = 3) \;=\; 0.$$

Tehty havainto yleistyy muotoon $\mathbb{P}(X = t) = 0$ kaikilla reaaliluvuilla t. Tämä silminnähden paradoksaalinen tulos selittyy sillä, että jatkuvan arvojoukon satunnaismuuttujalle $X = t$ tarkoittaa, että X:n arvo on t *äärettömän monen desimaalin tarkkuudella*. Odotusajan jakaumaa ei selvästikään voi esittää yksittäisten arvojen todennäköisyyksiä taulukoimalla, vaan tarvitaan jokin muu tapa. ■

Lukuarvoisen satunnaismuuttujan X *kertymäfunktio* määritellään kaavalla $F_X(t) = \mathbb{P}(X \leq t)$. Esimerkin 2.3 odotusajan kertymäfunktiolle voidaan johtaa kaava

$$F_X(t) \;=\; \begin{cases} 0, & t < 0, \\ \frac{t}{10}, & 0 \leq t \leq 10, \\ 1, & t > 1. \end{cases}$$



Kertymäfunktion avulla voi laskea tapahtumien todennäköisyyksiä hyödyntämällä todennäköisyyden yleisiä laskusääntöjä. Esimerkiksi erotuksen laskusäännön (1.3) mukaan

$$\begin{aligned} \mathbb{P}(s < X \leq t) &= \mathbb{P}(X \leq t) - \mathbb{P}(X \leq s) \\ &= F_X(t) - F_X(s). \end{aligned}$$

Vastakohdan laskusäännöstä (1.4) puolestaan seuraa

$$\mathbb{P}(X > t) \;=\; 1 - \mathbb{P}(X \leq t) \;=\; 1 - F_X(t).$$

Itse asiassa on mahdollista todistaa, että kertymäfunktio määrää lukuarvoisen satunnaismuuttujan jakauman yksikäsitteisesti. Useimmat käytännön laskut on kuitenkin hankala toteuttaa kertymäfunktion avulla. Paremman tavan tarjoavat tiheysfunktiot, joita tarkastellaan seuraavaksi.

2.3 Jakauman tiheysfunktio

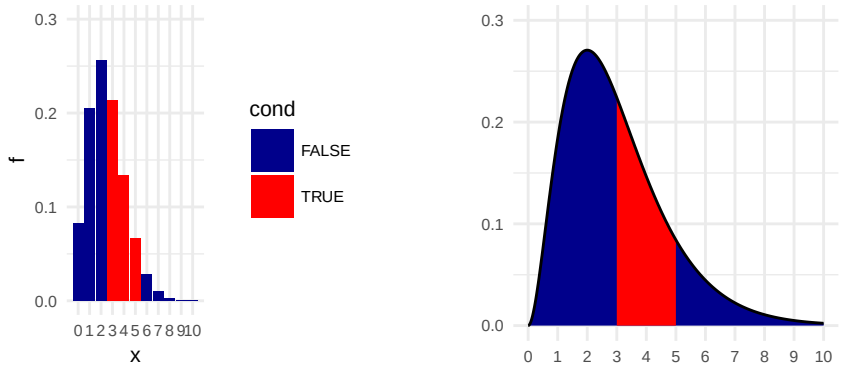
Satunnaismuuttujan X jakauma on *diskreetti*, jos sen arvojoukko on numeroituva[2] ja sen todennäköisyydet voidaan esittää funktion $f_X(x) \geq 0$ avulla muodossa

$$\mathbb{P}(X \in A) \;=\; \sum_{x \in A} f_X(x), \tag{2.1}$$

ja *jatkuva*, jos sen todennäköisyydet voidaan esittää funktion $f_X(x) \geq 0$ avulla muodossa

$$\mathbb{P}(X \in A) \;=\; \int_A f_X(x)\, dx. \tag{2.2}$$

Funktio $f_X(x)$ on X:n jakauman *tiheysfunktio*. Diskreetin satunnaismuuttujan tiheysfunktio tunnetaan myös termeillä pistemassafunktio ja (piste)todennäköisyysfunktio. Jatkuvan jakauman tiheysfunktio ei välttämättä ole jatkuva; tässä yhteydessä "jatkuva" viittaa jakauman kertymäfunktion absoluuttiseen jatkuvuuteen. Kuvassa 2.1 on esitetty todennäköisyyden laskeminen diskreetin ja jatkuvan jakauman tiheysfunktion avulla.



Kuva 2.1: Tapahtuman $3 \leq X \leq 5$ todennäköisyys lasketaan diskreetille jakaumalle punaisten pylväiden korkeuksien summana (vasen) ja jatkuvalle jakaumalle punaisen alueen pinta-alana (oikea).

Diskreetin satunnaismuuttujan tiheysfunktio voidaan aina kirjoittaa muodossa

$$f_X(x) \;=\; \mathbb{P}(X = x) \tag{2.3}$$

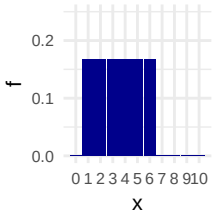
ja se toteuttaa ehdot

$$f_X(x) \geq 0 \qquad \text{ja} \qquad \sum_x f_X(x) \;=\; 1. \tag{2.4}$$

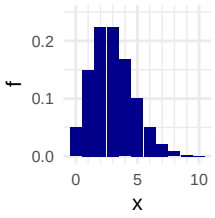
Vastaavasti mikä tahansa ehdot toteuttava (2.4) toteuttava funktio on jonkin diskreetin jakauman tiheysfunktio.

[2]Joukko on numeroituva, jos sen alkiot voidaan numeroida äärellisenä tai äärettömänä listana. Numeroituvia joukkoja: äärelliset joukot, kokonaisluvut, rationaaliluvut.

Esimerkki 2.4 (Noppa)**.** Yksittäisen nopanheiton tulos X on diskreetti satunnaismuuttuja, jonka tiheysfunktio on $f_X(x) = \frac{1}{6}$, $x \in \{1, 2, \ldots, 6\}$. Kyseinen jakauma on lukujoukon $\{1, \ldots, 6\}$ *diskreetti tasajakauma*. ■



Esimerkki 2.5 (Poisson-jakauma)**.** Lukujoukossa $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ on määritelty funktio $f(x) = e^{-3}\frac{3^x}{x!}$. Eksponenttifunktion sarjaesityksen perusteella $f(x)$ toteuttaa ehdot (2.4), joten se on erään diskreetin jakauman tiheysfunktio. Kyseinen jakauma on *Poisson-jakauma* parametrina 3. ■



Jatkuvan jakauman tiheysfunktiota *ei* voi kirjoittaa muodossa (2.3), sillä

$$\mathbb{P}(X = x) \;=\; \int_x^x f_X(t)\,dt \;=\; 0.$$

Tämä tarkoittaa sitä, että jatkuvalle satunnaismuuttujalle todennäköisyys saada arvo x äärettömän monen desimaalin tarkkuudella on nolla (vrt. esimerkki 2.3). Oikea tapa tulkita jatkuvan satunnaismuuttujan tiheysfunktio on *todennäköisyys suhteessa reaalilukujen esitystarkkuuteen*, nimittäin tiheysfunktion jatkuvuuspisteissä pätee pienillä[3] $h > 0$ arvoilla

$$f_X(x) \;\approx\; \frac{\mathbb{P}(X = x \pm h/2)}{h}, \tag{2.5}$$

missä merkintä $X = x \pm h/2$ tarkoittaa tapahtumaa $x - h/2 \le X \le x + h/2$. Jatkuvan jakauman tiheysfunktio toteuttaa ehdot

$$f_X(x) \ge 0 \qquad \text{ja} \qquad \int_{-\infty}^{\infty} f_X(x)\,dx \;=\; 1, \tag{2.6}$$

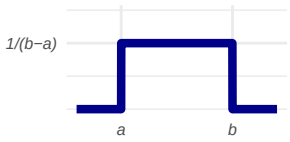
ja vastaavasti mikä tahansa ehdot (2.6) toteuttava funktio on jonkin jatkuvan jakauman tiheysfunktio. Jatkuvan jakauman kertymäfunktio määrittyy tiheysfunktiosta kaavalla

$$F_X(t) \;=\; \int_{-\infty}^{t} f_X(s)\,ds.$$

Vastaavasti $F_X'(t) = f_X(t)$ niissä pisteissä, joissa $F_X(t)$ on derivoituva.

Esimerkki 2.6. Valitaan vakiot $a < b$ ja tarkastellaan funktiota

$$f(t) \;=\; \begin{cases} \frac{1}{b-a}, & a < t < b, \\ 0, & \text{muuten.} \end{cases}$$

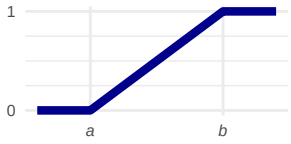


Tämä funktio toteuttaa ehdot (2.6), joten se on erään jatkuvan jakauman tiheysfunktio. Kyseinen jakauma on lukuvälin $[a, b]$ *jatkuva tasajakauma*. Sitä

[3] ao. lausekkeen "vasen puoli" $= \lim_{h \to 0}$ "oikea puoli"

vastaava kertymäfunktio saadaan integraalina

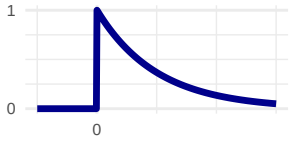
$$F(t) \;=\; \int_{-\infty}^{t} f(s)\,ds \;=\; \begin{cases} 0, & t < a, \\ \frac{t-a}{b-a}, & a \le t \le b, \\ 1, & t > b. \end{cases}$$



Sijoittamalla tähän $a = 0$ ja $b = 10$ havaitaan, että esimerkissä 2.3 tarkasteltu jakauma on välin $[0, 10]$ jatkuva tasajakauma. ■

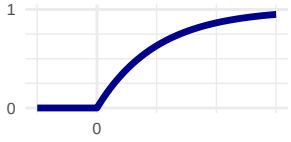
Esimerkki 2.7 (Eksponenttijakauma)**.** Valitaan vakio $\lambda > 0$ ja tarkastellaan funktiota

$$f(t) \;=\; \begin{cases} 0, & t < 0, \\ \lambda e^{-\lambda t}, & t \ge 0. \end{cases}$$



Tämä funktio toteuttaa ehdot (2.6), joten se on erään jatkuvan jakauman tiheysfunktio. Kyseinen jakauma on *eksponenttijakauma* parametrina λ.

$$F(t) \;=\; \int_{-\infty}^{t} f(s)\,ds \;=\; \begin{cases} 0, & t < 0, \\ 1 - e^{-\lambda t}, & t \ge 0. \end{cases}$$



■

2.4 Satunnaismuuttujien yhteisjakauma

Samaan satunnaisilmiöön liittyvien satunnaismuuttujien X ja Y *yhteisjakauma* on taulukko tai funktio, josta voidaan määrittää parin (X, Y) mahdolliset arvot ja niiden todennäköisyydet.

Esimerkki 2.8 (Kaksi nopanheittoa)**.** Mallinnetaan kahta nopanheittoa kuten esimerkissä 2.2 ja merkitään

$$\begin{aligned} X &= \text{“ensimmäisen heiton tulos”}, \\ Y &= \text{“toisen heiton tulos”}, \\ M &= \text{“heittotulosten maksimi”}. \end{aligned}$$

Määritä satunnaismuuttujien X ja Y yhteisjakauma. Määritä myös satunnaismuuttujien X ja M yhteisjakauma.

Parin (X, Y) mahdolliset arvot ovat tulojoukon $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$ lukuparit (x, y), jossa $x, y \in \{1, \ldots, 6\}$. Koska jokainen tulospari on yhtä todennäköinen, pätee kaikille tulojoukon lukupareille

$$\mathbb{P}(X = x,\, Y = y) \;=\; \frac{1}{36}.$$

Satunnaismuuttujien X ja Y yhteisjakauma voidaan näin ollen esittää taulukkona:

	Y					
X	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Myös parin (X, M) arvot sisältyvät tulojoukkoon $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$, mutta kaikki tulojoukon lukuparit eivät ole yhtä todennäköisiä. Esimerkiksi tapahtumaa $\{X = 3,\ M = 3\}$ vastaa perusjoukon alkiot $\{(3,1), (3,2), (3,3)\}$, joten $\mathbb{P}(X = 3,\ M = 3) = \frac{3}{36}$. Samalla tapaa kohta kohdalta päätellen voidaan todeta, että kaikille tulojoukon lukupareille (x, m) pätee

$$\mathbb{P}(X = x,\ M = m) \ = \ \begin{cases} \frac{1}{36}, & x < m, \\ \frac{x}{36}, & x = m, \\ 0, & x > m. \end{cases}$$

Satunnaismuuttujien X ja M yhteisjakauma voidaan siis esittää taulukkona:

	M					
X	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	0	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	0	0	$\frac{3}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	0	0	0	0	$\frac{5}{36}$	$\frac{1}{36}$
6	0	0	0	0	0	$\frac{6}{36}$

■

Satunnaismuuttujilla X ja Y on *diskreetti yhteisjakauma*, jos ne saavat arvoja numeroituvissa joukoissa ja niiden todennäköisyydet voidaan esittää funktion $f_{X,Y}(x, y) \geq 0$ avulla muodossa

$$\mathbb{P}\big((X, Y) \in A\big) \ = \ \sum_{(x,y)\in A} f_{X,Y}(x, y), \tag{2.7}$$

ja *jatkuva yhteisjakauma*, jos niiden todennäköisyydet voidaan esittää funktion $f_{X,Y}(x, y) \geq 0$ avulla muodossa

$$\mathbb{P}\big((X, Y) \in A\big) \ = \ \iint_A f_{X,Y}(x, y)\, dx\, dy. \tag{2.8}$$

Ylläolevissa yhtälöissä A tarkoittaa mielivaltaista[4] lukuparien joukkoa. Kaavoissa esiintyvä funktio $f_{X,Y}(x,y)$ on yhteisjakauman *tiheysfunktio*. Samanlaiset määritelmät ovat voimassa myös kolmelle ja useammalle satunnaismuuttujalle.

Diskreetin yhteisjakauman tiheysfunktio voidaan aina kirjoittaa muodossa

$$f_{X,Y}(x,y) \;=\; \mathbb{P}(X=x,\, Y=y) \tag{2.9}$$

ja se toteuttaa ehdot

$$f_{X,Y}(x,y) \geq 0 \qquad \text{ja} \qquad \sum_x \sum_y f_{X,Y}(x,y) \;=\; 1. \tag{2.10}$$

Vastaavasti mikä tahansa ehdot (2.10) toteuttava funktio on jonkin diskreetin yhteisjakauman tiheysfunktio. Satunnaismuuttujien X ja Y tiheysfunktiot saadaan yhteisjakauman tiheysfunktiosta kaavoilla

$$f_X(x) \;=\; \sum_y f_{X,Y}(x,y) \tag{2.11}$$

ja

$$f_Y(y) \;=\; \sum_x f_{X,Y}(x,y). \tag{2.12}$$

Kun diskreetti yhteisjakauma esitetään taulukkona, jonka rivejä ovat X:n arvot ja sarakkeita Y:n arvot, vastaavat $f_X(x)$:n arvot taulukon rivisummia ja $f_Y(y)$:n arvot taulukon sarakesummina. Esimerkissä 2.8 tarkasteltuja yhteisjakaumia

$$f_{X,Y}(x,y) \;=\; \frac{1}{36}, \qquad\qquad f_{X,M}(x,m) \;=\; \begin{cases} \frac{1}{36}, & x < m, \\ \frac{x}{36}, & x = m, \\ 0, & x > m, \end{cases}$$

kuvaavien taulukoiden rivi- ja sarakesummat on esitetty taulukoissa 2.1 ja 2.2.

Taulukon 2.2 rivisummat vastaavat joukon $\{1,\ldots,6\}$ tasajakaumaa eli yksittäisen nopanheiton tuloksia. Sarakesummat puolestaan vastaavat esimerkissä 2.2 johdettua kahden nopanheiton maksimin jakaumaa. Tästä syystä X:n ja Y:n jakaumia kutsutaan satunnaisvektorin (X,Y) *reunajakaumiksi* ja kaavojen (2.11) ja (2.12) määrittämiä funktioita funktion $f_{X,Y}(x,y)$ *reunatiheysfunktioiksi*.

Jatkuvan yhteisjakauman tiheysfunktiota ei voi kirjoittaa muodossa (2.9). Oikea tapa on tulkita $f_{X,Y}(x,y)$ todennäköisyytenä suhteessa reaalilukujen esitystarkkuuteen. Jatkuvan yhteisjakauman tiheysfunktion jatkuvuuspisteissä pätee lausekkeen (2.5) merkinnöin pienillä $h > 0$ arvoilla

$$f_{X,Y}(x,y) \;\approx\; \frac{\mathbb{P}(X = x \pm h/2,\, Y = y \pm h/2)}{h^2}. \tag{2.13}$$

[4] mitallista

	Y						
X	1	2	3	4	5	6	Yht
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
Yht	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Taulukko 2.1: Nopanheittojen tulosten X ja Y yhteisjakauma. Taulukon rivisummista saadaan X:n jakauma ja sarakesummista Y:n jakauma.

	M						
X	1	2	3	4	5	6	Yht
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	0	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	0	0	$\frac{3}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	0	0	0	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	0	0	0	0	$\frac{5}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{6}{36}$	$\frac{1}{6}$
Yht	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	

Taulukko 2.2: Ensimmäisen heiton X ja heittojen maksimin M yhteisjakauma. Taulukon rivisummista saadaan X:n jakauma ja sarakesummista M:n jakauma.

Jatkuvan yhteisjakauman tiheysfunktio toteuttaa ehdot

$$f_{X,Y}(x,y) \geq 0 \qquad \text{ja} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \;=\; 1, \tag{2.14}$$

ja vastaavasti jokainen ehdot toteuttava (2.6) toteuttava funktio on jonkin jatkuvan yhteisjakauman tiheysfunktio. Jatkuvaa yhteisjakaumaa noudattavien satunnaismuuttujien X ja Y jakaumat ovat jatkuvia, mutta käänteinen tulos ei yleisesti pidä paikkaansa. Satunnaismuuttujien X ja Y tiheysfunktiot saadaan yhteisjakauman tiheysfunktiosta kaavoilla

$$f_X(x) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \tag{2.15}$$

ja

$$f_Y(y) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx. \tag{2.16}$$

Myös jatkuvassa tapauksessa X:n ja Y:n jakaumia kutsutaan satunnaisvektorin (X,Y) *reunajakaumiksi* ja kaavojen (2.15) ja (2.16) määrittämiä funktioita funktion $f_{X,Y}(x,y)$ *reunatiheysfunktioiksi*.

Esimerkki 2.9 (Yksikköneliön tasajakauma)**.** Valitaan vakiot $a < b$ ja määritellään kahden muuttujan funktio

$$f_{X,Y}(x,y) \;=\; \begin{cases} \frac{1}{(b-a)^2}, & \text{kun } x \in (a,b) \text{ ja } y \in (a,b), \\ 0, & \text{muuten.} \end{cases}$$

Tämä funktio toteuttaa ehdot (2.6), joten se on joidenkin satunnaismuuttujien X ja Y yhteisjakauman tiheysfunktio. Integroimalla muuttujan y:n suhteen havaitaan, että

$$f_X(x) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \;=\; \begin{cases} \frac{1}{b-a}, & \text{kun } x \in (a,b), \\ 0, & \text{muuten.} \end{cases}$$

Vastaavasti integroimalla muuttujan x suhteen,

$$f_Y(y) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \;=\; \begin{cases} \frac{1}{b-a}, & \text{kun } y \in (a,b), \\ 0, & \text{muuten.} \end{cases}$$

Tiheysfunktiot $f_X(x)$ ja $f_Y(y)$ ovat molemmat samoja kuin esimerkissä 2.6, joten sekä X että Y noudattavat välin $[a,b]$ jatkuvaa tasajakaumaa. ■

2.5 Ehdolliset jakaumat

Satunnaismuuttujan Y *ehdollinen jakauma* tietyn tapahtuman suhteen on funktio tai taulukko, josta voidaan määrittää tapahtumien $Y \in A$ todennäköisyydet

kyseisen tapahtuman toteutuessa. Yleensä ehdollistava tapahtuma määrittyy jonkin toisen satunnaismuuttujan X kautta, jolloin ehdollisia jakaumia voi käsitellä ehdollisten tiheysfunktioiden avulla. Jos satunnaismuuttujien X ja Y diskreetillä tai jatkuvalla yhteisjakaumalla on tiheysfunktio $f_{X,Y}(x,y)$, niin satunnaismuuttujan Y *ehdollinen tiheysfunktio* satunnaismuuttujan X suhteen määritellään kaavalla

$$f_{Y|X}(y|x) \;=\; \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Kun $f_X(x) = 0$, ei ylläolevan kaavan oikea puoli ole määritelty; tällöin myös $f_{Y|X}(y|x)$ jätetään määrittelemättä. Kun $f_X(x) > 0$, havaitaan diskreetissä tapauksessa kaavan (2.11) avulla, että

$$f_{Y|X}(y|x) \geq 0 \quad \text{ja} \quad \sum_y f_{Y|X}(y|x) \;=\; 1,$$

ja jatkuvassa tapauksessa kaavan (2.15) avulla, että

$$f_{Y|X}(y|x) \geq 0 \quad \text{ja} \quad \int_{-\infty}^{\infty} f_{Y|X}(y|x)\,dy \;=\; 1.$$

Yhden muuttujan funktio $y \mapsto f_{Y|X}(y|x)$ on näin ollen jonkin jakauman tiheysfunktio. Kyseinen jakauma on satunnaismuuttujan Y *ehdollinen jakauma* tapahtuman $X = x$ suhteen. Ehdollisen jakauman tiheysfunktiolla voi laskea samaan tapaan kuin tavallisillakin tiheysfunktioilla, kun muuttujaksi valitaan y.

Diskreetissä tapauksessa havaitaan ehdollisen todennäköisyyden määritelmää sekä kaavoja (2.3) ja (2.9) käyttämällä, että

$$f_{Y|X}(y|x) \;=\; \mathbb{P}(Y = y \,|\, X = x).$$

Jatkuville jakaumille ei ylläoleva tulkinta ole mahdollinen, sillä tapahtumien $X = x$ ja $Y = y$ todennäköisyydet ovat nollia. Yhdistämällä kaavat (2.5) ja (2.13) havaitaan, että yhteisjakauman tiheysfunktion jatkuvuuspisteissä pienillä $h > 0$ arvoilla pätee

$$f_{Y|X}(y|x) \;\approx\; \frac{\mathbb{P}(Y = y \pm h/2 \;|\, X = x \pm h/2)}{h}.$$

2.6 Stokastinen riippuvuus ja riippumattomuus

Kaksi satunnaismuuttujaa ovat riippumattomat, jos informaatio toisen muuttujan arvosta ei vaikuta toisen muuttujan todennäköisyyksiin. Matemaattisesti ilmaistuna satunnaismuuttujat X ja Y ovat *riippumattomat*, jos kaikilla A ja B pätee

$$\mathbb{P}(X \in A,\, Y \in B) \;=\; \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B). \tag{2.17}$$

Silloin kun tapahtuman $X \in A$ todennäköisyys poikkeaa nollasta, voidaan ylläoleva yhtälö ilmaista myös yhtäpitävässä muodossa

$$\mathbb{P}(Y \in B \,|\, X \in A) \;=\; \mathbb{P}(Y \in B).$$

Riippumattomuus siis tarkoittaa sitä, että tieto tapahtuman $X \in A$ toteutumisesta ei vaikuta tapahtuman $Y \in B$ todennäköisyyteen, jolloin satunnaismuuttujaa X koskevista havainnoista ei ole hyötyä ennustettaessa satunnaismuuttujan Y arvoa.

Monen satunnaismuuttujan kokoelma puolestaan on riippumaton, jos mille tahansa siitä valituille satunnaismuuttujille $X_1, \ldots, X_n$ ja kaikille $A_1, \ldots, A_n$ pätee

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) \;=\; \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n). \tag{2.18}$$

Yllä mainitusta ehdosta seuraa, että X_i ja X_j ovat keskenään riippumattomat kaikilla $i \neq j$, mutta käänteinen implikaatio ei yleisesti pidä paikkaansa.

Lause 2.10. *Diskreettiä tai jatkuvaa yhteisjakaumaa noudattavat satunnaismuuttujat X ja Y ovat riippumattomat, jos ja vain jos niiden yhteisjakauman tiheysfunktio voidaan esittää muodossa*

$$f_{X,Y}(x, y) \;=\; f_X(x) f_Y(y). \tag{2.19}$$

Todistus. Todistetaan ensiksi diskreetti tapaus. (i) Jos X ja Y ovat riippumattomat, niin soveltamalla riippumattomuuden määritelmää (2.17) yhden alkion joukkoihin $A = \{x\}$ ja $B = \{y\}$ havaitaan, että

$$\begin{aligned} f_{X,Y}(x, y) &= \mathbb{P}(X = x,\, Y = y) \\ &= \mathbb{P}(X \in A,\, Y \in B) \\ &= \mathbb{P}(X \in A)\, \mathbb{P}(Y \in B) \\ &= \mathbb{P}(X = x)\, \mathbb{P}(Y = y) \\ &= f_X(x) f_Y(y). \end{aligned}$$

(ii) Oletetaan seuraavaksi, että yhteisjakauman tiheysfunktiolle on voimassa hajotelma (2.19). Koska tapahtuma "$X \in A$ ja $Y \in B$" toteutuu täsmälleen silloin kun, satunnaismuuttujien pari (X, Y) kuuluu tulojoukkoon $A \times B$,

$$\begin{aligned} \mathbb{P}(X \in A,\, Y \in B) &= \mathbb{P}((X, Y) \in A \times B) \\ &= \sum_{(x,y) \in A \times B} f_{X,Y}(x, y) \\ &= \sum_{(x,y) \in A \times B} f_X(x) f_Y(y) \\ &= \left(\sum_{x \in A} f_X(x) \right) \left(\sum_{y \in B} f_Y(y) \right) \\ &= \mathbb{P}(X \in A)\, \mathbb{P}(Y \in B). \end{aligned}$$

Koska ylläoleva yhtälö pätee kaikille A ja B, ovat X ja Y riippumattomat.

Jatkuvan yhteisjakauman tapauksessa ehdon (2.19) riittävyys voidaan perustella vaihtamalla summat integraaleiksi kohdassa (ii). Käänteisen seurausuhteen perustelemiseksi voidaan todeta, että jos X ja Y ovat riippumattomat, niin lausekkeen (2.13) merkinnöin kaikilla $h > 0$ pätee

$$\mathbb{P}(X = x \pm h/2,\, Y = y \pm h/2) \;=\; \mathbb{P}(X = x \pm h/2)\mathbb{P}(Y = y \pm h/2).$$

Jakamalla ylläolevan yhtälön molemmat puolet luvulla h^2 ja ottamalla rajaarvot kun $h \to 0$, voidaan tästä päätellä että (2.19) on voimassa funktion $f_{X,Y}(x, y)$ jatkuvuuspisteissä. Hajotelman (2.19) perustelu funktion $f_{X,Y}(x, y)$ epäjatkuvuuspisteille vaatii syvällisempää mittateorian tuntemusta ja se sivuutetaan tässä yhteydessä. □

Esimerkki 2.11 (Kaksi nopanheittoa)**.** Merkitään kahden nopanheiton tuloksia satunnaismuuttujilla X ja Y sekä tulosten maksimia satunnaismuuttujalla M. Ovatko satunnaismuuttujat X ja Y toisistaan riippuvat vai riippumattomat? Entä X ja M?

Intuitiivisesti on selvää, että X ja Y ovat toisistaan riippumattomat. Matemaattisesti tämän voi vahvistaa toteamalla, että yhtälö

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

pitää paikkansa, sillä yhteisjakauman taulukon 2.1 alkiot vastaavat rivi- ja sarakesummien alkioiden tuloja. Satunnaismuuttujat X ja M puolestaan ovat riippuvat, sillä esimerkiksi $\mathbb{P}(X = 2 \,|\, M = 1) = 0$ poikkeaa arvosta $\mathbb{P}(X = 2) = \frac{1}{6}$, joten

$$f_{X,M}(2, 1) \;\neq\; f_X(2) f_M(1).$$

Tämän voi havaita myös tarkastelemalla yhteisjakauman taulukosta 2.2 rivin 2 ja sarakkeen 1 alkioita. ■

Esimerkki 2.12 (Satunnaisotanta)**.** Korissa on 3 punaista ja 7 valkoista palloa. Korista poimitaan umpimähkään yksi pallo ja selvitetään sen väri. Sama toimenpide suoritetaan kaksi kertaa peräkkäin ja poimintojen tuloksia merkitään

$$X_1 \;=\; \begin{cases} 1, & \text{jos 1. pallo on punainen,} \\ 0, & \text{muuten,} \end{cases}$$

ja

$$X_2 \;=\; \begin{cases} 1, & \text{jos 2. pallo on punainen} \\ 0, & \text{muuten.} \end{cases}$$

Määritä satunnaismuuttujien X_1 ja X_2 yhteisjakauma.

Ylläoleva kysymys on huonosti asetettu, sillä vastaus riippuu siitä, palautetaanko poimittu pallo koriin ennen seuraavan poiminnan suorittamista. Satunnaismuuttujan X_1 todennäköisyydet ovat kuitenkin poimintatavasta huolimatta

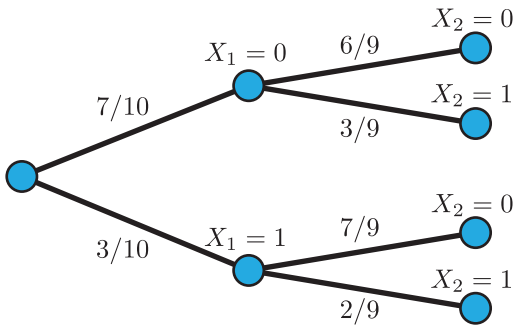
$f_{X_1}(0) = \frac{7}{10}$ ja $f_{X_1}(1) = \frac{3}{10}$. Jos poiminnat suoritetaan *palauttaen*, niin eri poimintakierrosten tulokset ovat toisistaan riippumattomat ja samoin jakautuneet, joten yhteisjakauma voidaan kirjoittaa muodossa

$$f_{X_1,X_2}(x,y) \;=\; f_{X_1}(x) f_{X_1}(y).$$

Jos taas poiminnat suoritetaan *palauttamatta*, niin tulokset X_1 ja X_2 riippuvat stokastisesti toisistaan eikä ylläolevaa kaavaa voi käyttää. Yleisen tulosäännön mukaan yhteisjakauma voidaan kuitenkin aina kirjoittaa muodossa

$$f_{X_1,X_2}(x,y) \;=\; f_{X_1}(x) f_{X_2|X_1}(y|x).$$

Riittää siis laskea ehdollisen jakauman $f_{X_2|X_1}(y|x)$ arvot. Tapahtuman $X_1 = 0$ toteutuessa korissa on toisen poimintakierroksen alussa 6 valkoista ja 3 punaista palloa, jolloin todennäköisyys saada valkoinen pallo on $\mathbb{P}(X_2 = 0 \,|\, X_1 = 0) = \frac{6}{9}$. Muut ehdolliset todennäköisyydet päätellään vastaavasti, ja ne on merkitty kuvan 2.2 puukaaviossa lehtisolmuihin johtavien linkkien yhteyteen.



Kuva 2.2: Satunnaisotanta palauttamatta. Tapahtuman $\{X_1 = 0, X_2 = 0\}$ todennäköisyydeksi voidaan kaaviosta lukea $f_{X_1,X_2}(0,0) = 7/10 \times 6/9 = 42/90$.

Satunnaismuuttujien yhteisjakaumat voidaan esittää ao. taulukkoina.

Palauttaen

	X_2		
X_1	0	1	Yht
0	$\frac{49}{100}$	$\frac{21}{100}$	$\frac{7}{10}$
1	$\frac{21}{100}$	$\frac{9}{100}$	$\frac{3}{10}$
Yht	$\frac{7}{10}$	$\frac{3}{10}$	

Palauttamatta

	X_2		
X_1	0	1	Yht
0	$\frac{42}{90}$	$\frac{21}{90}$	$\frac{7}{10}$
1	$\frac{21}{90}$	$\frac{6}{90}$	$\frac{3}{10}$
Yht	$\frac{7}{10}$	$\frac{3}{10}$	

Kummankin taulukon reunajakaumat ovat samat, mikä tarkoittaa että molemmat satunnaismuuttujat X_1 ja X_2 noudattavat jakaumaa $f_{X_i}(0) = 7/10$

ja $f_{X_i}(1) = 3/10$ poimintatavasta huolimatta. Muuttujien yhteisjakauma sen sijaan riippuu siitä, suoritetaanko poiminnat palauttaen vai palauttamatta. Tämä esimerkki osoittaa, että satunnaismuuttujien jakaumista f_{X_1} ja f_{X_2} ei voi päätellä niiden yhteisjakaumaa. ■

2.7 Yhteenveto

Alla olevassa taulukossa on tiivistelmä tämän luvun tärkeimmistä käsitteistä.

Diskreetti jakauma	**Jatkuva jakauma**
X:n arvot sisältyvät äärelliseen tai numeroituvasti äärettömään joukkoon	X:n arvot sisältyvät ylinumeroituvasti äärettömään reaalilukujen joukkoon
$\mathbb{P}(X = x) = f_X(x)$ kaikilla x	$\mathbb{P}(X = x) = 0$ kaikilla x
Jakauma määräytyy tiheysfunktiosta kaavalla $\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x)$	Jakauma määräytyy tiheysfunktiosta kaavalla $\mathbb{P}(X \in A) = \int_A f_X(x)\,dx$
Tiheysfunktion arvot ovat tarkkoja todennäköisyyksiä $f_X(x) = \mathbb{P}(X = x)$	Tiheysfunktion arvot ovat suhteellisia likiarvoisia todennäköisyyksiä $f_X(x) \approx h^{-1}\mathbb{P}(X = x \pm h/2)$
Esim. joukon $\{1, \ldots, 6\}$ tasajakauma	Esim. välin $[0, 10]$ tasajakauma

2.8 Kommentteja

Yleinen satunnaismuuttuja on *mitallinen funktio* $X : S \to S'$, mikä tarkoittaa että $B \mapsto X^{-1}(B) = \{s : X(s) \in B\}$ kuvaa maalijoukon sigma-algebran mitalliset joukot lähtöjoukon sigma-algebran mitallisiksi joukoiksi (ks. luku 1.9). Mitallisuus siis määritellään suhteessa lähtö- ja maalijoukon sigma-algebroihin, jotka yleensä oletetaan kontekstin pohjalta tunnetuiksi. Kaikki stokastiikan ja tilastotieteen sovelluksiin liittyvät funktiot ovat mitallisia, kunhan pohjalla olevat sigma-algebrat on valittu järkevästi. Niinpä mitallisuutta ei tässä monisteessa tarkastella sen tarkemmin, vaan jatkossa kaikki joukot ja funktiot oletetaan mitallisiksi ilman erillistä mainintaa.

Tämän luvun lopuksi vielä yksi esimerkkitapaus satunnaismuuttujan jakaumasta, joka ei ole diskreetti eikä jatkuva, vaan niiden sekoite.

Esimerkki 2.13. Merkitään X = satunnaisesti saapuvan matkustajan odotusaika (min) asemalla, jonne metroja saapuu tasaisin 10 min välein, ja jossa metrot pysähtyvät 1 min ajan. Määritä X:n jakauma.

Todennäköisyys että matkustaja asemalle saapuessaan näkee häntä odottavan metron on symmetrian perusteella $1/10$, ja tämän tapahtuman toteutuessa odotusaika on 0. Muussa tapauksessa odotusaika noudattaa jatkuvan välin $[0, 9]$ tasajakaumaa. Satunnaismuuttujan X jakauma ei ole diskreetti, sillä välin $[0, 9]$ lukuja ei voi numeroida listaan, eikä se ole jatkuva, sillä $\mathbb{P}(X = 0) = \frac{1}{10}$ poikkeaa nollasta. Jakauman kertymäfunktiolle voidaan kuitenkin johtaa lauseke

$$F_X(t) \;=\; \frac{1}{10} F_{X_0}(t) + \frac{9}{10} F_{X_1}(t),$$

missä

$$F_{X_0}(t) \;=\; \begin{cases} 0, & t < 0, \\ 1, & t \geq 0, \end{cases} \qquad F_{X_1}(t) \;=\; \begin{cases} 0, & t \leq 0, \\ \frac{t}{9}, & 0 < t < 9, \\ 1, & t \geq 9, \end{cases}$$

Tästä nähdään, että X:n jakauma on *diskreetin ja jatkuvan jakauman sekoitus*:

- X_0 on diskreetti satunnaismuuttuja, joka varmuudella saa arvon 0 (X:n jakauma ehdolla, että metro on odottamassa asemalla).
- X_1 on jatkuva satunnaismuuttuja, joka noudattaa välin $[0, 9]$ tasajakaumaa (X:n jakauma ehdolla, että metroa joudutaan odottamaan).

Näin ollen X:llä ei ole olemassa tiheysfunktiota tavanomaisessa mielessä. Yleistetyssä mielessä tiheysfunktion voi kuitenkin kirjoittaa viitemitan $\lambda(dx) = \delta_0(dx) + dx$ suhteen muodossa

$$f(x) \;=\; \begin{cases} \frac{1}{10}, & x = 0, \\ \frac{1}{9}, & 0 < x < 9, \\ 0, & \text{muuten}, \end{cases}$$

missä δ_0 on pisteen 0 Diracin mitta. Tällaisia yleisempiä mittoja ei tässä monisteessa käsitellä. Niistä voi lukea lisää esim. kirjoista [Kal02] tai [Wil91]. ■

Luku 3

Odotusarvo

3.1 Odotusarvon käsite ja suurten lukujen laki

Lukuarvoisen satunnaismuuttujan X *odotusarvo* määritellään tiheysfunktion $f_X(x)$ avulla diskreetille jakaumalle kaavalla[1]

$$\mathbb{E}(X) \;=\; \sum_x x\, f_X(x) \tag{3.1}$$

ja jatkuvalle jakaumalle kaavalla[2]

$$\mathbb{E}(X) \;=\; \int_{-\infty}^{\infty} x\, f_X(x)\, dx. \tag{3.2}$$

Esimerkki 3.1 (Noppa)**.** Nopanheiton jokainen tulos on yhtä todennäköinen, joten heittotulosta vastaavan diskreetin satunnaismuuttujan X tiheysfunktiolle pätee $f_X(x) = \frac{1}{6}$ kaikilla $x = 1, 2, \ldots, 6$. Näin ollen X:n odotusarvo on

$$\mathbb{E}(X) \;=\; \sum_{x=1}^{6} x\, f_X(x) \;=\; 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} \;=\; 3.5.$$

■

Esimerkki 3.2 (Jatkuva tasajakauma)**.** Välin $[a, b]$ jatkuvaa tasajakaumaa noudattavalla satunnaismuuttujalla X on tiheysfunktio

$$f_X(x) \;=\; \begin{cases} (b-a)^{-1}, & x \in [a, b], \\ 0, & \text{muuten.} \end{cases}$$

Integroimalla saadaan X:n odotusarvoksi

$$\mathbb{E}(X) \;=\; (b-a)^{-1} \int_a^b x\, dx \;=\; (b-a)^{-1} \frac{1}{2}(b^2 - a^2) \;=\; \frac{a+b}{2}.$$

■

[1] silloin kun oikean puolen summa suppenee (ks. luku 3.6)
[2] silloin kun oikean puolen integraali suppenee (ks. luku 3.6)

Odotusarvo voidaan tulkita X:n jakauman massakeskipisteenä: jos äärettömän pitkään ohueen palkkiin kohdistetaan massa $f_X(x)$ kohdassa x, niin silloin odotusarvo on se piste, johon tuettuna palkki pysyy tasapainossa. Fysikaalisen tulkinnan sijaan on kuitenkin tärkeämpää muodostaa mielikuva siitä, mitä odotusarvo kertoo satunnaismuuttujasta X. Odotusarvo *ei* tarkoita satunnaismuuttujan tyypillistä arvoa, koska esimerkiksi noppa ei milloinkaan voi saada arvoa 3.5. Tärkeän tulkinnan odotusarvon käsitteelle tarjoaa seuraava tulos, joka tunnetaan nimellä *suurten lukujen laki*[3].

Lause 3.3 (Suurten lukujen laki)**.** *Jos $X_1, X_2, \ldots$ ovat riippumattomia ja samoin jakautuneita satunnaislukuja odotusarvona μ, niin mielivaltaisen pienellä $\epsilon > 0$ tapahtuman*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \;=\; \mu \pm \epsilon \tag{3.3}$$

todennäköisyys lähestyy ykköstä suurilla n:n arvoilla[4]*.*

Todistus. Tulos seuraa erikoistapauksena lauseesta 5.8, joka esitetään luvussa 5, jossa pohjatiedoksi ensin tutustutaan keskihajonnan käsitteeseen. □

Ylläolevassa tuloksessa merkillepantavaa on se, että lausekkeen (3.3) vasen puoli on satunnaismuuttuja, mutta oikea puoli on tavallinen, ei-satunnainen luku. Keskiarvoon $\frac{1}{n}\sum_{k=1}^{n} X_k$ liittyvä satunnaisuus ja epävarmuus siis katoavat, kun summattavien määrä kasvaa suureksi. Tulosta merkitään usein

$$\frac{1}{n}\sum_{k=1}^{n} X_k \;\xrightarrow{\mathbb{P}}\; \mu$$

ja sanotaan että $\frac{1}{n}\sum_{k=1}^{n} X_k$ *suppenee stokastisesti* kohti lukua μ, kun $n \to \infty$. Suurten lukujen laki on yksi stokastiikan tärkeimmistä tuloksista, sillä siihen kiteytyy esim. rahoitus- ja vakuutusyhtiöiden toimintaperiaate: riskiä voidaan pienentää hajauttamalla varat useisiin toisistaan riippumattomiin kohteisiin. Suurten lukujen lain avulla saadaan odotusarvolle tulkinta:

> Satunnaismuuttujan odotusarvo $\mathbb{E}(X)$ on *likiarvo keskiarvolle*, joka lasketaan suuresta määrästä X:n tavoin jakautuneita riippumattomia satunnaislukuja.

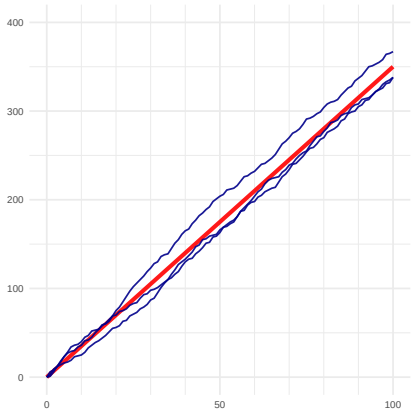
Esimerkki 3.4 (Noppapelin tuotto)**.** Noppapelissä voittaa kierroksella k silmäluvun X_k verran euroja. Yhden kierroksen tuoton odotusarvo on $\mathbb{E}(X_k) = 3.5$ EUR. Kertynyt tuotto suurelta määrältä kierroksia on suurten lukujen lain mukaan suurella todennäköisyydellä

$$\sum_{k=1}^{n} X_k \;=\; \left(\frac{1}{n}\sum_{i=k}^{n} X_k\right) n \;\approx\; 3.5n.$$

[3] suurten lukujen laista on myös vahva versio (ks. luku 3.6)

[4] Tarkemmin ilmaistuna $\lim_{n\to\infty} \mathbb{P}(|n^{-1}\sum_{k=1}^{n} X_k - \mu| \leq \epsilon) = 1$.

Kuvassa 3.1 on esitetty kolme simuloitua pelin toteumaa. Sadan pelikierroksen tuottokertymät ovat lähellä odotusarvoa 350 EUR, mutta poikkeavat siitä kuitenkin jonkun verran. ■



Kuva 3.1: Noppapelin tuottokertymän odotusarvo (punainen) ja kolme simuloitua toteumaa (sininen) pelikierrosten lukumäärän funktiona.

Mitä odotusarvo kertoo sellaisista satunnaismuuttujista, joista riippumattomia toistoja ei ole saatavilla, esim.

X = startup-yrityksen seuraavan vuoden liikevaihto,
Y = taloyhtiön materiaalivahingot ensi vuonna tapahtuvista tulipaloista.

Yksittäisen yrityksen perustajan näkökulmasta $\mathbb{E}(X)$ ei välttämättä ole erityisen tärkeä luku, mutta useaan toisistaan riippumattomasti toimivaan startup-yritykseen sijoittavan rahoittajan näkökulmasta tilanne on toinen. Vastaavasti $\mathbb{E}(Y)$ ei välttämättä ole yksittäisen taloyhtiön kannalta oleellinen luku, mutta samankaltaisia taloyhtiöitä vakuuttavan vakuutusyhtiön kannalta kylläkin.

Esimerkki 3.5 (Yksi miljoonasta)**.** Arvonnassa voittaa miljoona euroa todennäköisyydellä yksi miljoonasta. Yksittäisen arvan tuottoa kuvaavan satunnaismuuttujan X jakauma on ao. taulukossa

k	0	1000000
$\mathbb{P}(X = k)$	0.999999	0.000001

ja sen odotusarvo on

$$\mathbb{E}(X) \;=\; 0 \times 0.999999 \;+\; 1000000 \times 0.000001 \;=\; 1.$$

Tämä odotusarvo ei kerro kyseisen satunnaisilmiön luonteesta paljoakaan. Esimerkiksi jos X:n tavoin jakautuneita satunnaislukuja generoidaan toisistaan riippumattomasti, niin 10000 ensimmäistä satunnaislukua ovat kaikki nollia todennäköisyydellä $0.999999^{10000} \approx 99\%$. ■

3.2 Todennäköisyyden esiintyvyystulkinta

Kolikonheitosta puhuttaessa on tapana sanoa, että kruunan todennäköisyys on $\frac{1}{2}$. Yleensä tällä tarkoitetaan sitä, että pitkässä heittosarjassa odotetaan kruunien osuuden olevan lähellä puolikasta, mutta onko tälle intuitiolle matemaattisia takeita? Merkitään

$$I_k \;=\; \begin{cases} 1, & \text{jos heitolla } k \text{ saadaan kruuna,} \\ 0, & \text{muuten,} \end{cases}$$

jolloin n:llä heitolla saatujen kruunien lukumäärä on satunnaismuuttuja $S_n = \sum_{k=1}^{n} I_k$ ja kruunien suhteellinen esiintyvyys satunnaismuuttuja $\frac{S_n}{n}$. Kun kolikkoa heitetään tasaisen satunnaisesti, ovat satunnaismuuttujat $I_1, I_2, \ldots$ toisistaan riippumattomia ja samoin jakautuneita odotusarvonaan $\frac{1}{2}$. Näin ollen suurten lukujen lain mukaan kruunan suhteellinen esiintyvyys n heiton sarjassa toteuttaa suurella todennäköisyydellä

$$\frac{S_n}{n} \;=\; \frac{1}{2} \pm \epsilon,$$

missä $\epsilon > 0$ voidaan valita mielivaltaisen pieneksi, ja yllä olevan tapahtuman todennäköisyys on sitä lähempänä ykköstä, mitä suuremmaksi n kasvaa.

Vastaava päättely voidaan yleistää mielivaltaisille jakaumille, ja saatua tulosta kutsutaan *todennäköisyyden esiintyvyystulkinnaksi*.

Lause 3.6. *Jos $X_1, X_2, \ldots$ ovat keskenään riippumattomia X:n tavoin jakautuneita satunnaislukuja, niin mielivaltaisen arvojoukon B suhteellinen esiintyvyys tulossarjassa $(X_1, \ldots, X_n)$ toteuttaa mielivaltaisen pienelle $\epsilon > 0$*

$$\frac{\#\{k \le n : X_k \in B\}}{n} \;=\; \mathbb{P}(X \in B) \pm \epsilon$$

todennäköisyydellä, joka lähestyy ykköstä kun $n \to \infty$.

Todistus. Tulos perustellaan samalla argumentilla kuin yllä tehty kruunan suhteellisen esiintyvyyden analyysi, jossa vaihdetaan satunnaismuuttujan I_k paikalle tapahtuman $\{X_k \in B\}$ indikaattori, eli satunnaismuuttuja

$$I_k \;=\; \begin{cases} 1, & \text{jos } X_k \in B, \\ 0, & \text{muuten,} \end{cases}$$

Tällöin satunnaismuuttujat $I_1, I_2, \ldots$ ovat toisistaan riippumattomia ja samoin jakautuneita $\{0,1\}$-arvoisia satunnaismuuttujia. Koska $\mathbb{P}(I_k = 1) = \mathbb{P}(X_k \in B)$, havaitaan että

$$\mathbb{E}(I_k) \;=\; 0 \times \mathbb{P}(I_k = 0) + 1 \times \mathbb{P}(I_k = 1) \;=\; \mathbb{P}(X \in B).$$

Tulos siis seuraa soveltamalla suurten lukujen lakia (lause 3.3) satunnaismuuttujien keskiarvoon $\frac{1}{n}\sum_{k=1}^{n} I_k$. □

Esimerkki 3.7 (Kruunien lukumäärä)**.** Suurten lukujen lain perusteella kruunan suhteellinen esiintyvyys pitkässä heittosarjassa $(X_1, \ldots, X_n)$ on suurella todennäköisyydellä

$$\frac{\#\{k \leq n : X_k = \text{“kruuna”}\}}{n} \approx \frac{1}{2}.$$

Allaolevassa taulukossa on esitetty toteutunut kruunien lukumäärä, kun on simuloitu 8 kappaletta $n = 1000$ kolikonheiton sarjaa. Toteutuneet kruunien osuudet ovat kohtuullisen lähellä arvoa 0.5, mutta vaihtelevat silti jonkun verran kyseisen arvon molemmin puolin.

Simulaatio	1	2	3	4	5	6	7	8
Kruunien lkm	478	490	504	531	501	514	518	471
Kruunien osuus	0.478	0.490	0.504	0.531	0.501	0.514	0.518	0.471

3.3 Satunnaismuuttujan muunnos

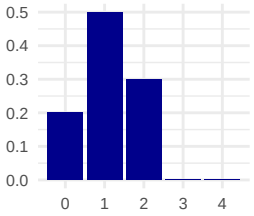
Jos X on perusjoukolla S määritelty satunnaismuuttuja ja $g(x)$ jokin X:n arvojoukolla määritelty funktio, niin

$$Y = g(X).$$

on samalla perusjoukolla S määritelty satunnaismuuttuja, joka voidaan tulkita yhdistettynä funktiona $Y(s) = g(X(s))$. Satunnaismuuttuja Y saa arvon $g(x)$ silloin kun X saa arvon x. Tarkastellaan seuraavaksi kahden esimerkin näkökulmasta, miten satunnaismuuttujan muunnoksen odotusarvon voi laskea.

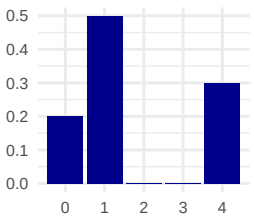
Esimerkki 3.8 (Diskreetin satunnaisluvun neliö)**.** Laske $\mathbb{E}(X^2)$, kun X:n jakauma on esitetty muodossa

k	0	1	2
$\mathbb{P}(X = k)$	0.2	0.5	0.3



Satunnaismuuttuja $Y = X^2$ on diskreetti satunnaisluku, jonka arvojoukko on $\{0, 1, 4\}$ ja jakauma on

k	0	1	4
$\mathbb{P}(Y = k)$	0.2	0.5	0.3



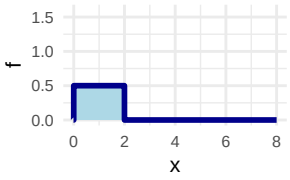
Näin ollen kysytty odotusarvo saadaan kaavasta

$$\mathbb{E}(Y) \;=\; 0 \times 0.2 \;+\; 1 \times 0.5 \;+\; 4 \times 0.3 \;=\; 1.7.$$

■

Esimerkki 3.9 (Jatkuvan satunnaisluvun kuutio)**.** Laske $\mathbb{E}(X^3)$, kun X noudattaa välin $[0, 2]$ tasajakaumaa tiheysfunktiona

$$f_X(t) \;=\; \begin{cases} \frac{1}{2}, & 0 < t < 2, \\ 0, & \text{muuten.} \end{cases}$$

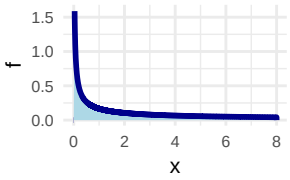


Satunnaisluvun $Y = X^3$ mahdolliset arvot sisältyvät joukkoon $[0, 8]$. Tiheysfunktion määrittämiseksi on ensiksi helpointa määrittää kertymäfunktio. Koska funktio $g(x) = x^3$ on kasvava, pätee $X^3 \leq t$ täsmälleen silloin kun $X \leq t^{1/3}$. Näin ollen kertymäfunktion arvot pisteissä $t \in [0, 8]$ saadaan laskettua kaavasta

$$F_Y(t) \;=\; \mathbb{P}(X^3 \leq t) \;=\; \mathbb{P}(X \leq t^{1/3}) \;=\; \int_0^{t^{1/3}} \frac{1}{2}\, ds \;=\; \frac{1}{2} t^{1/3}.$$

Derivoimalla havaitaan, että Y:n tiheysfunktio voidaan esittää muodossa

$$f_Y(t) \;=\; \begin{cases} \frac{1}{6} t^{-2/3}, & 0 < t < 8, \\ 0, & \text{muuten.} \end{cases}$$



Kysytty odotusarvo saadaan integroimalla

$$\mathbb{E}(Y) \;=\; \int_{-\infty}^{\infty} t\, f_Y(t) dt \;=\; \frac{1}{6} \int_0^8 t^{1/3} dt \;=\; \frac{1}{6} \left(\frac{3}{4} \times 8^{4/3} - \frac{3}{4} \times 0^{4/3} \right) \;=\; 2.$$

■

Satunnaismuuttujan X^3 odotusarvon laskeminen esimerkissä 3.9 osoittautui melko työlääksi, koska laskutehtävän yhteydessä samalla määritettiin kertymäfunktio ja tiheysfunktio. Usein ollaan kiinnostuneita pelkästään satunnaismuuttujan muunnoksen odotusarvosta, jolloin alla esitetty yleinen odotusarvon laskukaava on hyödyllinen.

Lause 3.10. *Mille tahansa satunnaismuuttujan X arvojoukolla määritellylle reaalifunktiolle g pätee diskreetin jakauman tapauksessa*

$$\mathbb{E}(g(X)) \;=\; \sum_x g(x)\, f_X(x) \tag{3.4}$$

ja jatkuvan jakauman tapauksessa

$$\mathbb{E}(g(X)) \;=\; \int_{-\infty}^{\infty} g(x)\, f_X(x) dx. \tag{3.5}$$

Todistus. Oletetaan ensiksi, että X on jakaumaltaan diskreetti, ja merkitään X:n arvojoukkoa $S_X = \{x : f_X(x) > 0\}$. Tällöin myös $Y = g(X)$ on jakaumaltaan diskreetti ja sen arvot sisältyvät joukkoon $S_Y = \{g(x) : x \in S_X\}$. Lisäksi havaitaan, että Y saa arvon y täsmälleen silloin, kun X:n arvo osuu joukkoon $B_y = \{x \in S_X : g(x) = y\}$. Näin ollen Y:n tiheysfunktio määräytyy kaavasta

$$f_Y(y) \;=\; \mathbb{P}(Y = y) \;=\; \mathbb{P}(X \in B_y) \;=\; \sum_{x \in B_y} f_X(x)$$

ja odotusarvo kaavasta

$$\mathbb{E}(Y) \;=\; \sum_{y \in S_Y} y\, f_Y(y) \;=\; \sum_{y \in S_Y} y \sum_{x \in B_y} f_X(x) \;=\; \sum_{y \in S_Y} \sum_{x \in B_y} y f_X(x).$$

Koska $g(x) = y$ kaikilla $x \in B_y$ ja koska joukot $B_y, y \in S_Y$ muodostavat joukon S_X osituksen, voidaan oikeanpuolimmainen summa kirjoittaa muodossa

$$\sum_{y \in S_Y} \sum_{x \in B_y} y f_X(x) \;=\; \sum_{y \in S_Y} \sum_{x \in B_y} g(x) f_X(x) \;=\; \sum_{x \in S_X} g(x) f_X(x) \;=\; \sum_{x} g(x) f_X(x),$$

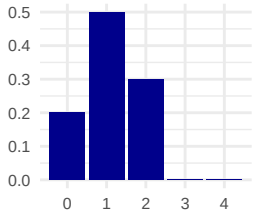
ja näin ollen yhtälö (3.4) pitää paikkansa.

Jatkuva tapaus voidaan perustella samaan tapaan. □

Seuraavaksi nähdään, miten esimerkkien 3.8 ja 3.9 odotusarvot voidaan laskea helposti odotusarvon muunnoskaavojen (3.4) ja (3.5) avulla.

Esimerkki 3.11 (Diskreetin satunnaisluvun neliö)**.** Laske $\mathbb{E}(X^2)$, kun X:n jakauma on esitetty muodossa

x	0	1	2
$f_X(x)$	0.2	0.5	0.3



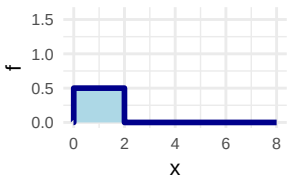
Soveltamalla odotusarvon muunnoskaavaa (3.4) funktioon $g(x) = x^2$ saadaan tulokseksi

$$\mathbb{E}(X^2) \;=\; \sum_{x=0}^{2} x^2 f_X(x) \;=\; 0^2 \times 0.2 + 1^2 \times 0.5 + 2^2 \times 0.3 \;=\; 1.7.$$

■

Esimerkki 3.12 (Jatkuvan satunnaisluvun kuutio)**.** Laske $\mathbb{E}(X^3)$, kun X noudattaa välin $[0, 2]$ tasajakaumaa tiheysfunktiona

$$f_X(t) \;=\; \begin{cases} \frac{1}{2}, & 0 < t < 2, \\ 0, & \text{muuten.} \end{cases}$$



Soveltamalla odotusarvon muunnoskaavaa (3.5) funktioon $g(x) = x^3$ saadaan tulokseksi

$$\mathbb{E}(X^3) \;=\; \int_{-\infty}^{\infty} x^3 \, f_X(x) dx \;=\; \int_0^2 x^3 \frac{1}{2}\, dx \;=\; \frac{1}{2}\left(\frac{1}{4}2^4 - \frac{1}{4}0^4\right) \;=\; 2.$$

■

Odotusarvo $\mathbb{E}(X)$ tunnetaan myös nimellä X:n *ensimmäinen momentti*. Vastaavasti luku $\mathbb{E}(X^2)$ on X:n *toinen momentti* ja samaan tapaan määritellään myös korkeamman kertaluvun momentit. Kuten yllä nähtiin, momentit voidaan laskea muunnoskaavoilla

$$\mathbb{E}(X^n) \;=\; \sum_x x^n f_X(x)$$

ja

$$\mathbb{E}(X^n) \;=\; \int_{-\infty}^{\infty} x^n f_X(x)\, dx.$$

Esimerkki 3.13 (Entropia)**.** Miten paljon informaatiota sisältyy viestiin, jossa paljastetaan että satunnaismuuttujan X arvo on x? On luontevaa ajatella, että viestissä on sitä enemmän informaatiota, mitä epätodennäköisemmästä tapahtumasta on kyse. Merkitään symbolilla $I(p)$ informaatiota, joka sisältyy viestiin tapahtumasta, jonka todennäköisyys on p. Tällöin $I(p)$:n tulee olla vähenevä p:n funktio. On myös luonteva olettaa, että jos tapahtumat $X = x$ ja $Y = y$ toteutuvat toisistaan riippumatta todennäköisyyksillä p ja q, niin tällöin näiden toteutumisen paljastava viesti sisältää informaatiota $I(p) + I(q)$ yksikköä. Koska tapahtuma $\{X = x, Y = y\}$ toteutuu todennäköisyydellä pq, on sitä vastaava informaatio $I(pq)$, joten

$$I(pq) \;=\; I(p) + I(q).$$

Voidaan näyttää, että kaikki ei-negatiiviset, vähenevät ja ylläolevan yhtälön toteuttavat funktiot ovat muotoa $I(p) = -c \log_2(p)$, missä c on positiivinen vakio. Yleensä valitaan $c = 1$, jolloin informaation yksikkönä on *bitti*.

Jos X on diskreetti satunnaismuuttuja tiheysfunktionaan $f(x)$, niin silloin viesti $\{X = x\}$ sisältää $-\log_2 f(x)$ bittiä informaatiota. Vastaavasti viesti, joka sisältää satunnaismuuttujan X tuntemattoman arvon, sisältää odotusarvoisesti

$$H(X) \;=\; -\sum_x f(x) \log_2 f(x)$$

bittiä informaatiota. Ylläoleva luku on satunnaismuuttujan X Shannonin *entropia* ja se voidaan tulkita X:n muunnoksen odotusarvona $\mathbb{E}g(X)$, missä $g(x) = -\log_2 f(x)$. Minkä tahansa n:n alkion joukossa tasajakaumaa noudattavan satunnaismuuttujan entropia on ylläolevan kaavan mukaan $H(X) = \log_2 n$. Näin ollen yhden kolikonheiton entropia on 1 bitti ja yhden nopanheiton entropia $\log_2 6 \approx 2.58$ bittiä. ■

3.4 Odotusarvon laskusääntöjä

Tärkeimmät odotusarvon laskusäännöt voidaan johtaa seuraavan tuloksen avulla, joka yleistää yhden muuttujan muunnoskaavat (lause 3.10) kahden muuttujan tapaukseen.

Lause 3.14. *Mille tahansa X:n ja Y:n arvojoukkojen tulojoukolla määritellylle reaalifunktiolle g pätee diskreetin yhteisjakauman tapauksessa*

$$\mathbb{E}(g(X,Y)) = \sum_x \sum_y g(x,y)\, f_{X,Y}(x,y) \tag{3.6}$$

ja jatkuvan yhteisjakauman tapauksessa

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\, dx\, dy. \tag{3.7}$$

Todistus. Kaava (3.6) seuraa yhden muuttujan muunnoskaavasta (3.4), kun tulkitaan pari $Z = (X,Y)$ diskreetiksi satunnaismuuttujaksi, jonka tiheysfunktio on X:n ja Y:n yhteisjakauman tiheysfunktio. Kaavan (3.7) perustelu vaatii teknisempiä mittateorian taustatietoja ja se sivuutetaan tässä yhteydessä. □

Lause 3.15. *Kaikille satunnaisluvuille X, Y ja kaikille reaaliluvuille a pätee*

$$\mathbb{E}(1) = 1, \tag{3.8}$$
$$\mathbb{E}(aX) = a\mathbb{E}(X), \tag{3.9}$$
$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y). \tag{3.10}$$

Todistus. Kaava (3.8) on selvä, kun sen vasemmalla puolella esiintyvä ykkönen tulkitaan diskreettinä satunnaismuuttujana, joka saa arvon yksi todennäköisyydellä yksi.

Oletetaan seuraavaksi, että X ja Y ovat diskreettejä, jolloin myös niiden yhteisjakauma on diskreetti. Tällöin soveltamalla odotusarvon muunnoskaavaa (3.6) funktioon $g(x,y) = ax + by$, ja yhteisjakauman reunatiheyksien laskentakaavoja (2.11)–(2.12) havaitaan, että

$$\begin{aligned}
\mathbb{E}(aX+bY) &= \sum_x \sum_y (ax+by)\, f_{X,Y}(x,y) \\
&= a\sum_x x \left(\sum_y f_{X,Y}(x,y) \right) + b \sum_y y \left(\sum_x f_{X,Y}(x,y) \right) \\
&= a\sum_x x\, f_X(x) + b\sum_y y\, f_Y(y) \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y).
\end{aligned}$$

Kaava (3.9) seuraa sijoittamalla ylläolevaan yhtälöön $b = 0$. Kaava (3.10) puolestaan seuraa sijoittamalla $a = 1$ ja $b = 1$.

Kaavojen todistaminen ei-diskreeteille satunnaismuuttujille sivuutetaan. □

3.5 Yhteenveto

Odotusarvo $\mathbb{E}(X)$ *ei* ole satunnaismuuttujan tyypillinen arvo, vaan se kertoo likiarvon keskiarvolle, joka lasketaan suuresta määrästä X:n tavoin jakautuneita riippumattomia satunnaismuuttujia. Odotusarvo on lineaarinen operaatio eli aina pätee

$$\begin{aligned} \mathbb{E}(aX) &= a\mathbb{E}(X), \\ \mathbb{E}(X+Y) &= \mathbb{E}(X)+\mathbb{E}(Y), \end{aligned}$$

huolimatta siitä ovatko X ja Y riippuvia vai riippumattomia. Diskreetin ja jatkuvan jakauman odotusarvot ja muunnosten odotusarvot voidaan laskea seuraavan taulukon mukaisilla kaavoilla.

Riippu-
mattomien
oa

Diskreetti jakauma	**Jatkuva jakauma**
Esim. joukon $\{1,\dots,6\}$ tasajakauma	Esim. välin $[0,10]$ tasajakauma
Jakauma määräytyy tiheysfunktiosta kaavalla	Jakauma määräytyy tiheysfunktiosta kaavalla
$\mathbb{P}(X\in A) = \sum_{x\in A} f_X(x)$	$\mathbb{P}(X\in A) = \int_A f_X(x)\,dx$
$\mathbb{E}(X) = \sum_x x\, f_X(x)$	$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f_X(x)dx$
$\mathbb{E}(g(X)) = \sum_x g(x)\, f_X(x)$	$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)dx$

3.6 Kommentteja

Diskreettiä jakaumaa noudattavan satunnaismuuttujan odotusarvo on olemassa silloin, kun summa $\sum_x x f_X(x)$ on hyvin määritelty, eli silloin kun vähintään toinen summista

$$\sum_x \max\{x,0\} f_X(x) \quad \text{ja} \quad \sum_x \max\{-x,0\} f_X(x)$$

on äärellinen. Vastaavasti jatkuvaa jakaumaa noudattavan satunnaismuuttujan odotusarvo on olemassa silloin, kun vähintään toinen integraaleista

$$\int_{-\infty}^{\infty} \max\{x,0\} f_X(x)\,dx \quad \text{ja} \quad \int_{-\infty}^{\infty} \max\{-x,0\} f_X(x)\,dx$$

on äärellinen. Odotusarvojen muunnoskaavoissa (lause (3.10) ja lause (3.14)) tulee myös olettaa, että odotusarvot ovat olemassa.

Käytännössä kaikilla stokastiikan sovelluksiin liittyvillä satunnaismuuttujilla on olemassa odotusarvo, ja lähes aina odotusarvo on äärellinen reaaliluku. Toisinaan kuitenkin kohdataan satunnaismuuttujia, joilla on olemassa odotusarvo, mutta odotusarvo on ääretön. Alla esimerkki tällaisesta tapauksesta.

Esimerkki 3.16 (Pietarin paradoksi)**.** Kasinolla on uhkapeli, jossa kolikkoa heitetään kunnes saadaan klaava. Pelin tuotto on

- 2 EUR, jos ensimmäinen klaava ilmestyy 1. heitolla
- 4 EUR, jos ensimmäinen klaava ilmestyy 2. heitolla
- 8 EUR, jos ensimmäinen klaava ilmestyy 3. heitolla
- ...

Paljonko olisit valmis maksamaan oikeudesta osallistua peliin?

Pelin tuotto on $g(T) = 2^T$, missä pelin kesto T on diskreetti satunnaisluku, jonka tiheysfunktio on $f_T(k) = (1/2)^k, k = 1, 2, 3, \ldots$ Pelin tuoton odotusarvo on

$$\mathbb{E}(g(T)) \;=\; 2^1 \times (1/2)^1 + 2^2 \times (1/2)^2 + 2^3 \times (1/2)^3 + \cdots \;=\; \infty.$$

Näin ollen pelistä ansaittava nettotuotto on positiivinen (ja vieläpä äärettömän suuri) huolimatta siitä, kuinka suuren summan joutuisi maksamaan oikeudesta osallistua peliin. ■

Luvussa 3.1 esitettyä suurten lukujen lakia (lause 3.3) vahvempi tulos on *vahva suurten lukujen laki*, jonka mukaan samojen oletusten vallitessa keskiarvo $n^{-1} \sum_{k=1}^{n} X_k$ lähestyy odotusarvoa μ todennäköisyydellä yksi. Suurten lukujen lakeja voidaan myös yleistää tapauksiin, joissa summattavat ovat riippuvia toisistaan. Vahvan suurten lukujen lain toteuttavaa satunnaisjonoa $X_1, X_2, \ldots$ kutsutaan *ergodiseksi*.

Luku 4

Keskihajonta ja korrelaatio

4.1 Jakauman varianssi ja keskihajonta

Edellisessä luvussa opittiin, että satunnaismuuttujan odotusarvo on X:n jakauman massakeskipiste eli X:n mahdollisten arvojen todennäköisyyksillä painotettu keskiarvo. Odotusarvo ei kuitenkaan kerro mitään jakauman *leveydestä*, kuten seuraava esimerkki vahvistaa.

Esimerkki 4.1 (Odotusarvon 1 satunnaismuuttujia)**.** Alla on kuvattu neljä diskreettiä jakaumaa, joilla kaikilla on sama odotusarvo, mutta ovat silti luonteeltaan hyvin erilaisia. Esimerkiksi X saa aina arvon 1, mutta Z ei milloinkaan. Satunnaismuuttuja W puolestaan saa lähes aina arvon 0.

x	1
$\mathbb{P}(X = x)$	1

x	0	1	2
$\mathbb{P}(Y = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

x	0	1	2
$\mathbb{P}(Z = x)$	$\frac{1}{2}$	0	$\frac{1}{2}$

x	0	1000000
$\mathbb{P}(W = x)$	0.999999	0.000001

■

Koska odotusarvo yksinään antaa varsin puutteellisen kuvan jakauman luonteesta, tarvitaan jokin toinen tunnusluku kuvaamaan jakauman leveyttä. Satunnaisluvun toteutunut poikkeama odotusarvosta $\mu = \mathbb{E}(X)$ on itsessään satunnaisluku $|X - \mu|$. Poikkeaman odotusarvo $\mathbb{E}(|X - \mu|)$ on luonteva jakauman leveyden mittari, sillä suurten lukujen lain perusteella se voidaan tulkita likiarvoksi keskiarvolle $\frac{1}{n}\sum_{k=1}^{n}|X_k - \mu|$, joka saadaan suuresta määrästä riippumattomia X:n tavoin jakautuneita satunnaislukuja. Näin määritelty jakauman itseinen keskipoikkeama on kuitenkin laskennan kannalta hankala suure, koska funktio $x \mapsto |x|$ ei ole derivoituva nollassa.

Käyttökelpoinen jakauman leveyttä kuvaava tunnusluku saadaan mittaamalla itseisen poikkeaman sijaan neliöllistä poikkeamaa. Satunnaismuuttujan X jakauman *varianssi* määritellään kaavalla

$$\mathrm{Var}(X) \;=\; \mathbb{E}[(X-\mu)^2],$$

missä $\mu = \mathbb{E}(X)$. Suurten lukujen lain (lause 3.3) avulla voidaan varianssi tulkita likiarvona keskiarvolle $\frac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2$, joka lasketaan suuresta määrästä riippumattomia X:n tavoin jakautuneita satunnaislukuja. Varianssi on optimoinnin ja numeerisen laskennan kannalta mukava tunnusluku, sillä funktiota $x \mapsto x^2$ on helppo derivoida ja se saavuttaa ainoan miniminsä nollassa. Varianssia on kuitenkin hankala tulkita intuitiivisesti, sillä se mittaa poikkeamien suuruutta neliöllisissä yksiköissä, esim. euroarvoisen osakekurssin varianssin mittayksikkö on neliöeuro. Tulkinnan helpottamiseksi varianssin ilmaisema tunnusluku normitetaan alkuperäisiin yksiköihin ottamalla neliöjuuri. Satunnaismuuttujan X jakauman *keskihajonta* (engl. standard deviation) määritellään kaavalla

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Luvun 3.3 muunnoskaavojen (3.4)–(3.5) mukaan satunnaisluvun varianssi saadaan tiheysfunktion $f_X(x)$ avulla laskettua kaavasta

$$\mathrm{Var}(X) \;=\; \begin{cases} \sum_x (x-\mu)^2 f_X(x) & \text{(diskreetti jakauma)}, \\ \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx & \text{(jatkuva jakauma)}, \end{cases} \tag{4.1}$$

ja tämän jälkeen keskihajonta saadaan ottamalla varianssin neliöjuuri.

Esimerkki 4.2 (Odotusarvon 1 satunnaismuuttujia)**.** Laske esimerkin 4.1 satunnaismuuttujien X, Y, Z, W jakaumien keskihajonnat.

Koska jokaisen esimerkin 4.1 satunnaismuuttujan odotusarvo on 1, niin soveltamalla ylläolevaa diskreetin jakauman keskihajonnan laskentakaavaa, saadaan esimerkiksi satunnaismuuttujan Y varianssiksi

$$\mathrm{Var}(Y) \;=\; (0-1)^2 \times \frac{1}{3} + (1-1)^2 \times \frac{1}{3} + (2-1)^2 \times \frac{1}{3} \;=\; \frac{2}{3},$$

joten keskihajonta on $\mathrm{SD}(Y) = \sqrt{\frac{2}{3}} \approx 0.8165$. Muut keskihajonnat voidaan laskea samaan tapaan, jolloin saadaan ao. taulukon mukaiset tulokset (neljän desimaalin tarkkuudella):

Satunnaismuuttuja	Odotusarvo	Keskihajonta
X	1	0.0000
Y	1	0.8165
Z	1	1.0000
W	1	999.9995

■

Seuraava tulos helpottaa varianssin, ja sitä kautta myös keskihajonnan, laskemista käytännön tilanteissa.

Lause 4.3. *Satunnaismuuttujan X jakauman varianssi voidaan laskea kaavasta*

$$\mathrm{Var}(X) \;=\; \mathbb{E}(X^2) - \big(\mathbb{E}X\big)^2. \tag{4.2}$$

Todistus. Merkitsemällä $\mu = \mathbb{E}(X)$, kirjoittamalla poikkeaman neliö muodossa $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$ ja soveltamalla odotusarvon lineaarisuutta havaitaan, että

$$\begin{aligned}
\mathbb{E}[(X-\mu)^2] &= \mathbb{E}[X^2 - 2\mu X + \mu^2] \\
&= \mathbb{E}[X^2] - \mathbb{E}[2\mu X] + \mathbb{E}[\mu^2] \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2\mathbb{E}[1] \\
&= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}[X^2] - \big(\mathbb{E}X\big)^2.
\end{aligned}$$

□

Esimerkki 4.4 (Metron odotusaika)**.** Jos seuraavan metron saapumiseen kuluva aika noudattaa välin $[0, 10]$ tasajakaumaa, niin saapumisajan odotusarvo on $\mu = 5$. Laske keskihajonta.

Merkitään odotusaikaa satunnaismuuttujana X, jolloin odotusajan toinen momentti on

$$\mathbb{E}(X^2) \;=\; \int_{-\infty}^{\infty} x^2 f(x)dx \;=\; \int_0^{10} x^2\,\frac{1}{10}dx \;=\; \frac{1}{10}\left(\frac{1}{3}10^3 - \frac{1}{3}0^3\right) \;=\; \frac{100}{3}.$$

Näin ollen kaavan (4.2) mukaan

$$\mathrm{Var}(X) \;=\; \mathbb{E}(X^2) - \mu^2 \;=\; \frac{100}{3} - 25 \;=\; \frac{25}{3},$$

joten $\mathrm{SD}(X) = \sqrt{\frac{25}{3}} \approx 2.89$. ■

Seuraava tulos kertoo, että keskihajonta on jakauman tunnuslukuna siirtoinvariantti: se ei muutu jos satunnaismuuttujaan lisätään tai vähennetään jokin vakio. Lisäksi se kertoo, että keskihajonta on nolla sellaisille satunnaismuuttujille, joissa ei ole lainkaan satunnaisvaihtelua. On myös mahdollista osoittaa, että jokaisen satunnaismuuttujan, jonka keskihajonta on nolla, täytyy olla vakio todennäköisyydellä yksi.

Lause 4.5. *Kaikilla reaaliluvuilla a ja b pätee*

$$\mathrm{SD}(1) \;=\; 0, \tag{4.3}$$
$$\mathrm{SD}(aX) \;=\; |a|\,\mathrm{SD}(X), \tag{4.4}$$
$$\mathrm{SD}(aX+b) \;=\; |a|\,\mathrm{SD}(X). \tag{4.5}$$

Todistus. Normitetulle ja siirretylle satunnaismuuttujalle $Y = aX + b$ pätee odotusarvon lineaarisuuden nojalla $\mathbb{E}(Y) = a\mu + b$, missä $\mu = \mathbb{E}(X)$. Näin ollen Y:n neliöity poikkeama odotusarvostaan voidaan kirjoittaa muodossa

$$(Y - \mathbb{E}(Y))^2 \;=\; (aX + b - (a\mu + b))^2 \;=\; a^2(X - \mu)^2.$$

Näin ollen

$$\mathbb{E}\Big((Y - \mathbb{E}(Y))^2\Big) \;=\; \mathbb{E}\Big(a^2(X - \mu)^2\Big) \;=\; a^2\mathbb{E}\Big((X - \mu)^2\Big),$$

mikä voidaan kirjoittaa myös muodossa

$$\operatorname{Var}(Y) \;=\; a^2 \operatorname{Var}(X).$$

Kaava (4.5) seuraa tästä ottamalla molemmilta puolilta neliöjuuret.

Kaava (4.3) saadaan sijoittamalla $a = 0$ ja $b = 1$ kaavaan (4.5). Kaava (4.4) saadaan vastaavasti sijoittamalla $b = 0$. □

4.2 Keskihajonta ja satunnaisvaihtelu

Seuraava venäläismatemaatikko Pafnuty Chebyshevin (1821–1894) mukaan nimetty epäyhtälö kiteyttää, mitä jakauman keskihajonta kertoo satunnaismuuttujan arvoista.

Lause 4.6 (Chebyshevin epäyhtälö)**.** *Satunnaismuuttujan X arvo osuu kahden keskihajonnan σ sisälle odotusarvostaan μ vähintään todennäköisyydellä $\frac{3}{4}$, ja yleisemmin*

$$\mathbb{P}(X = \mu \pm k\sigma) \;\geq\; 1 - \frac{1}{k^2} \qquad \textit{kaikilla } k \geq 1.$$

Todistus. Tehdään todistus ensiksi jatkuvalle jakaumalle, jolla on tiheysfunktio $f(x)$. Koska $(x - \mu)^2 > k^2\sigma^2$ joukossa $A = \{x : |x - \mu| > k\sigma\}$, saadaan kaavan (4.1) avulla jakauman varianssille alaraja

$$\begin{aligned}\sigma^2 \;=\; \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\,dx \;&\geq\; \int_A (x - \mu)^2 f_X(x)\,dx \\ &\geq\; \int_A k^2\sigma^2 f_X(x)\,dx \;=\; k^2\sigma^2\mathbb{P}(X \in A).\end{aligned}$$

Jakamalla ylläolevat termit luvulla $k^2\sigma^2$, havaitaan että $\mathbb{P}(X \in A) \;\leq\; \frac{1}{k^2}$, joten

$$\mathbb{P}(X = \mu \pm k\sigma) \;=\; 1 - \mathbb{P}(X \in A) \;\geq\; 1 - \frac{1}{k^2}.$$

Erityisesti yo. epäyhtälöstä seuraa, että X osuu $k = 2$ keskihajonnan sisälle odotusarvostaan vähintään todennäköisyydellä $\frac{3}{4}$. Diskreetin jakauman tapaus saadaan vaihtamalla ylläolevassa argumentissa integraalit summiksi. □

Esimerkki 4.7 (Tiedostopalvelin)**.** Palvelimelta ladatun satunnaisen tiedoston koko on odotusarvoltaan 1000 kB ja keskihajonnaltaan 200 kB. Onko todennäköistä vai epätodennäköistä, että satunnaisen tiedoston koko osuu (a) välille 600–1400 kB, (b) välille 800–1200 kB?

Chebyshevin epäyhtälön mukaan

$$\mathbb{P}(X \in [600, 1400]) \;=\; \mathbb{P}(X = \mu \pm 2\sigma) \;\geq\; \frac{3}{4},$$

joten tiedoston koko osuu välille 600–1400 kB melko suurella (yli 75%) todennäköisyydellä. Jälkimmäiseen (b)-kohdan kysymykseen on pelkän odotusarvon ja keskihajonnan pohjalta vaikea sanoa mitään, sillä tässä yhteydessä Chebyshevin epäyhtälö kertoo vain itsestään selvän tosiasian

$$\mathbb{P}(X \in [800, 1200]) \;=\; \mathbb{P}(X = \mu \pm \sigma) \;\geq\; 1 - \frac{1}{1^2} \;=\; 0.$$

Allaolevilla diskreeteillä jakaumilla on kaikilla sama odotusarvo $\mu = 1000$ ja sama keskihajonta $\sigma = 200$.

x	799	1000	1201
$f_1(x)$	0.495	0.01	0.495

x	599	1000	1401
$f_2(x)$	0.124	0.752	0.124

x	0	1000	5000
$f_3(x)$	0.008	0.990	0.002

Näiden jakaumien tapauksessa voidaan satunnaisen tiedoston koon todennäköisyys osua välille 800–1200 kB ja 600–1400 kB suoraan lukea kyseisen jakauman taulukosta. Saadut arvot (yhden prosenttiyksikön tarkkuudella) on tiivistetty allaolevaan taulukkoon:

Jakauma	$\mathbb{P}(800 \leq X \leq 1200)$	$\mathbb{P}(600 \leq X \leq 1400)$
$f_1(x)$	1%	100%
$f_2(x)$	75%	75%
$f_3(x)$	99%	99%

Taulukosta nähdään, että tiedoston koon todennäköisyys osua välille yhden keskihajonnan sisälle odotusarvosta voi olla hyvin pieni (noin 1%) tai hyvin suuri (noin 99%). Todennäköisyys osua välille *kahden* keskihajonnan sisälle odotusarvosta on kuitenkin kaikissa tapauksissa vähintään 75%, kuten Chebyshevin epäyhtälö vahvistaa. ■

4.3 Yhteisjakauman kovarianssi ja korrelaatio

Keskihajonta on jakauman tunnusluku, joka mittaa yhden muuttujan satunnaisvaihtelua, mutta ei kerro mitään useamman satunnaismuuttujan yhteisvaihtelusta. Yhteisvaihtelun suuntaa ja voimakkuutta voidaan kuvata yhteisjakauman kovarianssilla ja korrelaatiolla. Satunnaismuuttujien X ja Y yhteisjakauman *kovarianssi* määritellään kaavalla

$$\mathrm{Cov}(X, Y) \;=\; \mathbb{E}\Big[(X - \mu_X)(Y - \mu_Y)\Big], \tag{4.6}$$

jossa $\mu_X = \mathbb{E}[X]$ ja $\mu_Y = \mathbb{E}[Y]$. Varianssin tavoin myös kovarianssi mittaa vaihtelua neliöidyissä yksiköissä. Kovarianssia ei normiteta ottamalla neliöjuurta, sillä toisin kuin varianssi, kovarianssi voi olla negatiivinen. Luonteva tapa normittaa kovarianssi on jakaa se X:n ja Y:n keskihajonnoilla. Näin saatu yksikötön tunnusluku

$$\mathrm{Cor}(X,Y) \;=\; \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \tag{4.7}$$

on X:n ja Y:n yhteisjakauman *korrelaatio*. Se mittaa X:n ja Y:n yhteisvaihtelun suuntaa ja voimakkuutta normitetuissa yksiköissä.

Luvun 3 muunnoskaavojen (3.6)–(3.7) mukaan kovarianssi saadaan diskreetin yhteisjakauman tapauksessa laskettua kaavasta

$$\mathrm{Cov}(X,Y) \;=\; \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y) \tag{4.8}$$

ja jatkuvan yhteisjakauman tapauksessa kaavasta

$$\mathrm{Cov}(X,Y) \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)\,dx\,dy. \tag{4.9}$$

Seuraava laskukaava helpottaa kovarianssin laskemista käytännön tilanteissa.

Lause 4.8. *Kovarianssi voidaan laskea kaavasta*

$$\mathrm{Cov}(X,Y) \;=\; \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{4.10}$$

Todistus. Kirjoittamalla $(X-\mu_X)(Y-\mu_Y) = XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y$ ja soveltamalla odotusarvon lineaarisuutta havaitaan, että

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{E}[XY] - \mu_X\mathbb{E}[Y] - \mu_Y\mathbb{E}[X] + \mu_X\mu_Y \\
&= \mathbb{E}[XY] - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= \mathbb{E}[XY] - \mu_X\mu_Y.
\end{aligned}$$

□

Kovarianssi tarvitaan satunnaismuuttujien summien ja lineaarikombinaatioiden keskihajontojen laskemiseen. Laskemista helpottavat kovarianssin hyvät algebralliset ominaisuudet, jotka on listattu alla.

Lause 4.9. *Kovarianssille pätee*

$$\mathrm{Cov}(Y,X) \;=\; \mathrm{Cov}(X,Y), \tag{4.11}$$

$$\mathrm{Cov}(aX,Y) \;=\; a\,\mathrm{Cov}(X,Y), \tag{4.12}$$

$$\mathrm{Cov}(X+Y,Z) \;=\; \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z), \tag{4.13}$$

ja yleisesti

$$\mathrm{Cov}\Big(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\Big) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j\,\mathrm{Cov}(X_i,Y_j). \tag{4.14}$$

Todistus. Symmetrisyys (4.11) seuraa suoraan kovarianssin määritelmästä (4.6).

Kovarianssin laskentakaavan (4.10) ja odotusarvon lineaarisuuden perusteella puolestaan havaitaan, että

$$\begin{aligned}\mathrm{Cov}(aX+bY,Z) &= \mathbb{E}[(aX+bY)Z]-\mathbb{E}[aX+bY]\mathbb{E}[Z]\\ &= a\mathbb{E}[XZ]+b\mathbb{E}[YZ]-(a\mathbb{E}[X]+b\mathbb{E}[Y])\mathbb{E}[Z]\\ &= a\mathbb{E}[XZ]-a\mathbb{E}[X]\mathbb{E}[Z]+b\mathbb{E}[YZ]-b\mathbb{E}[Y]\mathbb{E}[Z]\\ &= a\,\mathrm{Cov}(X,Z)+b\,\mathrm{Cov}(Y,Z).\end{aligned}$$

Sijoittamalla tähän $b=0$ ja $Z=Y$ saadaan kaava (4.12). Sijoittamalla tähän $a=1$ ja $b=1$ saadaan kaava (4.13). Yleinen summakaava (4.14) seuraa soveltamalla kaavoja (4.11)–(4.13) sopivasti peräkkäin. □

4.4 Korrelaatio ja stokastinen riippuvuus

Lause 4.10. *Stokastisesti riippumattomille satunnaismuuttujille X ja Y pätee*

$$\mathrm{Cov}(X,Y)=0 \quad \textit{ja} \quad \mathrm{Cor}(X,Y)=0.$$

Todistus. Tarkastellaan satunnaismuuttujia X ja Y, joiden yhteisjakauma on diskreetti ja tiheysfunktio $f_{X,Y}(x,y)$. Kun X ja Y ovat stokastisesti riippumattomat, pätee lauseen 2.10 mukaan $f_{X,Y}(x,y)=f_X(x)f_Y(y)$. Soveltamalla muunnoksen odotusarvon laskukaavaa (3.6) funktioon $g(x,y)=xy$, saadaan

$$\begin{aligned}\mathbb{E}(XY) &= \sum_x\sum_y xy\, f_{X,Y}(x,y)\\ &= \sum_x\sum_y xy\, f_X(x)f_Y(y)\\ &= \left(\sum_x x f_X(x)\right)\left(\sum_y y f_Y(y)\right) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Yhdistämällä ylläoleva tulos kovarianssin laskukaavaan (4.10), nähdään että

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)\\ &= \mathbb{E}(X)\mathbb{E}(Y)-\mathbb{E}(X)\mathbb{E}(Y)\\ &= 0.\end{aligned}$$

Näin ollen myös $\mathrm{Cor}(X,Y)=0$. □

Lauseen 4.10 mukaan stokastinen riippumattomuus takaa korreloimattomuuden, mutta sama ei päde kääntäen. Seuraava esimerkki osoittaa, että korrelaatio voi olla nolla tilanteessa, jossa muuttujat X ja Y ovat stokastisesti riippuvat — jopa niin vahvasti, että ne määräytyvät tarkasti toisistaan epälineaarisen funktion välityksellä.

Esimerkki 4.11 (Satunnaismuuttuja ja sen neliö)**.** Laske satunnaislukujen X ja Y korrelaatio, kun X noudattaa joukon $\{-1, 0, 1\}$ tasajakaumaa ja $Y = X^2$.

Lasketaan ensiksi yhteisjakauman kovarianssi ja sen pohjustukseksi odotusarvot $\mathbb{E}(X), \mathbb{E}(Y)$ ja $\mathbb{E}(XY)$. Odotusarvon määritelmästä

$$\mathbb{E}(X) \;=\; \frac{1}{3} \times (-1) + \frac{1}{3} \times 0 + \frac{1}{3} \times 1 \;=\; 0,$$

ja muunnoskaavan (3.4) mukaan

$$\begin{aligned} \mathbb{E}(Y) \;=\; \mathbb{E}(X^2) \;&=\; \frac{1}{3} \times (-1)^2 + \frac{1}{3} \times 0^2 + \frac{1}{3} \times 1^2 \;=\; \frac{2}{3}, \\ \mathbb{E}(XY) \;=\; \mathbb{E}(X^3) \;&=\; \frac{1}{3} \times (-1)^3 + \frac{1}{3} \times 0^3 + \frac{1}{3} \times 1^3 \;=\; 0. \end{aligned}$$

Näin ollen laskukaavan (4.10) avulla

$$\mathrm{Cov}(X, Y) \;=\; \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \;=\; 0 - 0 \times \frac{2}{3} \;=\; 0,$$

joten myös $\mathrm{Cor}(X, Y) = 0$. ■

4.5 Korrelaatio ja lineaarinen riippuvuus

Korrelaation olemusta parhaiten kuvaa seuraava tulos, jonka mukaan korrelaation ääriarvot +1 ja -1 vastaavat determinististä lineaarista riippuvuutta.

Lause 4.12. *Satunnaismuuttujien X ja Y korrelaatiolle pätee aina*

$$-1 \;\leq\; \mathrm{Cor}(X, Y) \;\leq\; +1.$$

Lisäksi

$$\mathrm{Cor}(X, Y) \;=\; \begin{cases} +1, & \textit{jos ja vain jos } Y = aX + b \textit{ jollain } a > 0, \\ -1, & \textit{jos ja vain jos } Y = aX + b \textit{ jollain } a < 0. \end{cases}$$

Todistus. Merkitään X:n ja Y:n odotusarvoja μ_X, μ_Y ja keskihajontoja σ_X, σ_Y, sekä yhteisjakauman korrelaatiota $\rho = \mathrm{Cor}(X, Y)$. Tarkastellaan funktiota

$$p(t) \;=\; \mathbb{E}\Big[(tU + V)^2\Big], \tag{4.15}$$

jossa $U = X - \mu_X$ ja $V = Y - \mu_Y$. Kun huomataan, että $\mathbb{E}(U^2) = \sigma_X^2$, $\mathbb{E}(V^2) = \sigma_Y^2$ ja $\mathbb{E}(UV) = \mathrm{Cov}(X, Y) = \rho\sigma_X\sigma_Y$, voidaan ylläoleva funktio myös kirjoittaa muodossa

$$p(t) \;=\; \sigma_X^2 t^2 + 2\rho\sigma_X\sigma_Y t + \sigma_Y^2.$$

Funktio $p(t)$ on siis polynomi, jonka nollakohdat määräytyvät toisen asteen yhtälön ratkaisukaavasta

$$t \;=\; \frac{-2\rho\sigma_X\sigma_Y \pm \sqrt{D}}{2\sigma_X^2},$$

jonka diskriminantti on

$$D \;=\; (2\rho\sigma_X\sigma_Y)^2 - 4\sigma_X^2\sigma_Y^2 \;=\; 4\sigma_X^2\sigma_Y^2\left(\rho^2-1\right).$$

Koska yhtälön (4.15) mukaan $p(t) \geq 0$ kaikilla t, havaitaan että diskriminantti toteuttaa $D \leq 0$. Tämä puolestaan on mahdollista vain silloin kun $|\rho| \leq 1$.

Oletetaan seuraavaksi, että $|\rho| = 1$. Tällöin $D = 0$, ja polynomilla $p(t)$ on yksikäsitteinen nollakohta pisteessä

$$t \;=\; \frac{-2\rho\sigma_X\sigma_Y}{2\sigma_X^2} \;=\; -\rho\frac{\sigma_Y}{\sigma_X}.$$

Tässä pisteessä yhtälön (4.15) mukaan pätee $\mathbb{E}[(tU+V)^2] = 0$. Ei-negatiiviselle satunnaismuuttujalle $(tU+V)^2$ tämä on mahdollista vain, jos $(tU+V)^2 = 0$ todennäköisyydellä yksi. Varmuudella siis pätee

$$-\rho\frac{\sigma_Y}{\sigma_X}U + V \;=\; 0,$$

eli

$$Y - \mu_Y \;=\; \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X).$$

Ylläolevan yhtälön voi kirjoittaa muodossa $Y = aX+b$, kun valitaan $a = \rho\frac{\sigma_Y}{\sigma_X}$ ja $b = \mu_Y - a\mu_X$. Tällöin vakion a merkki määräytyy korrelaation ρ merkistä. □

Esimerkki 4.13 (Kaksi lineaarikombinaatiota)**.** Merkitään

$$\begin{aligned} X &= a_1U_1 + a_2U_2, \\ Y &= b_1U_1 + b_2U_2, \end{aligned}$$

missä U_1 ja U_2 ovat stokastisesti riippumattomat ja keskihajonnaltaan yksi. Määritä X:n ja Y:n korrelaatio.

Lasketaan ensiksi X:n ja Y:n kovarianssi. Koska U_1 ja U_2 ovat stokastisesti riippumattomat, pätee lauseen 4.10 mukaan $\mathrm{Cov}(U_1,U_2) = \mathrm{Cov}(U_2,U_1) = 0$. Näin ollen kovarianssin lineaarisuudesta (4.14) seuraa

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathrm{Cov}(a_1U_1 + a_2U_2,\; b_1U_1 + b_2U_2) \\ &= a_1b_1\,\mathrm{Cov}(U_1,U_1) + (a_1b_2 + a_2b_1)\,\mathrm{Cov}(U_1,U_2) + a_2b_2\,\mathrm{Cov}(U_2,U_2) \\ &= a_1b_1\,\mathrm{Cov}(U_1,U_1) + a_2b_2\,\mathrm{Cov}(U_2,U_2) \\ &= a_1b_1\,\mathrm{Var}(U_1) + a_2b_2\,\mathrm{Var}(U_2). \end{aligned}$$

Koska U_1 ja U_2 ovat keskihajonnoiltaan ja näin myös variansseiltaan yksi, päätellään että

$$\mathrm{Cov}(X,Y) \;=\; a_1b_1 + a_2b_2.$$

Korrelaation määrittämiseksi tulee vielä laskea X:n ja Y:n keskihajonnat. Samaan tapaan kuin yllä, voidaan päätellä että

$$\begin{aligned} \mathrm{Cov}(X,X) &= a_1^2 + a_2^2, \\ \mathrm{Cov}(Y,Y) &= b_1^2 + b_2^2, \end{aligned}$$

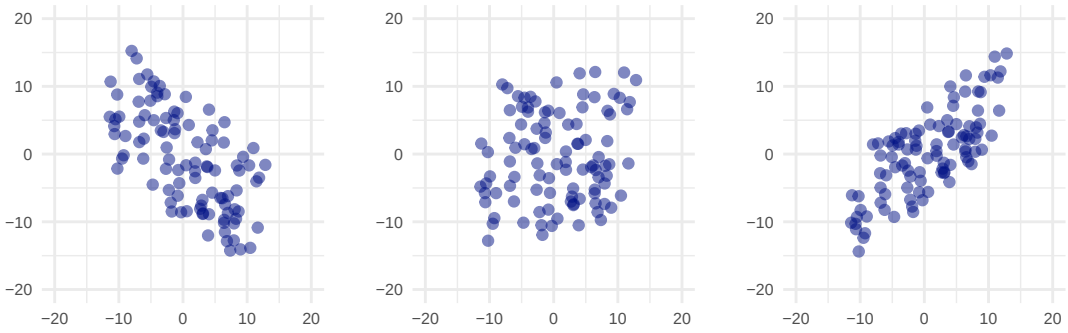
joten $\mathrm{SD}(X) = \mathrm{Cov}(X,X)^{1/2} = (a_1^2 + a_2^2)^{1/2}$ ja $\mathrm{SD}(Y) = (b_1^2 + b_2^2)^{1/2}$. Näin ollen kysytty korrelaatio on

$$\mathrm{Cor}(X,Y) \;=\; \frac{a_1 b_1 + a_2 b_2}{(a_1^2 + a_2^2)^{1/2}(b_1^2 + b_2^2)^{1/2}}.$$

Jos kertoimet $a = (a_1, a_2)$ ja $b = (b_1, b_2)$ tulkitaan 2-ulotteisiksi vektoreiksi, voidaan korrelaatio kirjoittaa vektoreiden pistetulon ja pituuksien avulla myös muodossa

$$\mathrm{Cor}(X,Y) \;=\; \frac{a \cdot b}{||a||\,||b||}.$$

Alla on simuloitu 100 lukuparia X:n ja Y:n yhteisjakaumasta kolmessa eri tapauksessa, kun U_1 ja U_2 arvottu välin $[-\sqrt{3}, \sqrt{3}]$ jatkuvasta tasajakaumasta. Vasemmalla, kun korrelaatio on reilusti negatiivinen ($\rho = -0.60$), X:n suuret arvot toteutuvat usein käsi kädessä Y:n pienten arvojen kanssa. Oikealla, kun korrelaatio on reilusti positiivinen ($\rho = 0.80$), X:n suuret arvot voidaan usein rinnastaa Y:n suuriin arvoihin.



$\rho = -0.60$
$a = (1,7)$, $b = (5,-5)$

$\rho = 0.28$
$a = (1,7)$, $b = (7,1)$

$\rho = 0.80$
$a = (1,7)$, $b = (5,5)$

4.6 Yhteenveto

Odotusarvo ja keskihajonta ovat kaksi tärkeintä yhden muuttujan jakaumaa kuvaavaa tunnuslukua, joita usein merkitään $\mu_X = \mathbb{E}(X)$ ja $\sigma_X = \mathrm{SD}(X)$. Odotusarvo kuvastaa jakaumasta tuotettujen satunnaislukujen keskiarvoista sijaintia ja keskihajonta niiden normitettua keskiarvoista poikkeamaa odotusarvosta. Odotusarvo ja keskihajonta voidaan laskea ao. kaavojen avulla.

Diskreetti jakauma	Jatkuva jakauma
$\mu \;=\; \sum_x x\, f(x)$	$\mu \;=\; \int x\, f(x) dx$
$\sigma \;=\; \left(\sum_x (x-\mu)^2 f(x)\right)^{1/2}$	$\sigma \;=\; \left(\int (x-\mu)^2 f(x)\, dx\right)^{1/2}$

Satunnaismuuttujan arvo osuu kahden keskihajonnan sisälle odotusarvostaan vähintään todennäköisyydellä 75%.

Satunnaismuuttujien yhteisjakauman korrelaatio kuvaa yhteisjakaumasta tuotettujen satunnaisten lukuparien keskiarvoista lineaarista riippuvuutta. Korrelaatiota merkitään usein symbolilla $\rho_{X,Y} = \mathrm{Cor}(X, Y)$ ja se voidaan laskea kovarianssin normitettuna versiona

$$\mathrm{Cor}(X, Y) \;=\; \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)}.$$

Korrelaatio kuuluu aina välille $[-1, 1]$ ja saa arvon nolla silloin, kun X ja Y ovat stokastisesti riippumattomat. Korrelaatio voi olla nolla myös toisistaan stokastisesti riippuville satunnaismuuttujille.

Satunnaismuuttujien yhteisjakauman kovarianssin voi laskea diskreetissä tapauksessa kaavalla

$$\mathrm{Cov}(X, Y) \;=\; \sum_x \sum_y (x - \mu_X)(y - \mu_Y)\, f(x, y)$$

ja jatkuvassa tapauksessa kaavalla

$$\mathrm{Cov}(X, Y) \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, f(x, y)\, dxdy.$$

Kovarianssi on operaattorina symmetrinen ja molempien argumenttiensa suhteen lineaarinen:

$$\begin{aligned} \mathrm{Cov}(Y, X) \;&=\; \mathrm{Cov}(X, Y), \\ \mathrm{Cov}\Big(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\Big) \;&=\; \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j\, \mathrm{Cov}(X_i, Y_j) \end{aligned}$$

Luku 5

Satunnaismuuttujien summa ja keskiarvo

5.1 Satunnaismuuttujien summa

Kahden satunnaismuuttujan summa $X + Y$ on satunnaismuuttuja, jonka jakauma voidaan määrittää X:n ja Y:n yhteisjakaumasta $f_{X,Y}(x, y)$. Summan tiheysfunktioksi saadaan

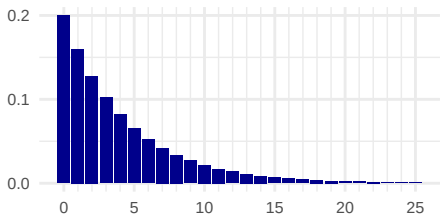
$$f_{X+Y}(s) = \begin{cases} \sum_x f_{X,Y}(x, s-x) & \text{(diskreetti yhteisjakauma),} \\ \int_{-\infty}^{\infty} f_{X,Y}(x, s-x)\, dx & \text{(jatkuva yhteisjakauma).} \end{cases}$$

Jos summan termit ovat stokastisesti riippumattomat, voidaan yllä olevat kaavat kirjoittaa tiheysfunktioiden $f_X(x)$ ja $f_Y(y)$ avulla[1] muodossa

$$f_{X+Y}(s) = \begin{cases} \sum_x f_X(x) f_Y(s-x) & \text{(diskreetti yhteisjakauma),} \\ \int_{-\infty}^{\infty} f_X(x) f_Y(s-x)\, dx & \text{(jatkuva yhteisjakauma).} \end{cases} \tag{5.1}$$

Esimerkki 5.1 (Kahden satunnaismuuttujan summa)**.** Satunnaismuuttujat X_1 ja X_2 ovat toisistaan riippumattomat ja noudattavat lukujoukon $\{0, 1, 2, \dots\}$ geometrista jakaumaa parametrina $a = 4/5$ ja tiheysfunktiona

$$f(x) = (1-a)a^x.$$



Määritä satunnaismuuttujan $X_1 + X_2$ jakauma.

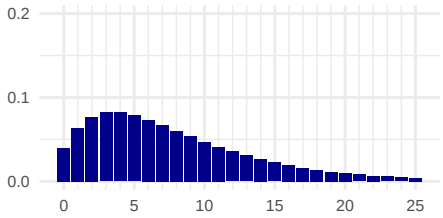
[1] Kaavan (5.1) yhtälöt voidaan tulkita tiheysfunktioiden f_X ja f_Y konvoluutioina. Yleisesti funktioiden f ja g *konvoluutio* $h = f \star g$ määritellään diskreetissä tilanteessa kaavalla $h(z) = \sum_x f(x)g(z-x)$ ja jatkuvassa tilanteessa kaavalla $h(z) = \int f(x)g(z-x)\, dx$.

Satunnaismuuttujan $X_1 + X_2$ arvojoukko on $\{0, 1, 2, \dots\}$ ja tiheysfunktio saadaan määritettyä summakaavasta (5.1). Koska $f(x) = 0$ pisteissä $x < 0$,

$$f_{X_1+X_2}(s) \;=\; \sum_x f(x)f(s-x) \;=\; \sum_{x=0}^{s}(1-a)a^x(1-a)a^{s-x}.$$

Sieventämällä oikeanpuolimmainen lauseke nähdään, että summan jakauma voidaan esittää tiheysfunktiona

$$f_{X_1+X_2}(s) \;=\; (1-a)^2(s+1)a^s.$$



Monen satunnaismuuttujien summa $S_n = X_1 + \cdots + X_n$ ja keskiarvo $n^{-1}S_n$ ovat satunnaismuuttujia, joiden avulla mallinnetaan satunnaisotannan havaintojen esiintyvyyksiä, kohinaisten mittausten keskiarvoja sekä talouden tuottoja kustannuskertymiä. Silloin kun summan termit ovat stokastisesti riippumattomia ja satunnaismuuttujan X tavoin jakautuneita, voidaan summan S_n jakauma määrittää X:n jakaumasta. Yksinkertaisimmassa tilanteessa summan termit ovat $\{0, 1\}$-arvoisia ja jakautuneet tiheysfunktion

$$f(x) \;=\; (1-p)^{1-x}p^x \;=\; \begin{cases} 1-p, & x = 0, \\ p, & x = 1, \end{cases}$$

mukaan. Tämä on *Bernoulli-jakauma* parametrina $p \in [0, 1]$, missä parametri p kertoo tapahtuman $X = 1$ todennäköisyyden. Tällöin summa S_n saa arvon x täsmälleen silloin, kun summattavista x saavat arvon 1 ja loput $n - x$ saavat arvon 0. Koska n:stä summattavasta voidaan valita $\binom{n}{x}$ tavalla x arvon 1 saavaa termiä, havaitaan että summan S_n jakauma noudattaa tiheysfunktiota

$$f(x) \;=\; \binom{n}{x}p^x(1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

Tämä on *binomijakauma* parametreina $n \geq 1$ ja $p \in [0, 1]$. Stokastisesti riippumattomien ja samoin jakautuneiden $\{0, 1\}$-arvoisten satunnaismuuttujien summan jakauma on siis aina binomijakauma.

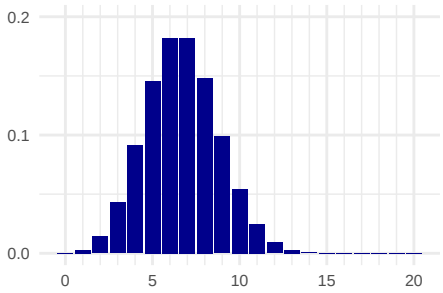
Esimerkki 5.2. Monivalintakokeessa on 20 kysymystä, joista jokaisessa pitää valita yksi oikea vastaus kolmen vaihtoehdon joukosta. Mikä on todennäköisyys saada kokeesta umpimähkään arvaamalla vähintään 19 oikein?

Oikeiden vastausten lukumäärä voidaan esittää summana $S_n = X_1 + \cdots + X_n$, jossa $n = 20$ ja

$$X_i \;=\; \begin{cases} 1, & \text{jos kysymyksen } i \text{ vastaus on oikein,} \\ 0, & \text{muuten.} \end{cases}$$

Umpimähkään arvatessa ovat yksittäisten kysymysten vastaukset toisistaan riippumattomat, ja yksittäinen vastaus on oikein todennäköisyydellä $\frac{1}{3}$. Näin ollen termit $X_1, \ldots, X_{20}$ ovat toisistaan riippumattomat ja Bernoulli-jakautuneet parametrina $p = \frac{1}{3}$. Tämän seurauksena summa S_n noudattaa binomijakaumaa parametreina $n = 20$ ja $p = \frac{1}{3}$ ja tiheysfunktiona

$$f(x) \;=\; \binom{20}{x}(1/3)^x(1-1/3)^{20-x}.$$



Todennäköisyys saada vähintään 19 oikein on siis

$$\begin{aligned}\mathbb{P}(S_n \geq 19) &= f(19) + f(20)\\ &\approx 11.47 \times 10^{-9} + 0.29 \times 10^{-9}\\ &\approx 12 \times 10^{-9}.\end{aligned}$$

Tiheysfunktion arvot pisteissä $x \geq 17$ ovat niin pieniä, että ne eivät näy yllä olevassa tiheysfunktion kuvaajassa. ■

Yleisessä tapauksessa, jossa summattavat eivät ole binaariarvoisia, ovat summan jakauman määrittämiseen tarvittavat konvoluutiokaavat ovat yleensä niin monimutkaisia, että summan jakauman lauseketta ei voi kirjoittaa siistissä suljetussa muodossa. Silloin kun summattavien määrä on suuri, voidaan summan jakaumaa kuitenkin arvioida hyvin tarkasti normaali- tai Poisson-jakauman avulla. Tässä luvussa opitaan soveltamaan normaali- ja Poisson-jakaumia käytännön tilanteissa esiintyvien summien ja keskiarvojen analysoimiseen.

5.2 Summan keskihajonta

Luvussa 3 esitetty suurten lukujen laki (lause 3.3) kertoo, että keskiarvo suuresta määrästä riippumattomia X:n tavoin jakautuneita satunnaislukuja (odotusarvo μ, keskihajonta σ) on suurella todennäköisyydellä likimain

$$\frac{1}{n}\sum_{i=1}^{n} X_i \;\approx\; \mu.$$

Suurten lukujen laki ei kuitenkaan kerro sitä, miten tarkka kyseinen arvio on, eikä sitä, miten summattavien lukumäärä n ja summattavien keskihajonta σ vaikuttavat arvion tarkkuuteen. Arvion tarkkuutta voidaan tutkia tarkastelemalla keskihajontaa

$$\mathrm{SD}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n}\,\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right).$$

Tämän auki laskemiseksi tarvitaan laskentakaava summan keskihajonnalle. Tarkastellaan ensiksi kahden muuttujan tapausta seuraavassa esimerkissä.

Esimerkki 5.3 (Kahden satunnaismuuttujan summa)**.** Mitä voidaan sanoa summan $X+Y$ keskihajonnasta, kun tunnetaan odotusarvot $\mu_X = 1$ ja $\mu_Y = 1$ sekä keskihajonnat $\sigma_X = 2$ ja $\sigma_Y = 3$?

Kovarianssin lineaarisuuden ja symmetrisyyden perusteella

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{Cov}(X+Y,\, X+Y)\\ &= \mathrm{Cov}(X,X) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,Y)\\ &= \mathrm{Var}(X) + 2\,\mathrm{Cov}(X,Y) + \mathrm{Var}(Y).\end{aligned}$$

Ottamalla yllä olevan yhtälön molemmilta puolilta neliöjuuret ja kirjoittamalla oikean puolen kovarianssitermi muodossa $\mathrm{Cov}(X,Y) = \rho\sigma_X\sigma_Y$, missä $\rho = \mathrm{Cor}(X,Y)$ on X:n ja Y:n korrelaatio, saadaan summan keskihajonnalle kaava

$$\sigma_{X+Y} = \left(\sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2\right)^{1/2}. \tag{5.2}$$

Summan keskihajontaa *ei* siis voi laskea tuntematta korrelaatiota. Soveltamalla kaavaan (5.2) korrelaation rajoja $-1 \le \rho \le 1$, saadaan summan keskihajonnalle kuitenkin estimaatit

$$|\sigma_X - \sigma_Y| \le \sigma_{X+Y} \le \sigma_X + \sigma_Y,$$

jotka kysymyksenasettelun lukuarvoilla vastaavat tapausta $1 \le \sigma_{X+Y} \le 5$. Jos X ja Y voidaan olettaa stokastisesti riippumattomiksi, voidaan kaavaan (5.2) sijoittaa $\rho = 0$, jolloin

$$\sigma_{X+Y} = \left(\sigma_X^2 + \sigma_Y^2\right)^{1/2},$$

mikä kysymyksenasettelun lukuarvoilla tuottaa $\sigma_{X+Y} \approx 3.61$. ■

Edellisessä esimerkissä johdettu summan keskihajonnan lauseke (5.2) yleistyy melko pienellä vaivalla myös kahta useamman satunnaismuuttujan summille.

Lause 5.4. *Satunnaismuuttujien $X_1, \ldots, X_n$ summan keskihajonta saadaan kaavasta*

$$\mathrm{SD}\left(\sum_i X_i\right) = \left(\sum_i \sigma_i^2 + \sum_i \sum_{j:j\neq i} \sigma_i\sigma_j\rho_{i,j}\right)^{1/2}, \tag{5.3}$$

jossa $\sigma_i = \mathrm{SD}(X_i)$ ja $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$.

Todistus. Kovarianssin lineaarisuudesta ja korrelaation määritelmästä seuraa

$$\begin{aligned}\mathrm{Var}\left(\sum_i X_i\right) &= \mathrm{Cov}\left(\sum_i X_i, \sum_j X_j\right)\\ &= \sum_i \sum_j \mathrm{Cov}(X_i, X_j)\\ &= \sum_i \mathrm{Cov}(X_i, X_i) + \sum_i \sum_{j:j\neq i} \mathrm{Cov}(X_i, X_j)\\ &= \sum_i \sigma_i^2 + \sum_i \sum_{j:j\neq i} \sigma_i\sigma_j\rho_{i,j},\end{aligned}$$

joten väite seuraa ottamalla yllä olevasta yhtälöstä neliöjuuret. □

Tärkeä erityistapaus yllä olevasta tuloksesta on tilanne, jossa $X_1, \ldots, X_n$ ovat korreloimattomia ($\rho_{i,j} = 0$) ja samoin jakautuneita ($\sigma_i = \sigma$), jolloin kaava (5.3) pelkistyy muotoon

$$\mathrm{SD}\Big(\sum_{i=1}^{n} X_i\Big) \;=\; \sigma\sqrt{n}. \tag{5.4}$$

Yllä oleva kaava on yksi stokastiikan tärkeimpiä tuloksia, sillä se kertoo tarkasti, miten riippumattomien ja samoin jakautuneiden satunnaismuuttujien summan keskihajonta käyttäytyy suhteessa summattavien lukumäärään. Erityisen merkillepantavaa on se, että suurilla n:n arvoilla on summan keskihajonta mitättömän pieni suhteessa summan odotusarvoon

$$\mathbb{E}\Big(\sum_{i=1}^{n} X_i\Big) \;=\; \mu n.$$

Esimerkki 5.5 (Noppapeli)**.** Pelataan n kierrosta noppapeliä, jossa yksittäisellä kierroksella voittaa nopan silmäluvun mukaisen määrän euroja. Laske kertyneen tuoton $S = X_1 + \cdots + X_n$ odotusarvo ja keskihajonta tapauksissa $n = 10, 100, 1000$.

Yhden kierroksen tuoton odotusarvo on

$$\mu_X \;=\; \frac{1}{6}\times 1 + \frac{1}{6}\times 2 + \cdots + \frac{1}{6}\times 6 \;=\; 3.5$$

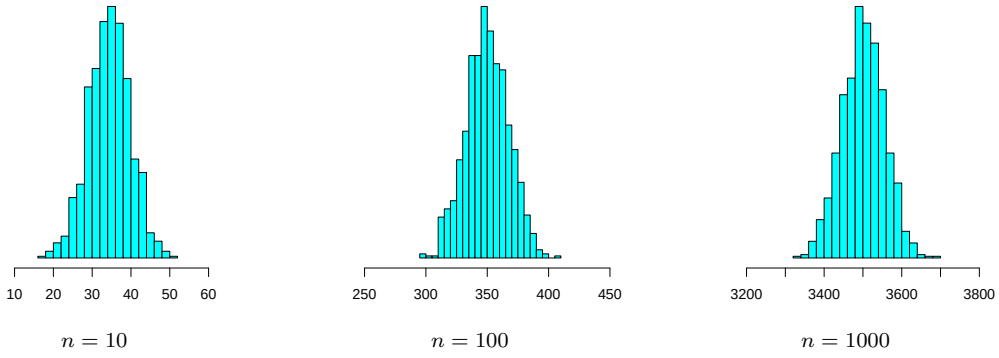
ja keskihajonta kahden desimaalin tarkkuudella

$$\sigma_X \;=\; \left(\frac{1}{6}\times(1-\mu_X)^2 + \frac{1}{6}\times(2-\mu_X)^2 + \cdots + \frac{1}{6}\times(6-\mu_X)^2\right)^{1/2} \;=\; 1.71.$$

Koska pelikierrokset ovat stokastisesti riippumattomat ja samoin jakautuneet, saadaan kertyneen tuoton odotusarvoksi $\mu_S = \mu_X n$ ja keskihajonnaksi $\sigma_S = \sigma_X\sqrt{n}$. Tulokset eri n:n arvoilla ovat alla.

n	μ_S	σ_S
10	35	5.4
100	350	17.1
1000	3500	54.0

Alla olevassa kuvassa on simuloimalla tuotettuja kertyneen tuoton S_n jakaumia. Jokaisessa kuvassa havaitaan, että käytännössä kaikki simuloidut arvot sisältyvät neljän keskihajonnan sisään odotusarvosta. Chebyshevin epäyhtälön (lause 4.6) mukaan tiedetään, että näin tapahtuu vähintään todennäköisyydellä $\frac{15}{16} = 93.75\%$.



Esimerkki 5.6 (Lentoyhtiö)**.** 300 lentolippua myydään lennolle, jossa on 290 matkustajapaikkaa. Arviolta 5% lipun ostaneista jää saapumatta lennolle, toisistaan riippumattomasti. Millä todennäköisyydellä kaikki saapujat mahtuvat lennolle?

Lennolle saapuvien matkustajien lukumäärä voidaan kirjoittaa satunnaismuuttujien summana $T = X_1 + \cdots + X_{300}$, jossa

$$X_i \;=\; \begin{cases} 1, & \text{jos lentolipun } i \text{ ostaja saapuu lennolle,} \\ 0, & \text{muuten.} \end{cases}$$

Indikaattorimuuttujan X_i odotusarvo on

$$\mu_X \;=\; 0.05 \times 0 + 0.95 \times 1 \;=\; 0.95$$

ja keskihajonta

$$\sigma_X \;=\; \Big(0.05 \times (0 - \mu_X)^2 + 0.95 \times (1 - \mu_X)^2\Big)^{1/2} \;=\; 0.218.$$

Koska satunnaismuuttujat $X_1, X_2, \ldots$ ovat stokastisesti riippumattomat ja samoin jakautuneet, saadaan satunnaismuuttujan T odotusarvoksi $\mu_T = \mu_X \times 300 = 285$ ja keskihajonnaksi $\sigma_T = \sigma_X \times \sqrt{300} = 3.77$. Kaikki saapujat mahtuvat lennolle silloin, kun $N \leq 290$. Tämän tapahtuman todennäköisyyttä voidaan Chebyshevin epäyhtälön avulla arvioida muodossa

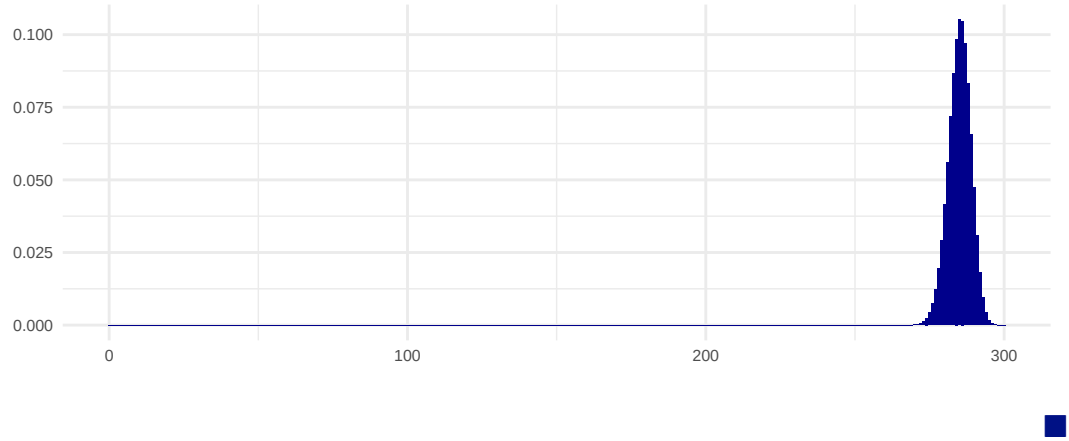
$$\mathbb{P}(T \leq 290) \;\geq\; \mathbb{P}(T \in [280, 290]) \;=\; \mathbb{P}(T = \mu_T \pm 1.32\sigma_T) \;\geq\; 1 - \frac{1}{1.32^2} \approx 42.6\%.$$

Kaikki saapujat mahtuvat siis lennolle vähintään todennäköisyydellä 42.6%. Tämä alaraja kuulostaa hyvin pessimistiseltä arviolta. Koska T on riippumattomien ja samoin jakautuneiden $\{0, 1\}$-arvoisten satunnaismuuttujien summa, tunnetaan sen jakauma itse asiassa tarkasti. Kuten kappaleessa 5.1 todettiin, noudattaa T binomijakaumaa parametreina $n = 300$ ja $p = 0.95$. Tietokoneella voidaan laskea tarkka todennäköisyys

$$\mathbb{P}(T \leq 290) \;=\; 93.5\%.$$

Binomijakaumalle Chebyshevin epäyhtälö antaa siis ylipessimistisiä arvioita[2]

Alla on kuva satunnaismuuttujan T jakauman tiheysfunktiosta. Tiheysfunktion arvot ovat aidosti positiivisia kaikilla $x \in \{0, 1, \ldots, 300\}$, mutta tähtitieteellisen pieniä kun $x \leq 250$, joten ne eivät näy kuvassa.



5.3 Satunnaismuuttujien keskiarvo ja suurten lukujen laki

Summan keskihajonnan avulla voidaan todistaa yleisempi versio aiemmasta suurten lukujen laista (lause 3.3). Summattavien ei tarvitse olla stokastisesti riippumattomia, vaan riittää että ne ovat korreloimattomia. Lasketaan ensiksi satunnaismuuttujien keskiarvon $\frac{1}{n}\sum_{i=1}^{n} X_i$ odotusarvo ja keskihajonta.

Lause 5.7. *Jos satunnaismuuttujat $X_1, \ldots, X_n$ ovat korreloimattomia ja kaikilla on sama odotusarvo μ ja keskihajonta σ, niin tällöin satunnaismuuttujan $\frac{1}{n}\sum_{i=1}^{n} X_i$ odotusarvo on μ ja keskihajonta $\frac{\sigma}{\sqrt{n}}$.*

Todistus. Odotusarvon lineaarisuuden (lause 3.15) perusteella voidaan satunnaismuuttujan $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ odotusarvoksi laskea

$$\mathbb{E}(M_n) \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mu \;=\; \mu.$$

Koska korrelaatiot $\rho_{i,j} = \mathrm{Cor}(X_i, X_j)$ ovat nollia, saadaan kaavan (5.3) avulla

$$\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right) \;=\; \left(\sum_{i}\mathrm{SD}(X_i)^2\right)^{1/2} \;=\; \left(\sum_{i}\sigma^2\right)^{1/2} \;=\; \sigma\sqrt{n}.$$

[2] riippumattomien satunnaismuuttujien summille saadaan tarkempia estimaatteja ns. Chernoffin epäyhtälön avulla

Keskihajonnan perusominaisuuksien (lause 4.5) mukaan

$$\mathrm{SD}(M_n) \;=\; \mathrm{SD}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n}\,\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n}\sigma\sqrt{n} \;=\; \frac{\sigma}{\sqrt{n}}.$$

□

Lause 5.8 (Suurten lukujen laki)**.** *Jos satunnaismuuttujat $X_1, X_2, \ldots$ ovat korreloimattomia, ja kaikilla on sama odotusarvo μ ja keskihajonta σ, niin mielivaltaisen pienellä $\epsilon > 0$, tapahtuman*

$$n^{-1}\sum_{k=1}^{n} X_k \;=\; \mu \pm \epsilon \tag{5.5}$$

todennäköisyys lähestyy ykköstä suurilla n:n arvoilla[3]*.*

Todistus. Lauseen 5.7 mukaan satunnaismuuttujan $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ odotusarvo on $\mu_{M_n} = \mu$ ja keskihajonta $\sigma_{M_n} = \sigma n^{-1/2}$. Kun merkitään $k = \frac{\epsilon n^{1/2}}{\sigma}$, voidaan tapahtuma (5.5) lausua muodossa

$$M_n \;=\; \mu_{M_n} \pm k\sigma_{M_n},$$

ja Chebyshevin epäyhtälön tämän tapahtuman todennäköisyys on vähintään

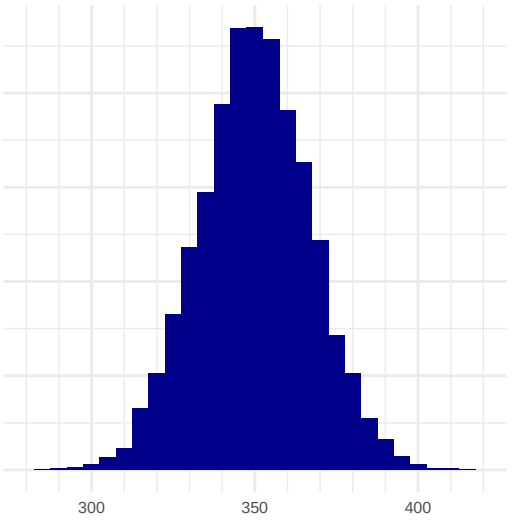
$$\mathbb{P}(M_n \;=\; \mu_{M_n} \pm k\sigma_{M_n}) \;\geq\; 1 - \frac{1}{k^2} \;=\; 1 - \frac{\sigma^2}{\epsilon^2 n}.$$

Väite seuraa, koska yllä olevan epäyhtälön oikea puoli lähestyy ykköstä, kun n kasvaa. □

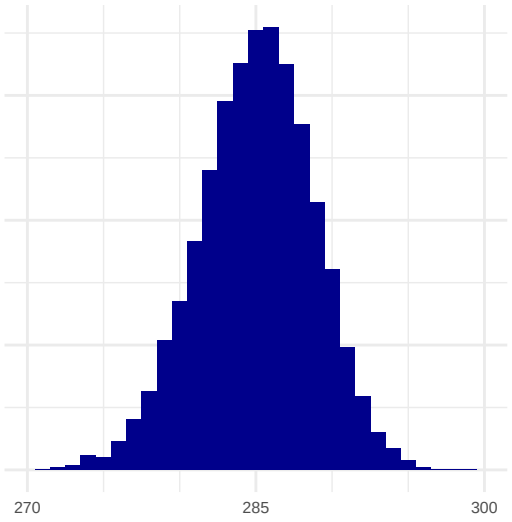
5.4 Summan normaaliapproksimaatio

Esimerkissä 5.5 simuloitu sadan nopanheiton summan $S = S_{100}$ ja esimerkissä 5.6 simuloitu kolmensadan indikaattorimuuttujan summa T ovat muodoltaan samankaltaiset, kuten alla oleva kuva osoittaa.

[3] Tarkemmin ilmaistuna $\lim_{n\to\infty} \mathbb{P}(|n^{-1}\sum_{k=1}^{n} X_k - \mu| \leq \epsilon) = 1$.



$S = S_{100}$ (esimerkki 5.5)



T (esimerkki 5.6)

Jakaumat ovat jopa yllättävän samankaltaiset, sillä noppapelin tuottokertymä $S = S_{100}$ ja lennolle saapuvien lukumäärä T liittyvät täysin erilaisiin kontekstei-hin. Ainoa kyseisiä satunnaismuuttujia yhdistävä tekijä on se, että molemmat voidaan tulkita stokastisesti riippumattomien satunnaismuuttujien summana.

Jakaumien muotoa voi tarkemmin vertailla piirtämällä normitettujen satun-naismuuttujien

$$\tilde{S} = \frac{S - \mu_S}{\sigma_S} \quad \text{ja} \quad \tilde{T} = \frac{T - \mu_T}{\sigma_T}$$

jakaumat. Ne on esitetty kuvassa 5.1. Punaisella piirretty jakaumien muotoa tarkasti approksimoiva funktio on

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{5.6}$$

Kyseinen Gaussin kellokäyränä tunnettu funktio on positiivinen ja integroituu ykköseksi, joten se on erään jatkuvan jakauman tiheysfunktio. Tiheysfunktion (5.6) määrittämä jatkuva jakauma on nimeltään *normitettu normaalijakauma*.

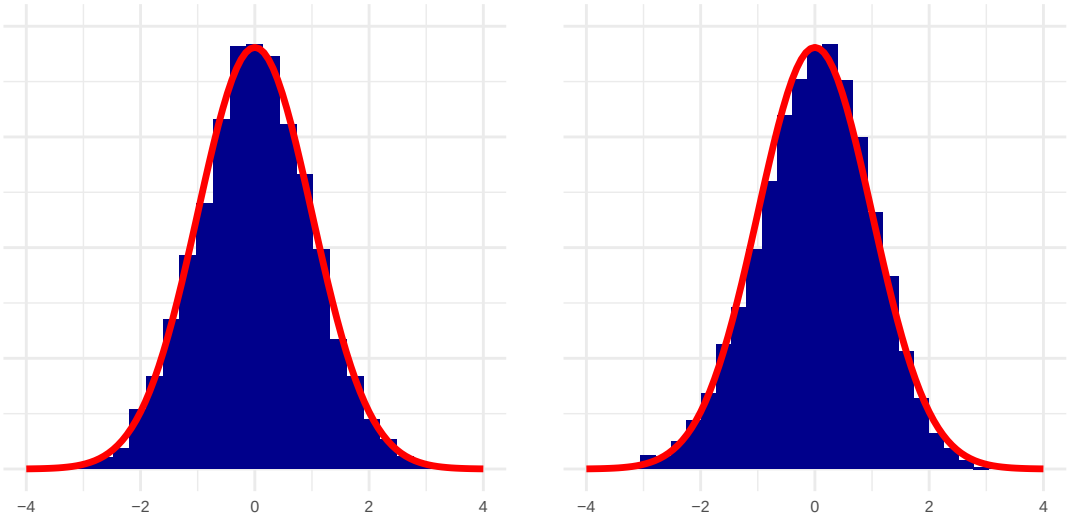
Normitettujen jakaumien samankaltaisuus on universaali matematiikan laki, joka koskee *kaikkia* stokastisesti riippumattomia satunnaismuuttujien summia. Tämä tärkeä tulos tunnetaan nimellä *keskeinen raja-arvolause*.

Lause 5.9 (Keskeinen raja-arvolause)**.** *Jos summan $S_n = X_1 + \cdots X_n$ termit ovat stokastisesti riippumattomia ja samoin jakautuneita satunnaismuuttujia, joilla on odotusarvo μ_X ja keskihajonta $0 < \sigma_X < \infty$, niin normitettu summa*

$$\tilde{S}_n = \frac{S_n - \mu_{S_n}}{\sigma_{S_n}},$$

jossa $\mu_{S_n} = \mu_X n$ ja $\sigma_{S_n} = \sigma_X \sqrt{n}$, noudattaa suurilla n arvoilla likimain normitettua normaalijakaumaa.

Todistus sivuutetaan tässä yhteydessä.



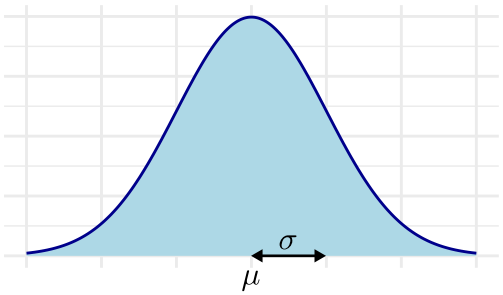
$\tilde{S}$ (esimerkki 5.5) $\tilde{T}$ (esimerkki 5.6)

Kuva 5.1: Normitettujen satunnaismuuttujien $\tilde{S}$ ja $\tilde{T}$ simuloidut jakaumat.

5.5 Normaalijakauma

Yleinen *normaalijakauma* parametreina $\mu \in (-\infty, \infty)$ ja $\sigma \in (0, \infty)$ on yhden muuttujan jatkuva jakauma, jonka tiheysfunktio on

$$f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Tiheysfunktiota sopivasti osittain integroimalla voidaan vahvistaa, että

$$\mu \;=\; \int_{-\infty}^{\infty} x f(x)\,dx \qquad \text{ja} \qquad \sigma \;=\; \left(\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \right)^{1/2},$$

joten parametri μ on normaalijakauman odotusarvo ja parametri σ sen keskihajonta. Normaalijakauman kertymäfunktiota tarkastelemalla havaitaan myös, että jos X on normaalijakautunut parametrein μ_X ja σ_X, niin tällöin $Y = a+bX$ on normaalijakautunut parametrein $\mu_Y = a + b\mu_X$ ja $\sigma_Y = |b|\sigma_X$. Tästä seuraa, että normitettu satunnaismuuttuja

$$Z \;=\; \frac{X - \mu_X}{\sigma_X} \tag{5.7}$$

noudattaa normitettua normaalijakaumaa odotusarvona 0 ja keskihajontana 1. Vastaavasti mikä tahansa parametrin μ ja σ normaalijakautunut satunnaismuuttuja voidaan esittää muodossa

$$X = \mu + \sigma Z, \tag{5.8}$$

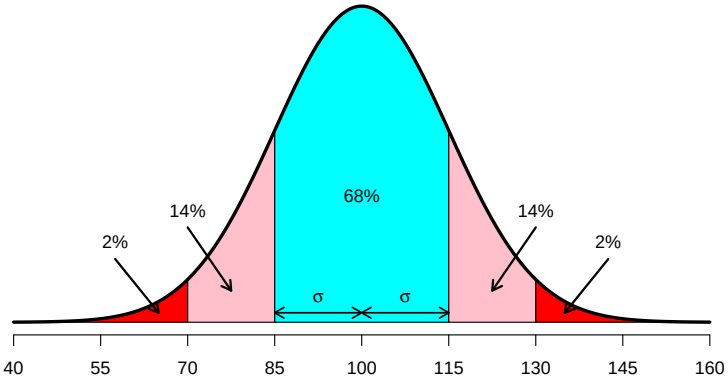
missä Z noudattaa normitettua normaalijakaumaa.

Normaalijakauman kertymäfunktiota ei voi esittää siistissä suljetussa muodossa, joten siihen liittyvät todennäköisyydet lasketaan kertymäfunktion taulukoiden (liite B) tai numeeristen ohjelmistojen (R: `pnorm()`, Excel: `NORM.DIST()`) avulla. Normaalijakauman taulukoissa yleensä raportoidaan vain normitetun normaalijakauman kertymäfunktion arvot, sillä muut normaalijakaumat voidaan palauttaa normitettuun tapaukseen kaavojen (5.7)–(5.8) avulla.

Esimerkki 5.10 (Älykkyysosamäärä)**.** Yhdeksäsluokkalaisten älykkyysosamäärä noudattaa likimain normaalijakaumaa ($\mu = 100$, $\sigma = 15$). Millä todennäköisyydellä satunnaisesti valitun yhdeksäsluokkalaisen älykkyysosamäärä on

(a) yli 130?

(b) välillä 85–115?



Normitettu satunnaismuuttuja $Z = \frac{X-\mu}{\sigma}$ noudattaa normitettua normaalijakaumaa, joten

$$\mathbb{P}(X > 130) = \mathbb{P}\left(\frac{X-\mu}{\sigma} > \frac{130-100}{15}\right) = \mathbb{P}(Z > 2).$$

Normitetun normaalijakauman symmetrian ja jatkuvuuden perusteella pätee $\mathbb{P}(Z > 2) = \mathbb{P}(Z < -2) = \mathbb{P}(Z \leq -2)$. Vastaukseksi (a)-kohtaan saadaan normaalijakauman taulukoista $\mathbb{P}(Z \leq -2) \approx 0.023$.

Samaan tapaan

$$\begin{aligned}
\mathbb{P}(85 \leq X \leq 115) &= \mathbb{P}\left(\frac{85-100}{15} \leq \frac{X-\mu}{\sigma} \leq \frac{115-100}{15}\right) \\
&= \mathbb{P}(-1 \leq Z \leq 1) \\
&= \mathbb{P}(-1 < Z \leq 1) \\
&= \mathbb{P}(Z \leq 1) - \mathbb{P}(Z \leq -1),
\end{aligned}$$

joten (b)-kohdan vastaukseksi saadaan normaalijakauman taulukoista $\mathbb{P}(Z \leq 1) - \mathbb{P}(Z \leq -1) \approx 0.683$. ■

Esimerkki 5.11 (Noppapeli)**.** Arvioi normaalijakauman avulla, millä todennäköisyydellä esimerkin 5.5 noppapelissä 100 pelikierrokselta kertynyt tuotto on

(a) välillä 316–384 EUR?

(b) yli 500 EUR?

Merkitään kertynyttä tuottoa $S_{100} = X_1 + \cdots + X_{100}$. Koska yhden kierroksen tuoton odotusarvo ja keskihajonta (yhden desimaalin tarkkuudella) ovat $\mu_X = 3.5$ ja $\sigma_X = 1.7$, ja tuotot ovat stokastisesti riippumattomat, on 100 pelikierroksen tuoton odotusarvo

$$\mu_{S_{100}} \;=\; 3.5 \times 100 \;=\; 350$$

ja keskihajonta

$$\sigma_{S_{100}} \;=\; 1.7 \times \sqrt{100} \;=\; 17.$$

Kun normitetun tuottokertymän $\frac{S_{100}-350}{17}$ jakaumaa arvioidaan normitettua normaalijakaumaa noudattavalla satunnaismuuttujalla Z, saadaan tulokseksi

$$\begin{aligned}
\mathbb{P}(316 \leq S_{100} \leq 384) &= \mathbb{P}\left(-2 \leq \frac{S_{100} - 350}{17} \leq 2\right) \\
&\approx \mathbb{P}(-2 \leq Z \leq 2) \\
&= 1 - 2\mathbb{P}(Z \leq -2) \\
&\approx 95.4\%.
\end{aligned}$$

ja

$$\begin{aligned}
\mathbb{P}(S_{100} > 500) &= \mathbb{P}\left(\frac{S_{100} - 350}{17} > 8.82\right) \\
&\approx \mathbb{P}(Z > 8.82) \\
&= \mathbb{P}(Z \leq -8.82) \\
&\approx 6 \times 10^{-19}.
\end{aligned}$$

■

Esimerkki 5.12 (Lentoyhtiö)**.** Arvioi normaalijakauman avulla, millä todennäköisyydellä esimerkissä 5.6 kaikki lennolle saapuvat matkustajat mahtuvat lennolle.

Esimerkissä 5.6 johdettiin lennolle saapuvien matkustajien lukumäärän T odotusarvoksi $\mu_T = 285$ ja keskihajonnaksi $\sigma_T = 3.77$. Lennolle saapuvien matkustajien normitettu lukumäärä on satunnaismuuttuja

$$\frac{T - \mu_T}{\sigma_T} \;=\; \frac{T - 285}{3.77}.$$

Kun satunnaismuuttujan $\frac{T-285}{3.77}$ jakaumaa arvioidaan normitettua normaalijakaumaa noudattavalla satunnaismuuttujalla Z, havaitaan että kaikki matkustajat mahtuvat lennolle todennäköisyydellä

$$\begin{aligned}
\mathbb{P}(T \leq 290) &= \mathbb{P}\left(\frac{T-285}{3.77} \leq \frac{290-285}{3.77}\right) \\
&= \mathbb{P}\left(\frac{T-285}{3.77} \leq 1.33\right) \\
&\approx \mathbb{P}(Z \leq 1.33) \\
&= 90.8\%.
\end{aligned}$$

Hieman tarkemman arvion saa huomaamalla, kokonaislukuarvoiselle satunnaismuuttujalle T pätee $\mathbb{P}(T \leq 290) = \mathbb{P}(T \leq 290.5)$, jolloin

$$\begin{aligned}
\mathbb{P}(T \leq 290) &= \mathbb{P}(T \leq 290.5) \\
&= \mathbb{P}\left(\frac{T-285}{3.77} \leq \frac{290.5-285}{3.77}\right) \\
&= \mathbb{P}\left(\frac{T-285}{3.77} \leq 1.46\right) \\
&\approx \mathbb{P}(Z \leq 1.46) \\
&= 92.8\%.
\end{aligned}$$

Näin saatu ns. jatkuvuuskorjaus tuottaa hieman tarkemman arvion, sillä tapahtuman tarkka todennäköisyys on binomijakauman mukaan $\mathbb{P}(T \leq 290) = 93.5\%$. ■

5.6 Poisson-approksimaatio

Keskeinen raja-arvolause kertoo, että stokastisesti riippumattomien ja samoin jakautuneiden satunnaismuuttujien summa $S_n = X_1 + \cdots X_n$ noudattaa suurilla n:n arvoilla likimain normaalijakaumaa parametrein $\mu_X n$ ja $\sigma_X\sqrt{n}$, kunhan summattavien keskihajonta σ_X on aidosti positiivinen ja äärellinen. Tietyissä tilanteissa tarvitaan arvioita satunnaismuuttujien summalle, jossa σ_X on hyvin lähellä nollaa. Tällöin normaaliapproksimaation tarkkuus on heikko.

Esimerkki 5.13. Suositun uutissivuston www-palvelimelle saapuu keskimäärin $\lambda = 2.6$ sivupyyntöä sekunnissa. Arvioi todennäköisyys, jolla seuraavan sekunnin aikana palvelimelle saapuu yli 10 sivupyyntöä.

Luonnollinen malli sekunnin aikana saapuville sivupyynnöille on satunnaismuuttujien summa $S_n = \sum_{i=1}^{n} X_i$, missä n on uutissivustoa seuraavien käyttäjien lukumäärä ja

$$X_i = \begin{cases} 1, & \text{jos käyttäjältä } i \text{ saapuu sivupyyntö,} \\ 0, & \text{muuten.} \end{cases}$$

Summattavien indikaattorimuuttujien odotusarvo on $\mu_X = p$ ja keskihajonta $\sigma_X = (p(1-p))^{1/2}$, missä $p = \mathbb{P}(X_i = 1)$. Näin ollen saapuvien sivupyyntöjen odotusarvo voidaan kirjoittaa muodossa $\mathbb{E}(S_n) = np$. Parametreja n ja p ei tehtävänannon pohjalta tunneta, mutta tunnetun odotusarvon λ pohjalta voidaan ratkaista $p = \frac{\lambda}{n}$. Kun uutissivustoa seuraavien käyttäjien lukumäärä n on suuri, on summattavien keskihajonta likimain

$$\sigma_X \;=\; (p(1-p))^{1/2} \;\approx\; \lambda^{1/2} n^{-1/2}.$$

Koska σ_X on hyvin lähellä nollaa, ei normaaliapproksimaation tarkkuudelle ole takeita. ■

Yllä olevan esimerkin tilanteeseen sopiva approksimoiva jakauma on lukujoukon $\{0, 1, 2, \dots\}$ diskreetti jakauma tiheysfunktiona

$$f(x) \;=\; e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Tämä jakauma on *Poisson-jakauma* parametrina $\lambda > 0$. Jakauma on nimetty ranskalaismatemaatikko Siméon Denis Poissonin (1781–1840) mukaan. Seuraava tulos tunnetaan nimellä pienten lukujen laki.

Lause 5.14. *Jos summan $S_n = X_1 + \cdots X_n$ termit ovat stokastisesti riippumattomia ja samoin jakautuneita $\{0,1\}$-arvoisia satunnaismuuttujia odotusarvona $\mu_X \approx \lambda/n$, niin S_n noudattaa suurilla n likimain Poisson-jakaumaa parametrina λ.*

Todistus. Yllä olevien oletusten vallitessa S_n noudattaa binomijakaumaa parametreina n ja $p = \mu_X$, joten

$$\mathbb{P}(S_n = x) \;=\; \binom{n}{x} p^x (1-p)^{n-x}.$$

Kun n on suuri, yllä esiintyvä binomikerroin on likimain

$$\binom{n}{x} \;=\; \frac{1}{x!} \prod_{k=0}^{x-1} (n-k) \;=\; \frac{n^x}{x!} \prod_{k=0}^{x-1} \left(1 - \frac{k}{n}\right) \;\approx\; \frac{n^x}{x!}.$$

Lisäksi kun $p \approx \frac{\lambda}{n}$, pätee $p^x \;\approx\; \left(\frac{\lambda}{n}\right)^x$, ja kaavan $\lim_{n\to\infty} \left(1 + \frac{t}{n}\right)^n = e^t$ avulla

$$(1-p)^{n-x} \;\approx\; \left(1 - \frac{\lambda}{n}\right)^{n-x} \;=\; \left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^{n} \;\approx\; e^{-\lambda}.$$

Yhdistämällä nämä kolme arviota havaitaan, että

$$\mathbb{P}(S_n = x) \;=\; \binom{n}{x} p^x (1-p)^{n-x} \;\approx\; \frac{n^x}{x!} \left(\frac{\lambda}{n}\right)^x e^{-\lambda} \;=\; e^{-\lambda} \frac{\lambda^x}{x!}.$$

□

Binomijakaumaa parametreina n ja p voidaan siis arvioida kahdella eri jakaumalla:

(i) normaalijakauma parametrein $\mu = np$ ja $\sigma = (np(1-p))^{1/2}$, tarkka silloin kun n on suuri ja p ei kovin lähellä nollaa eikä ykköstä

(ii) Poisson-jakauma parametrina $\lambda = np$, tarkka silloin kun n on suuri ja p lähellä nollaa.

Esimerkki 5.15. Suositun uutissivuston www-palvelimelle saapuu keskimäärin $\lambda = 2.6$ sivupyyntöä sekunnissa. Arvioi todennäköisyys, jolla seuraavan sekunnin aikana palvelimelle saapuu yli 10 sivupyyntöä.

Saapuvien sivupyyntöjen lukumäärää on luonnollista arvioida binomijakaumalla parametreina n ja $p \approx \frac{\lambda}{n}$. Lauseen 5.14 mukaan suurella n kyseinen binomijakauma on likimain Poisson-jakauma parametrina λ. Kysytty todennäköisyys on siis arviolta

$$\mathbb{P}(S_n > 10) \;=\; 1 - \mathbb{P}(S_n \leq 10) \;\approx\; \sum_{x=0}^{10} e^{-\lambda}\frac{\lambda^x}{x!} \;\approx\; 0.000087.$$

■

5.7 Yhteenveto

Satunnaismuuttujien summan $S_n = \sum_{i=1}^n X_i$ odotusarvo ja keskihajonta määräytyvät ao. taulukon kaavoista.

Summan termit	$\mathbb{E}(\sum_i X_i)$	$\mathrm{SD}(\sum_i X_i)$
Yleiset	$\sum_i \mu_i$	$\left(\sum_i \sigma_i^2 + \sum_i \sum_{j:j\neq i} \sigma_i\sigma_j\rho_{i,j}\right)^{1/2}$
Korreloimattomat	$\sum_i \mu_i$	$\left(\sum_i \sigma_i^2\right)^{1/2}$
Korreloimattomat ja samoin jakautuneet	μn	$\sigma\sqrt{n}$

Jos satunnaismuuttujien summan $S_n = X_1 + \cdots X_n$ termit ovat stokastisesti riippumattomia ja samoin jakautuneita, odotusarvona μ_X ja keskihajontana σ_X, niin summan odotusarvo on

$$\mu_{S_n} \;=\; \mu_X n$$

ja keskihajonta

$$\sigma_{S_n} \;=\; \sigma_X\sqrt{n}.$$

Silloin kun σ_X on aidosti positiivinen ja äärellinen, noudattaa normitettu summa $\frac{S_n - \mu_{S_n}}{\sigma_{S_n}}$ suurilla n likimain normitettua normaalijakaumaa, joten jakauman näkökulmasta

$$S_n \;\approx\; \mu_{S_n} + \sigma_{S_n} Z,$$

missä Z noudattaa normitettua normaalijakaumaa. Jos summattavat ovat $\{0, 1\}$-arvoisia, on summan tarkka jakauma binomijakauma parametreina n ja $p = \mu_X$. Kun p ei ole liian lähellä nollaa tai ykköstä, voidaan kyseistä binomijakaumaa arvioida yo. normaalijakaumaa käyttäen. Pienillä $p \approx \lambda/n$ arvioilla parempi arvio saadaan Poisson-jakaumasta parametrina $\lambda > 0$.

Luku 6

Datajoukkojen jakaumat, tunnusluvut ja kuvaajat

6.1 Datajoukko ja datakehikko

Tässä monisteessa *datajoukko*[1] tarkoittaa järjestettyä listaa keskenään samantyyppisiä alkioita, esimerkiksi lukuja, merkkijonoja tai näistä muodostettuja listoja. Moniulotteinen datajoukko on datajoukko, jonka alkiot ovat järjestettyjä listoja. Moniulotteinen datajoukko esitetään yleensä *datakehikkona* (engl. data frame) eli taulukkona, jonka jokainen rivi vastaa yhtä moniulotteisen datajoukon alkiota, ja jonka sarakkeita kutsutaan datajoukon *muuttujiksi*.

Esimerkki 6.1. Alla oleva datakehikko kuvastaa fiktiivisen kurssin kurssipalautteesta koostettua neliulotteista datajoukkoa

((12345A, 5, 1, 5), (98759K, 1, 5, 2), (33312K, 4, 4, 3), (23453B, 4, 4, 3), (21453U, 3, 3, 3)),

jossa on yksi merkkijonoarvoinen muuttuja (opiskelijanumero) ja kolme lukuarvoista muuttujaa (yleisarvio, työläys, hyödyllisyys).

Opiskelijanumero	Yleisarvio	Työläys	Hyödyllisyys
12345A	5	1	5
98759K	1	5	2
33312K	4	4	3
23453B	4	4	3
21453U	3	3	5

Tämä datajoukko voidaan myös tulkita neljän yksiulotteisen datajoukon listana, esimerkiksi muuttujaa "Yleisarvio" vastaa datajoukko $(5, 1, 4, 4, 3)$. ■

6.2 Datajoukon keskiarvo ja keskihajonta

Havaittua datajoukkoa on usein tapana kuvailla raportoimalla siitä laskettuja yksittäisiä lukuarvoja. Tällaista lukua kutsutaan *tunnusluvuksi* (engl. statistic).

[1] Datajoukko *ei* ole tarkassa matemaattisessa mielessä joukko, sillä datajoukossa sama alkio voi esiintyä monta kertaa.

Lukuarvoisen datajoukon $\vec{x} = (x_1, \ldots, x_n)$ sijaintia kuvaavista tunnusluvuista yleisin on datajoukon *keskiarvo* (engl. mean)

$$m(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{6.1}$$

Hajontaa kuvaavista tunnusluvuista yleisin on datajoukon *keskihajonta* (engl. standard deviation)

$$\mathrm{sd}(\vec{x}) \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\vec{x}))^2\right)^{1/2}. \tag{6.2}$$

Datajoukon keskihajonta $\mathrm{sd}(\vec{x})$ kuvaa arvojen $x_1, \ldots, x_n$ keskiarvoista poikkeamaa keskiarvosta $m(\vec{x})$. Datajoukon *varianssi* määritellään keskihajonnan ja neliönä $\mathrm{var}(\vec{x}) = \mathrm{sd}(\vec{x})^2$. Sekin mittaa datajoukon hajontaa keskiarvon ympärillä, mutta käytännön kannalta vaikeammin tulkittavissa neliöllisissä yksiköissä.

Esimerkki 6.2 (Keskiarvon yksi datajoukkoja)**.** Laske keskiarvo ja keskihajonta seuraaville datajoukoille:

$$\begin{aligned}
\vec{x} &= (1,1,1,1,1),\\
\vec{y} &= (0,0,1,1,2,2),\\
\vec{z} &= (0,2,0,2,0,2,0,2,0,2),\\
\vec{w} &= (\underbrace{0,0,0,0,\ldots,0,0,0,0}_{\text{666666 kpl}},1000000,\underbrace{0,0,\ldots,0,0}_{\text{333333 kpl}}).
\end{aligned}$$

Datajoukon $\vec{x}$ keskiarvo selvästikin on 1. Koska datajoukon $\vec{x}$ kaikki alkiot ovat ykkösiä, havaitaan kaavasta (6.2), että $\mathrm{sd}(\vec{x}) = 0$. Datajoukon $\vec{y}$ keskiarvo on

$$m(\vec{y}) \;=\; \frac{1}{6}\,(0+0+1+1+2+2) \;=\; 1$$

ja keskihajonta

$$\mathrm{sd}(\vec{y}) = \left(\frac{1}{6}\Big((0-1)^2+(0-1)^2+(1-1)^2+(1-1)^2+(2-1)^2+(2-1)^2\Big)\right)^{1/2},$$

jonka lukuarvoksi saadaan $\mathrm{sd}(\vec{y}) = \sqrt{2/3} \approx 0.8165$. Vastaavaan tapaan voidaan laskea myös datajoukkojen $\vec{z}$ ja $\vec{w}$ tunnusluvut. Datajoukoilla on sama keskiarvo, mutta eri keskihajonnat. Tulokset on esitetty alla olevassa taulukossa.

Datajoukko	Keskiarvo	Keskihajonta
$\vec{x}$	1	0.0000
$\vec{y}$	1	0.8165
$\vec{z}$	1	1.0000
$\vec{w}$	1	999.9995

Silloin kun käsiteltävä datajoukko edustaa pientä otosta suuresta populaatiosta[2], on usein tapana laskea keskihajonnan sijaan datajoukon *otoskeskihajonta* $\mathrm{sd}_s(\vec{x})$ (engl. sample standard deviation), joka saadaan korvaamalla luku $\frac{1}{n}$ luvulla $\frac{1}{n-1}$ kaavassa (6.2). Vastaavasti datajoukon *otosvarianssi* määritellään kaavalla $\mathrm{var}_s(\vec{x}) = \mathrm{sd}_s(\vec{x})^2$. Keskihajonta ja otoskeskihajonta voidaan muuntaa toisikseen kaavalla

$$\mathrm{sd}(\vec{x}) \;=\; \left(1 - \frac{1}{n}\right)^{1/2} \mathrm{sd}_s(\vec{x}), \tag{6.3}$$

josta myös nähdään, että suurille datajoukoille ei ole väliä, kumpaa keskihajonnan kaavaa käytetään. Tilasto-ohjelmistoja käytettäessä tulee kuitenkin varmistaa, kumpaa keskihajontaa ohjelmistot oletusarvoisesti laskevat. Liitteessä C on kuvattu, miten nämä lasketaan R:llä, Pythonilla ja Excelillä.

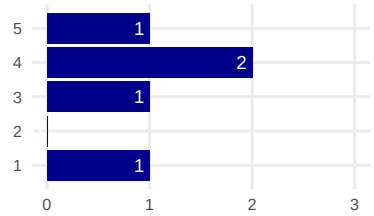
6.3 Empiirinen jakauma

Suuresta datajoukosta $\vec{x} = (x_1, \ldots, x_n)$ on hankala muodostaa mielikuvaa pelkästään tarkastelemalla sitä vastaavaa datakehikkoa. Toisaalta myös yksittäiset tunnusluvut antavat yleensä hyvin vajavaisen kuvan datajoukon olemuksesta. Silloin kannattaa tarkastella eri arvojen esiintyvyyksiä. Arvon x *esiintyvyys* eli frekvenssi

$$n_{\vec{x}}(x) \;=\; \#\{i : x_i = x\}$$

on datajoukossa $\vec{x} = (x_1, \ldots, x_n)$ arvoltaan x olevien alkioiden lukumäärä. Yksiulotteiselle datajoukolle eri arvojen esiintyvyydet on tapana raportoida esiintyvyystaulukkona tai vaakasuuntaisena palkkikaaviona. Esimerkin 6.1 datakehikon muuttujaa "Yleisarvio" vastaavan datajoukon $(5, 1, 4, 4, 3)$ esiintyvyystaulukko on esitetty alla.

x	1	2	3	4	5
$n_{\vec{x}}(x)$	1	0	1	2	1



Taulukko 6.1: Esimerkin 6.1 muuttujaa 'Yleisarvio' vastaavan datajoukon $(5, 1, 4, 4, 3)$ esiintyvyystaulukko ja sitä vastaava palkkikaavio.

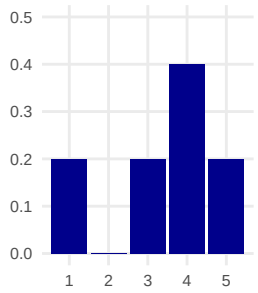
Kun halutaan vertailla arvojen esiintyvyyksiä erisuuruisissa datajoukoissa, on absoluuttisten lukumäärien sijaan suositeltavaa tarkastella suhteellisia esiintyvyyksiä. Arvon x *suhteellinen esiintyvyys*

$$f_{\vec{x}}(x) \;=\; \frac{n_{\vec{x}}(x)}{n} \tag{6.4}$$

[2]Syy tähän on perusteltu luvussa 6.8.

kertoo, mikä osuus datajoukon $\vec{x}$ alkioista on arvoltaan x. Suhteelliset esiintyvyydet on tapana raportoida taulukkona tai pylväskaaviona. Taulukossa 6.1 esitetyn datajoukon suhteelliset esiintyvyydet on esitetty alla.

x	1	2	3	4	5
$f_{\vec{x}}(x)$	$\frac{1}{5}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$



Taulukko 6.2: Esimerkin 6.1 muuttujaa 'Yleisarvio' vastaavan datajoukon $(5, 1, 4, 4, 3)$ suhteellinen esiintyvyystaulukko ja sitä vastaava pylväskaavio.

Yllä olevan taulukon suhteelliset esiintyvyydet ovat epänegatiivisia ja summautuvat ykköseksi. Tämä ominaisuus on voimassa yleisesti. Tästä seuraa, että kaavan (6.4) määrittämä funktio $f_{\vec{x}}(x)$ on jonkin diskreetin jakauman tiheysfunktio. Kyseinen diskreetti jakauma on datajoukon $\vec{x} = (x_1, \ldots, x_n)$ *empiirinen jakauma* ja funktio $f_{\vec{x}}(x)$ sitä vastaava *empiirinen tiheysfunktio*.

Seuraava tulos tarjoaa intuitiivisen tulkinnan datajoukon empiiriselle jakaumalle. Sen mukaan empiirinen jakauma voidaan tulkita todennäköisyysjakaumana satunnaismuuttujalle, joka saadaan valitsemalla datajoukosta yksi alkio tasaisen satunnaisesti. Datajoukon empiirinen tiheysfunktio $f_{\vec{x}}(x)$ kertoo siis todennäköisyyden, jolla datajoukosta tasaisen satunnaisesti valittu alkio on arvoltaan x.

Lause 6.3. *Datajoukosta $\vec{x} = (x_1, \ldots, x_n)$ tasaisen satunnaisesti valittu alkio X on diskreetti satunnaismuuttuja, joka noudattaa datajoukon $\vec{x}$ empiiristä jakaumaa tiheysfunktiona $f_X(x) = f_{\vec{x}}(x)$ ja toteuttaa*

$$\mathbb{E}(X) = m(\vec{x}), \tag{6.5}$$
$$\mathrm{SD}(X) = \mathrm{sd}(\vec{x}), \tag{6.6}$$
$$\mathrm{Var}(X) = \mathrm{var}(\vec{x}). \tag{6.7}$$

Lisäksi mielivaltaiselle funktiolle g pätee

$$\mathbb{E}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(x_i). \tag{6.8}$$

Todistus. Datajoukosta tasaisen satunnaisesti poimittu alkio voidaan kirjoittaa satunnaismuuttujana $X = x_I$, jossa satunnaismuuttuja I noudattaa indeksijoukon $\{1, \ldots, n\}$ tasajakaumaa. Satunnaismuuttuja X saa arvon x täsmälleen silloin, kun satunnaismuuttuja I kuuluu joukkoon $A = \{i : x_i = x\}$, jonka koko on $\#A = n_{\vec{x}}(x)$. Näin ollen X:n tiheysfunktio $f_X(x)$ toteuttaa

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}(I \in A) = \frac{\#A}{n} = \frac{n_{\vec{x}}(x)}{n} = f_{\vec{x}}(x).$$

Perustellaan seuraavaksi kaava (6.8). Satunnaismuuttujan muunnoksen odotusarvon yleistä laskukaavaa (3.4) ja empiirisen tiheysfunktion määritelmää (6.4) soveltamalla nähdään, että mielivaltaiselle funktiolle g pätee

$$\mathbb{E}[g(X)] \;=\; \sum_x g(x) f_X(x) \;=\; \sum_x g(x) f_{\vec{x}}(x) \;=\; \frac{1}{n} \sum_x g(x) n_{\vec{x}}(x).$$

Todetaan seuraavaksi, että

$$\sum_x g(x) n_{\vec{x}}(x) \;=\; \sum_{i=1}^{n} g(x_i),$$

sillä $n_{\vec{x}}(x)$ kertoo lukumäärän, kuinka monta kertaa arvo x esiintyy summassa $\sum_{i=1}^{n} g(x_i)$. Yhdistämällä kaksi yllä olevaa yhtälöä saadaan todistetuksi (6.8).

Perustellaan seuraavaksi kaavat (6.5)–(6.7). Sijoittamalla $g(x) = x$ kaavaan (6.8) havaitaan, että

$$\mathbb{E}(X) \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i \;=\; m(\vec{x}).$$

Näin ollen on todistettu yhtälö (6.5). Soveltamalla kaavaa (6.8) uudelleen, tällä kertaa funktioon $g(x) = (x - m(\vec{x}))^2$, tuottaa tulokseksi

$$\mathbb{E}\Big((X - \mathbb{E}(X))^2\Big) \;=\; \mathbb{E}\Big((X - m(\vec{x}))^2\Big) \;=\; \mathbb{E}\big(g(X)\big) \;=\; \frac{1}{n} \sum_{i=1}^{n} (x_i - m(\vec{x}))^2.$$

Tämä tarkoittaa, että kaava (6.7) pätee. Kaava (6.6) seuraa tästä ottamalla neliöjuuret molemmin puolin. □

Esimerkki 6.4. Määritä empiirinen jakauma ja laske sen avulla keskiarvo ja keskihajonta datajoukolle $\vec{y} = (0, 0, 1, 1, 2, 2)$.

Datajoukon koko on $n = 6$ ja kaikkien siihen sisältyvien arvojen esiintyvyys on $n_{\vec{y}}(y) = 2$. Arvojen suhteelliset esiintyvyydet voidaan siis taulukoida muodossa:

y	0	1	2
$f_{\vec{y}}(y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Tiheysfunktion $f_{\vec{y}}(y)$ mukaan jakautuneelle satunnaismuuttujalle Y pätee odotusarvon ja varianssin yleisten määritelmien (3.1) ja (4.1) perusteella

$$\mathbb{E}(Y) \;=\; \sum_{y=0}^{2} y f_{\vec{y}}(y) \;=\; 0 \times \frac{1}{3} + 1 \times \frac{1}{3} + 2 \times \frac{1}{3} \;=\; 1$$

ja

$$\mathrm{Var}(Y) \;=\; \sum_{y=0}^{2} (y-1)^2 f_{\vec{y}}(y) \;=\; (0-1)^2 \times \frac{1}{3} + (1-1)^2 \times \frac{1}{3} + (2-1)^2 \times \frac{1}{3} \;=\; \frac{2}{3},$$

joten Y:n keskihajonta on $\mathrm{SD}(Y) = \sqrt{\mathrm{Var}(Y)} = \sqrt{\frac{2}{3}} \approx 0.8165$. Lauseen 6.3 perusteella datajoukon $\vec{y}$ keskiarvoksi saadaan $m(\vec{y}) = 1$ ja keskihajonnaksi $\mathrm{sd}(\vec{y}) \approx 0.8165$. Sama tulos saatiin esimerkissä 6.2 laskettua pidemmällä tavalla suoraan määritelmästä (6.2).

Todetaan myös, että tulos on sama, mitä esimerkin 4.1 satunnaismuuttujalle Y laskettiin esimerkissä 4.2. Näin pitääkin olla, sillä datajoukon $\vec{y}$ empiirinen jakauma on sama kuin esimerkin 4.1 satunnaismuuttujan Y jakauma. ■

6.4 Kahden muuttujan datajoukon tunnuslukuja

Kahden muuttujan datajoukko on järjestetty lista pareja, jota tässä monisteessa merkitään

$$\vec{xy} \;=\; ((x_1, y_1), \ldots, (x_n, y_n)).$$

Vaihtoehtoisesti voidaan kahden muuttujan datajoukko voidaan tulkita myös parina $(\vec{x}, \vec{y})$, jossa $\vec{x} = (x_1, \ldots, x_n)$ ja $\vec{y} = (y_1, \ldots, y_n)$ ovat samankokoisia yhden muuttujan datajoukkoja. Kahden muuttujan datajoukkoa voidaan kuvailla laskemalla yksittäisten muuttujien keskiarvot $m(\vec{x})$ ja $m(\vec{y})$ sekä keskihajonnat $\mathrm{sd}(\vec{x})$ ja $\mathrm{sd}(\vec{y})$. Nämä eivät kuitenkaan kerro mitään datajoukon muuttujien yhteisvaihtelusta tai keskinäisistä riippuvuuksista. Muuttujien yhteisvaihtelua kuvaava tunnusluku on datajoukon *kovarianssi*

$$\mathrm{cov}(\vec{x}, \vec{y}) \;=\; \frac{1}{n} \sum_{i=1}^{n} (x_i - m(\vec{x}))(y_i - m(\vec{y})), \tag{6.9}$$

joka mittaa x- ja y-muuttujien yhteisvaihtelun suuntaa ja voimakkuutta. Kuten varianssi, myös kovarianssi mittaa vaihtelua käytännön kannalta hankalisti tulkittavissa neliöllisissä yksiköissä. Kovarianssi normitetaan jakamalla yllä oleva lauseke datajoukkojen keskihajonnoilla, jolloin saadaan datajoukon *korrelaatio*

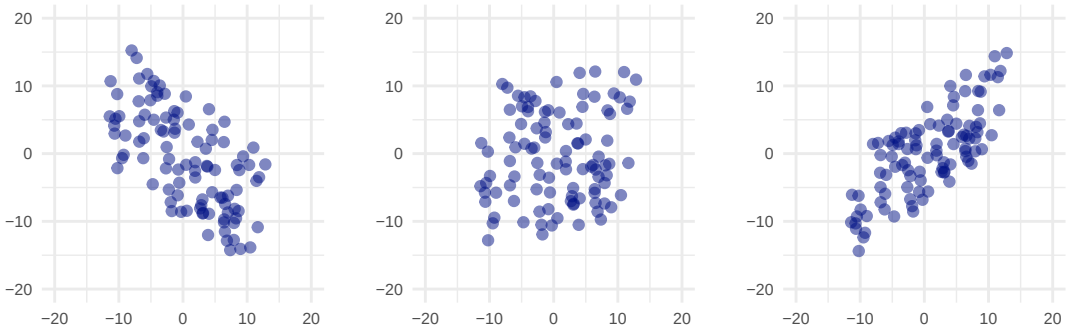
$$\mathrm{cor}(\vec{x}, \vec{y}) \;=\; \frac{\mathrm{cov}(\vec{x}, \vec{y})}{\mathrm{sd}(\vec{x})\,\mathrm{sd}(\vec{y})}. \tag{6.10}$$

Voidaan todistaa, että että korrelaatio aina toteuttaa ehdot

$$-1 \;\leq\; \mathrm{cor}(\vec{x}, \vec{y}) \;\leq\; +1.$$

Datajoukon korrelaatiota kutsutaan myös nimellä Pearsonin korrelaatiokerroin erotuksena muista, järjestyslukuihin perustuvista korrelaatiokertoimista.

Kaksiulotteinen datajoukko voidaan visualisoida hajontakuviona piirtämällä datajoukon lukuparit (x, y)-tasoon. Alla on esitetty hajontakaaviot kolmelle kaksiulotteiselle sadan alkion datajoukolle sekä niiden korrelaatiot.



$\text{cor}(\vec{x}, \vec{y}) = -0.645$ $\text{cor}(\vec{x}, \vec{y}) = 0.144$ $\text{cor}(\vec{x}, \vec{y}) = 0.751$

Aivan kuten keskihajonnankin kohdalla, silloin kun käsiteltävä datajoukko edustaa pientä otosta suuresta populaatiosta[3], on usein tapana laskea kovarianssin sijaan *otoskovarianssi* $\text{cov}_s(\vec{x}, \vec{y})$ (engl. sample covariance), joka saadaan korvaamalla luku $\frac{1}{n}$ luvulla $\frac{1}{n-1}$ kaavassa (6.9). Datajoukon kovarianssi ja otoskovarianssi voidaan muuntaa toisikseen kaavalla

$$\text{cov}_s(\vec{x}, \vec{y}) \;=\; \left(1 - \frac{1}{n}\right) \text{cov}(\vec{x}, \vec{y}),$$

josta myös nähdään, että suurille datajoukoille ei ole väliä, kumpaa kovarianssin kaavaa käytetään. Tilasto-ohjelmistoja käytettäessä tulee kuitenkin huolellisesti varmistaa, kumpaa keskihajontaa ohjelmistot oletusarvoisesti laskevat. Liitteessä C on kuvattu, miten nämä lasketaan R:llä, Pythonilla ja Excelillä.

6.5 Ristitaulukko ja empiirinen yhteisjakauma

Kahden muuttujan datajoukkoja voidaan tehokkaasti kuvailla laskemalla arvoparien esiintyvyydet. Arvoparin (x, y) *esiintyvyys* (engl. frequency)

$$n_{\vec{xy}}(x, y) \;=\; \#\{i : x_i = x \text{ ja } y_i = y\}$$

on datajoukossa arvoltaan (x, y) olevien alkioiden lukumäärä. Esimerkin 6.1 muuttujat "Yleisarvio" ja "Hyödyllisyys" voidaan koostaa datajoukoksi ((5,5), (1,2), (4,3), (4,3), (3,3)). Sen arvoparien esiintyvyydet voidaan taulukoida muodossa

	y					
x	1	2	3	4	5	Yht
1	0	1	0	0	0	1
2	0	0	0	0	0	0
3	0	0	1	0	0	1
4	0	0	2	0	0	2
5	0	0	0	0	1	1
Yht	0	1	3	0	1	

[3]Syy tähän on perusteltu luvussa 6.8.

Yllä oleva esitys on muuttujien x ja y esiintyvyyksien *ristitaulukko* (engl. contingency table) ja tällaista esitysmenetelmää kutsutaan ristiintaulukoimiseksi (engl. cross tabulation). Ristitaulukon rivisummista saadaan muuttujan x esiintyvyydet (vrt. taulukko 6.1) ja sarakesummista muuttujan y esiintyvyydet.

Arvoparin (x, y) *suhteellinen esiintyvyys* datajoukossa $x\vec{y}$ määritellään kaavalla.

$$f_{x\vec{y}}(x, y) = \frac{n_{x\vec{y}}(x, y)}{n}. \tag{6.11}$$

Datajoukon ((5,5), (1,2), (4,3), (4,3), (3,3)) suhteelliset esiintyvyydet voidaan taulukoida muodossa

	y					
x	1	2	3	4	5	Yht
1	0	$\frac{1}{5}$	0	0	0	$\frac{1}{5}$
2	0	0	0	0	0	0
3	0	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$
4	0	0	$\frac{2}{5}$	0	0	$\frac{2}{5}$
5	0	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$
Yht	0	$\frac{1}{5}$	$\frac{3}{5}$	0	$\frac{1}{5}$	

Aivan kuin yksiulotteisillekin datajoukoille, myös kaksiulotteisen datajoukon suhteelliset esiintyvyydet $f_{x\vec{y}}(x, y)$ ovat epänegatiivisia ja summautuvat ykköseksi. Näin ollen ylläoleva taulukko vastaa erään diskreetin yhteisjakauman tiheysfunktiota. Kyseinen diskreetti jakauma on datajoukon $x\vec{y}$ *empiirinen yhteisjakauma*, ja kaavan (6.11) määrittämä funktio $f_{x\vec{y}}(x, y)$ sitä vastaava tiheysfunktio. Empiirisen yhteisjakauman rivisummista saadaan datajoukon $(x_1, \ldots, x_n)$ empiirinen jakauma (vrt. taulukko 6.2) ja sarakesummista datajoukon $(y_1, \ldots, y_n)$ empiirinen jakauma.

Seuraava tulos tarjoaa todennäköisyystulkinnan empiiriselle yhteisjakaumalle. Sen mukaan empiirinen jakauma voidaan tulkita datajoukosta satunnaisotannalla valitun parin yhteisjakaumana, jolloin empiirinen tiheysfunktio $f(x, y)$ kertoo todennäköisyyden, jolla datajoukosta satunnaisesti valitun parin arvot ovat x ja y.

Lause 6.5. *Datajoukosta $x\vec{y} = ((x_1, y_1), \ldots, (x_n, y_n))$ tasaisen satunnaisesti valittu pari (X, Y) on diskreetti satunnaismuuttuja, joka noudattaa datajoukon $x\vec{y}$ empiiristä jakaumaa tiheysfunktiona $f_{X,Y}(x, y) = f_{x\vec{y}}(x, y)$ ja toteuttaa*

$$\begin{aligned} \mathbb{E}(X) &= m(\vec{x}), & \mathbb{E}(Y) &= m(\vec{y}), \\ \mathrm{SD}(X) &= \mathrm{sd}(\vec{x}), & \mathrm{SD}(Y) &= \mathrm{sd}(\vec{y}), \\ \mathrm{Var}(X) &= \mathrm{var}(\vec{x}), & \mathrm{Var}(Y) &= \mathrm{var}(\vec{y}), \end{aligned} \tag{6.12}$$

sekä

$$\mathrm{Cor}(X, Y) = \mathrm{cor}(\vec{x}, \vec{y}), \tag{6.13}$$

$$\mathrm{Cov}(X, Y) = \mathrm{cov}(\vec{x}, \vec{y}). \tag{6.14}$$

Lisäksi mielivaltaiselle kahden muuttujan funktiolle g pätee

$$\mathbb{E}[g(X,Y)] \;=\; \frac{1}{n}\sum_{i=1}^{n} g(x_i,y_i). \tag{6.15}$$

Todistus. Tarkastelun kohteena oleva datajoukko $\vec{xy}$ voidaan tulkita yksiulotteisena datajoukkona $\vec{z}=(z_1,\ldots,z_n)$, jonka alkiot koostuvat lukupareista $z_i=(x_i,y_i)$. Satunnaisesti valittu lukupari puolestaan voidaan esittää satunnaismuuttujana $Z=(X,Y)$, joka saadaan valitsemalla tasaisen satunnaisesti alkio listasta $\vec{z}=(z_1,\ldots,z_n)$. Tällöin lauseen 6.3 mukaan

$$f_Z(z) \;=\; f_{\vec{z}}(z),$$

jossa $f_{\vec{z}}(z)$ on arvon z suhteellinen esiintyvyys datajoukossa $(z_1,\ldots,z_n)$. Koska lukuparin $z=(x,y)$ suhteelliselle esiintyvyydelle pätee $f_{\vec{z}}(z) = f_{\vec{xy}}(x,y)$, havaitaan tästä että

$$f_{X,Y}(x,y) \;=\; \mathbb{P}(X=x,Y=y) \;=\; \mathbb{P}(Z=z) \;=\; f_{\vec{z}}(z) \;=\; f_{\vec{xy}}(x,y).$$

Satunnaisen lukuparin (X,Y) jakauma on siis datajoukon empiirinen yhteisjakauma $f_{\vec{xy}}(x,y)$.

Todistetaan seuraavaksi kaava (6.15). Tulkitaan $g(x,y)$ yhden muuttujan funktiona $\tilde{g}(z)=g(x,y)$, jonka syötteenä ovat lukuparit muotoa $z=(x,y)$. Soveltamalla kaavaa (6.8) datajoukosta $\vec{z}=(z_1,\ldots,z_n)$ satunnaisesti poimittuun alkioon Z havaitaan, että

$$\mathbb{E}[g(X,Y)] \;=\; \mathbb{E}[\tilde{g}(Z)] \;=\; \frac{1}{n}\sum_{i=1}^{n}\tilde{g}(z_i) \;=\; \frac{1}{n}\sum_{i=1}^{n} g(x_i,y_i).$$

Näin ollen kaava (6.15) on tosi.

Kaavat (6.12) voidaan todistaa sijoittamalla kaavaan (6.15) ensiksi funktiot $g(x,y)=x$ ja $g(x,y)=y$, ja etenemällä sen jälkeen samaan tapaan kuin lauseen 6.3 todistuksessa. Kun kaavaa (6.15) sovelletaan funktioon $g(x,y) = (x-m(\vec{x})(y-m(\vec{y}))$, saadaan tulokseksi

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}(X))]\mathbb{E}[(Y-\mathbb{E}(Y))] \\
&= \mathbb{E}[(X-m(\vec{x}))]\mathbb{E}[(Y-m(\vec{y}))] \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-m(\vec{x})(y_i-m(\vec{y})) \\
&= \mathrm{cov}(\vec{x},\vec{y}).
\end{aligned}$$

Näin ollen kaava (6.14) pätee. Jakamalla yllä olevan yhtälön molemmat puolet termillä $\mathrm{SD}(X)\,\mathrm{SD}(Y) = \mathrm{sd}(\vec{x})\,\mathrm{sd}(\vec{y})$ havaitaan, että myös kaava (6.13) pätee. □

6.6 Kvantiilit

Lukuarvoisen datajoukon *kvantiili* tasolla $p \in (0,1)$ on tunnusluku $Q(p)$, jonka avulla pilkotaan datajoukko kahtia niin, että alkioista suurin piirtein osuus p sijaitsee luvun $Q(p)$ alapuolella ja loput alkioista luvun $Q(p)$ yläpuolella. Tasojen 0.25, 0.5 ja 0.75 kvantiileja kutsutaan *kvartiileiksi* ja ne tunnetaan nimillä *alakvartiili*, *mediaani* ja *yläkvartiili*. Tasojen $0.01, 0.02, \ldots$ kvantiileja puolestaan kutsutaan *prosentiileiksi*. Yleisesti ottaen kvantiilit määritellään järjestämällä datajoukon $(x_1, \ldots, x_n)$ alkiot suuruusjärjestykseen muodossa

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Luku $x_{(k)}$ on datajoukon k:nnes *järjestystunnusluku*. Tason $p \in (0,1)$ kvantiili määritellään yleensä järjestystunnuslukujen painotettuna keskiarvona

$$Q(p) \;=\; (1-\gamma)x_{(j)} + \gamma x_{(j+1)},$$

jossa[4] $j = \lfloor np + (1-p) \rfloor$ ja $\gamma = np + (1-p) - j$. Ylläoleva kuvaus tulkittuna p:n funktioksi on datajoukon *kvantiilifunktio*[5]. Kvantiilifunktion voi tulkita helpoiten piirtämällä sen kuvaaja seuraavasti:

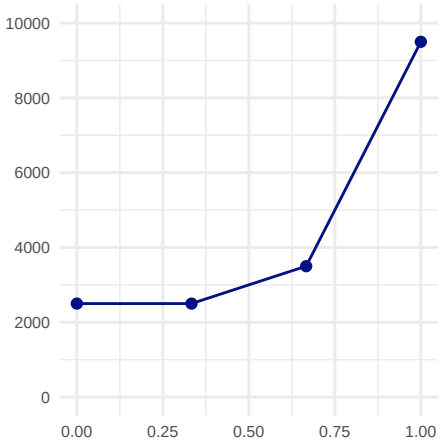
- Jaetaan vaaka-akselin yksikköväli $n-1$ tasapituiseen väliin päätepisteinä luvut $p_k = (k-1)/(n-1)$, $k = 1, \ldots, n$.
- Piirretään tasoon pisteet $(p_k, x_{(k)})$ ja yhdistetään ne viivoilla.

Esimerkki 6.6. Pienessä yrityksessä työskentelee neljä henkilöä, joiden bruttopalkat ovat 2500, 3500, 2500, 9500 (eur/kk). Laske bruttopalkkojen järjestystunnusluvut, piirrä kvantiilifunktio, ja määritä kvantiilifunktion avulla palkkajakauman alakvartiili, mediaani ja yläkvartiili.

Datajoukon $(2500, 3500, 2500, 9500)$ järjestystunnusluvut ovat $x_{(1)} = 2500$, $x_{(2)} = 2500$, $x_{(3)} = 3500$ ja $x_{(4)} = 9500$. Jaetaan vaaka-akselin yksikköväli kolmeen yhtäpitkään osaväliin päätepisteinä $p_1 = 0$, $p_2 = \frac{1}{3}$, $p_3 = \frac{2}{3}$ ja $p_4 = 1$. Kvantiilifunktion kuvaaja saadaan piirtämällä tasoon pisteet $(p_1, x_{(1)}), \ldots (p_4, x_{(4)})$ ja yhdistämällä ne viivoilla.

[4] $\lfloor x \rfloor$ on luku x pyöristettynä alaspäin kokonaisluvuksi.

[5] Kvantiilifunktio määritellään eri yhteyksissä hieman eri tavoin, esim. R-ohjelmisto tarjoaa kahdeksan vaihtoehtoista tapaa kvantiilifunktion laskemiseen.



Kvantiilifunktion kuvaajasta luetaan: alakvartiili $Q(0.25) = 2500$, mediaani $Q(0.5) = 3000$ ja yläkvartiili $Q(0.75) = 5000$. Tässä datajoukossa mediaani 3000 on reilusti pienempi kuin keskiarvo 4500. ■

6.7 Histogrammi

Silloin kun datajoukko sisältää suuren määrän arvoja, saattaa tarkka esiintyvyystaulukko tai empiirinen jakauma olla liian yksityiskohtainen, jotta sen voisi selkeästi hahmottaa. Tällöin on tapana karkeistaa dataa osittamalla arvojoukko pienempään määrään lukuvälejä. Näin saadaan datajoukon luokiteltu esiintyvyystaulukko. Luokitellun esiintyvyystaulukon suhteellisia osuuksia esittävä kuvaaja on datajoukon *histogrammi*. Histogrammi piirretään yleensä näin:

- Yksi pylväs per luokka
- Pylvään leveys = luokkavälin leveys
- Pylvään korkeus = datapisteiden suhteellinen osuus jaettuna palkin leveydellä

Seuraava esimerkki valaisee asiaa.

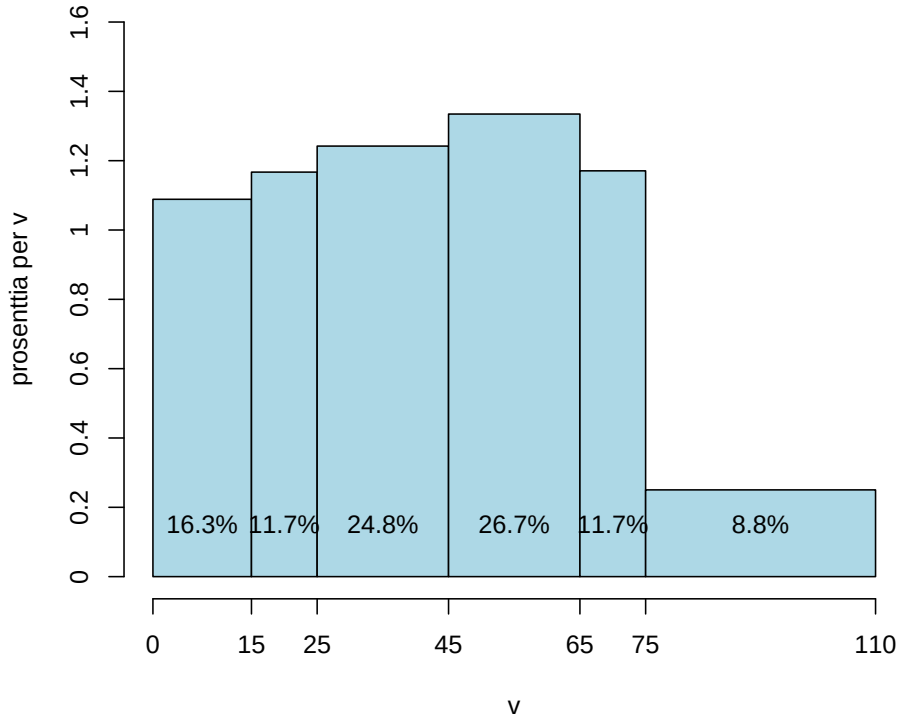
Esimerkki 6.7 (Suomalaisten ikärakenne)**.** Suomalaisten ikärakenne 31.12.2015 sisältää $n = 5\ 487\ 308$ miljoonaa datapistettä[6]. Ei ole järkeä piirtää jokaista pistettä kuvaajaan, vaan jaetaan datapisteet luokkiin.

Ikä (v)	Lukumäärä
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

[6]Lähde: Tilastokeskus

Esim: Suomalaiset

- 1. pylväs käsittää suomalaiset, joiden ikä on 0–14 vuotta
- 1. pylvään leveys = 15 v
- Datapisteiden lkm luokassa 1 on 896023 ja suhteellinen osuus $896023/5487308 \approx 16.3\%$
- Pylvään korkeus = $16.3/15 \approx 1.09$ (yksikkönä % per vuosi).



6.8 Kommentteja ja lisätietoa

Alla on todistettu tulos, joka selittää, miksi pienille datajoukoille on toisinaan tapana laskea varianssin ja kovarianssin sijaa otosvarianssi ja otoskovarianssi.

Lause 6.8. *Jos* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *ovat keskenään riippumattomia ja satunnaisen lukuparin* (X, Y) *kanssa samoin jakautuneita satunnaislukupareja, niin datajoukon* $\vec{X}\vec{Y}$ *kovarianssille pätee*

$$\mathbb{E}\left[\operatorname{cov}(\vec{X}, \vec{Y})\right] \;=\; \left(1 - \frac{1}{n}\right) \operatorname{Cov}(X, Y).$$

Vastaavasti

$$\mathbb{E}\left[\mathrm{var}(\vec{X})\right] \;=\; \left(1-\frac{1}{n}\right)\mathrm{Var}(X).$$

Todistus. Datajoukkojen kovarianssille pätee yleisesti kaava

$$\mathrm{cov}(\vec{X},\vec{Y}) \;=\; \frac{1}{n}\sum_{i=1}^{n} X_iY_i - m(\vec{X})m(\vec{Y}). \tag{6.16}$$

Lasketaan seuraavaksi odotusarvot yhtälön (6.16) oikealla puolella esiintyville termeille. Odotusarvon lineaarisuuden mukaan ensimmäinen termi on odotusarvoltaan

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_iY_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_iY_i) \;=\; \mathbb{E}(XY). \tag{6.17}$$

Yhtälön (6.16) oikeanpuolimmaisen termin odotusarvo pitää analysoida huolellisesti, sillä satunnaismuuttujat $m(\vec{X})$ ja $m(\vec{Y})$ ovat yleisesti ottaen toisistaan riippuvia. Näiden tulo voidaan avata muotoon

$$m(\vec{X})m(\vec{Y}) \;=\; \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j:j\neq i} X_iY_j + \frac{1}{n^2}\sum_{i=1}^{n} X_iY_i.$$

Riippumattomuuden nojalla $\mathbb{E}(X_iY_j) = \mathbb{E}(X_i)\mathbb{E}(Y_j)$ aina kun $j \neq i$. Näin ollen odotusarvon lineaarisuutta käyttämällä

$$\begin{aligned}\mathbb{E}\Big[m(\vec{X})m(\vec{Y})\Big] \;&=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j:j\neq i}\mathbb{E}(X_iY_j) + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(X_iY_i)\\ &=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j:j\neq i}\mathbb{E}(X_i)\mathbb{E}(Y_j) + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(X_iY_i),\end{aligned}$$

josta seuraa, että

$$\mathbb{E}\Big[m(\vec{X})m(\vec{Y})\Big] \;=\; \left(1-\frac{1}{n}\right)\mathbb{E}(X)\mathbb{E}(Y) + \frac{1}{n}\mathbb{E}(XY). \tag{6.18}$$

Yhdistämällä kaavat (6.17)–(6.18) kaavaan (6.16) havaitaan nyt, että

$$\begin{aligned}\mathbb{E}\,\mathrm{cov}(\vec{X},\vec{Y}) \;&=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_iY_i\right) - \mathbb{E}\Big[m(\vec{X})m(\vec{Y})\Big]\\ &=\; \mathbb{E}(XY) - \Big(1-\frac{1}{n}\Big)\mathbb{E}(X)\mathbb{E}(Y) - \frac{1}{n}\mathbb{E}(XY)\\ &=\; \left(1-\frac{1}{n}\right)\Big(\mathbb{E}(XY) - \mathbb{E}(XY)\Big).\end{aligned}$$

Tästä seuraa väite, sillä $\mathbb{E}(XY) - \mathbb{E}(XY) = \mathrm{Cov}(X,Y)$ kaavan (4.10) perusteella.

Datajoukon varianssia koskeva väite seuraa kovarianssia koskevasta tuloksesta, kun huomataan että $\mathrm{var}(\vec{x}) = \mathrm{cov}(\vec{x},\vec{x})$ ja $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$. □

Luku 7

Parametrien estimointi

7.1 Parametriset jakaumat

Tarkastellaan tuntematonta datalähdettä, joka tuottaa toisistaan stokastisesti riippumattomia ja tiheysfunktion $f(x)$ mukaan jakautuneita satunnaislukuja. Yleensä tiheysfunktiota ei tunneta, mutta toisinaan tiheysfunktion rakenteellinen muoto voidaan kuitenkin päätellä kontekstista. Esimerkiksi binaarisen datalähteen tiheysfunktio tunnetaan yhtä parametria vaille, kuten alla oleva esimerkki vahvistaa.

Esimerkki 7.1 (Binaarinen datalähde)**.** Binaariselle $\{0, 1\}$-arvoisia satunnaislukuja tuottavalle datalähteelle pätee $f(x) = 0$ aina kun $x \neq 0$ tai $x \neq 1$. Koska diskreetin jakauman tiheysfunktion arvot summautuvat ykköseksi, voidaan tästä päätellä että $f(0) = 1 - f(1)$. Tiheysfunktio voidaan siis kirjoittaa muodossa

$$f(x) \;=\; \begin{cases} 1-p, & x = 0, \\ p, & x = 1, \\ 0, & \text{muuten}, \end{cases}$$

missä $p = f(1)$ on arvon 1 todennäköisyys. Ylläolevan tiheysfunktion määrittämä jakauma on *Bernoulli-jakauma* parametrina p. Yllä tehdyn päättelyn mukaan siis jokainen $\{0, 1\}$-arvoinen datalähde noudattaa Bernoulli-jakaumaa. Bernoulli-jakaumaa parametrina p noudattavan satunnaisluvun X odotusarvo on

$$\mathbb{E}(X) \;=\; \sum_x x f(x) \;=\; 0 \times (1-p) + 1 \times p \;=\; p. \tag{7.1}$$

Koska myös $\mathbb{E}(X^2) = \sum_x x^2 f(x) = p$, saadaan keskihajonnaksi kaavan (4.2) avulla

$$\mathrm{SD}(X) \;=\; \left(\mathbb{E}(X^2) - (\mathbb{E}(X))^2\right)^{1/2} \;=\; (p(1-p))^{1/2}. \tag{7.2}$$

■

Taulukkoon 7.1 on listattu tärkeimpiä yhden muuttujan jakaumia, joissa parametrien lukumäärä on yksi tai kaksi. Kun datalähteen tiheysfunktio tunnetaan tiettyjä parametreja vaille ja datalähteestä on havaittu arvot $x_1, x_2, \ldots, x_n$,

Malli	**Parametrit**	**Arvojoukko**	**Tiheysfunktio**
Bernoullijakauma	p	$\{0,1\}$	$f_p(x) = (1-p)^{1-x}p^x$
Binomijakauma	n,p	$\{0,1,\ldots,n\}$	$f_{n,p}(x) = \binom{n}{x}(1-p)^{n-x}p^x$
Eksponenttijakauma	λ	$(0,\infty)$	$f_\lambda(x) = \lambda e^{-\lambda x}$
Jatkuva tasajakauma	a,b	$[a,b]$	$f_{a,b}(x) = \frac{1}{b-a}$
Normaalijakauma	μ,σ	$(-\infty,\infty)$	$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Taulukko 7.1: Yhden muuttujan parametrisia jakaumia.

jää tehtäväksi määrittää tuntemattomien parametrien arvot. Kun havaittuja arvoja on rajallinen määrä, on parametrien tarkka määrittäminen mahdotonta. Tällöin paras, mitä voidaan tehdä, on muodostaa tuntemattomille parametriarvoille valistunut arvaus. Systemaattisia menetelmiä valistuneiden arvausten muodostamiseksi kutsutaan parametrien estimoinniksi.

7.2 Suurimman uskottavuuden estimointi

Tarkastellaan datalähdettä, josta on havaittu lukuarvot $x_1,\ldots,x_n$, ja jonka jakauman tiheysfunktio oletetaan parametria θ vaille tunnetuksi. Kun halutaan muodostaa valistunut arvaus tuntemattoman parametrin θ arvolle, voidaan tarkastella miten tiheysfunktio

$$f(x_1,\ldots,x_n \,|\, \theta)$$

käyttäytyy parametrin eri arvoilla. Tässä kohtaa on muodostettu uusi näkökulma tiheysfunktion olemukseen. Nimittäin lukuarvot $x_1,\ldots,x_n$ ovat nyt tunnettuja ja parametri θ on tuntematon. Näin tulkittua tiheysfunktiota kutsutaan parametrin θ *uskottavuusfunktioksi* (engl. likelihood function) ja sitä merkitään

$$L(\theta) \;=\; f(x_1,\ldots,x_n \,|\, \theta).$$

Mitä suurempi ylläolevan uskottavuusfunktion arvo on, sitä enemmän on aihetta uskoa, että havaitut lukuarvot $x_1,\ldots,x_n$ ovat peräisin parametrin θ mukaisesta datalähteestä. Silloin kun datalähteen tuottamat satunnaismuuttujat voidaan olettaa toisistaan riippumattomiksi, voidaan uskottavuusfunktio kirjoittaa muodossa

$$L(\theta) \;=\; f(x_1 \,|\, \theta)\cdots f(x_n \,|\, \theta). \tag{7.3}$$

Luonnollinen tapa tuntemattoman parametrin estimoimiseksi on etsiä parametri, jolle uskottavuusfunktion arvo on suurin mahdollinen. Näin saatu luku θ^* on parametrin θ *suurimman uskottavuuden estimaatti* havaitun datajoukon $(x_1,\ldots,x_n)$ suhteen.

Esimerkki 7.2 (Jatkuva tasajakauma)**.** Datalähteen satunnaisluvut noudattavat jatkuvan välin $[0, \theta]$ tasajakaumaa, jonka parametri θ on tuntematon. On havaittu arvot $x_1, \ldots, x_n$. Määritä suurimman uskottavuuden estimaatti parametrille θ.

Yksittäinen satunnaisluku noudattaa jatkuvan välin $[0, \theta]$ tasajakaumaa tiheysfunktiona

$$f(x \,|\, \theta) \;=\; \begin{cases} \theta^{-1}, & x \in [0, \theta], \\ 0, & \text{muuten}, \end{cases}$$

joten parametrin θ uskottavuusfunktion havaitun datajoukon $(x_1, \ldots, x_n)$ suhteen on

$$L(\theta) \;=\; \prod_{i=1}^{n} f(x_i \,|\, \theta) \;=\; \begin{cases} \theta^{-n}, & x_1, \ldots, x_n \in [0, \theta], \\ 0, & \text{muuten}. \end{cases}$$

Tämä voidaan myös kirjoittaa muodossa

$$L(\theta) \;=\; \begin{cases} \theta^{-n}, & \theta \geq \max\{x_1, \ldots, x_n\}, \\ 0, & \text{muuten}. \end{cases}$$

Luonnostelemalla funktion $L(\theta)$ kuvaaja havaitaan, että $L(\theta)$ saa suurimman arvonsa pisteessä $\theta^* = \max\{x_1, \ldots, x_n\}$. Tämä on parametrin θ suurimman uskottavuuden estimaatti. ■

Uskottavuusfunktion $L(\theta)$ maksimointi on usein helpompaa logaritmisen muunnoksen avulla. Parametrin θ *logaritminen uskottavuusfunktio* määritellään kaavalla

$$\ell(\theta) \;=\; \log L(\theta),$$

missä log tarkoittaa luonnollista logaritmia. Koska logaritmi on aidosti kasvava funktio, saavuttaa $L(\theta)$ maksiminsa samoissa pisteissä, joissa $\ell(\theta)$ saavuttaa oman maksiminsa.

Esimerkki 7.3 (Hirmumyrskyt)**.** Eräälle trooppiselle saarelle on 2000-luvulla iskenyt hirmumyrsky vuosina 2000, 2009, 2011 ja 2017. Saarelle saapuvien hirmumyrskyjen väliaikoja (vuosina) mallinnetaan käyttämällä lukujoukon $\{1, 2, \ldots\}$ geometrista jakaumaa tiheysfunktiona

$$f(x \,|\, \theta) \;=\; (1 - \theta)^{x-1}\theta.$$

Määritä parametrin θ suurimman uskottavuuden estimaatti ja ennusta sen avulla, millä todennäköisyydellä saarelle iskee seuraava hirmumyrsky viimeistään vuonna 2020.

Parametrin θ uskottavuusfunktio havaittujen väliaikojen $x_1 = 9$, $x_2 = 2$ ja $x_3 = 6$ suhteen on

$$\begin{aligned} L(\theta) \;&=\; f(9 \,|\, \theta) f(2 \,|\, \theta) f(6 \,|\, \theta) \\ &=\; (1 - \theta)^{9-1}\theta \times (1 - \theta)^{2-1}\theta \times (1 - \theta)^{6-1}\theta \\ &=\; (1 - \theta)^{14}\theta^3. \end{aligned}$$

Vastaava logaritminen uskottavuusfunktio $\ell(\theta) = \log L(\theta)$ on näin ollen

$$\ell(\theta) \;=\; 14\log(1-\theta) + 3\log\theta,$$

jonka derivaatta on

$$\ell'(\theta) \;=\; -14\frac{1}{1-\theta} + 3\frac{1}{\theta}.$$

Derivaatan nollakohta löytyy pisteestä $\theta = \frac{3}{17}$. Tämä on logaritmisen uskottavuusfunktion ja näin ollen myös uskottavuusfunktion maksimikohta, sillä $\ell''(\theta) \leq 0$. Näin ollen suurimman uskottavuuden estimaatti on $\theta^* = \frac{3}{17}$.

Kun seuraavan hirmumyrskyn saapumisaikaa merkitään satunnaismuuttujalla X, on todennäköisyys että seuraava hirmumyrsky iskee viimeistään vuonna 2020

$$\mathbb{P}(X \leq 3) \;=\; \sum_{x=1}^{3} f(x\,|\,\theta) \;=\; \theta + (1-\theta)\theta + (1-\theta)^2\theta.$$

Sijoittamalla tähän $\theta = \frac{3}{17}$ saadaan ennusteeksi $\mathbb{P}(X \leq 3) \approx 0.44$. ■

7.3 Binaarimallin estimointi

Tarkastellaan datalähdettä, joka tuottaa toisistaan riippumattomia $\{0,1\}$-arvoisia satunnaislukuja. Esimerkin 7.1 mukaan satunnaisluvut noudattavat Bernoullijakaumaa parametrina $p = f(1)$ ja tiheysfunktiona

$$f(x\,|\,p) \;=\; \begin{cases} 1-p, & x = 0, \\ p, & x = 1, \\ 0, & \text{muuten.} \end{cases}$$

Seuraava tulos kertoo, miten suurimman uskottavuuden estimaatti lasketaan binaarimallille. Suurimman uskottavuuden estimoinnin näkökulmasta ei ole väliä, miten nollat ja ykköset ovat sijoittuneet datajoukossa $(x_1, \ldots, x_n)$, vaan riittää tietää ykkösten suhteellinen osuus.

Lause 7.4. *Binaariselle $\{0,1\}$-arvoiselle datalähteelle parametrin $p = f(1)$ suurimman uskottavuuden estimaatti datajoukon $\vec{x} = (x_1, \ldots, x_n)$ suhteen on ykkösten osuus datajoukossa eli*

$$p^* \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Todistus. Bernoullijakauman tiheysfunktio voidaan kirjoittaa kompaktissa muodossa

$$f(x\,|\,p) \;=\; (1-p)^{1-x}p^x,$$

jonka avulla binaarimallin uskottavuusfunktio havaitun datajoukon $(x_1, \ldots, x_n)$ suhteen saadaan muotoon

$$L(p) \;=\; f(x_1 \,|\, p) \cdots f(x_n \,|\, p) \;=\; \prod_{i=1}^{n} (1-p)^{1-x_i} p^{x_i}.$$

Tätä vastaava logaritminen uskottavuusfunktio voidaan sieventää muotoon

$$\begin{aligned} \ell(p) \;&=\; \sum_{i=1}^{n} \left((1-x_i) \log(1-p) + x_i \log(p) \right) \\ &=\; n(1-m(\vec{x})) \log(1-p) + nm(\vec{x}) \log(p), \end{aligned}$$

missä $m(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i$ on ykkösten suhteellinen osuus havaitussa datajoukossa ja samalla kyseisen datajoukon keskiarvo. Derivoimalla

$$\ell'(p) \;=\; n(1-m(\vec{x}))(1-p)^{-1}(-1) + nm(\vec{x})p^{-1}$$

ja ratkaisemalla yhtälö $\ell'(p) = 0$ havaitaan että derivaatan nollakohta on $p = m(\vec{x})$. Derivoimalla toisen kerran voidaan tarkistaa, että $\ell''(p) \leq 0$. Näin ollen $p^* = m(\vec{x})$ on suurimman uskottavuuden estimaatti. □

Esimerkki 7.5 (Vialliset komponentit)**.** Tuotantolinjalla valmistetaan komponentteja menetelmällä, jonka seurauksena yksittäinen komponentti on viallinen todennäköisyydellä p, muista riippumattomasti. Kun tarkastettiin 200 komponentin erä, havaittiin 22 viallista. Määritä suurimman uskottavuuden estimaatti tuntemattoman parametrin p arvolle.

Koska komponentit ovat viallisia toisistaan riippumattomasti, vastaa datalähde binaarimallia, jossa 0=“ehjä” ja 1=“viallinen”. Lauseen 7.4 mukaan parametrin $p = f(1)$ suurimman uskottavuuden estimaatti on ykkösten osuus havaitussa datajoukossa eli $p^* = \frac{22}{200} = 11\%$. Suurimman uskottavuuden estimaatti viallisten komponenttien osuudelle *koko tuotannossa* on siis sama kuin viallisten komponenttien osuus tarkastetussa erässä. ■

Esimerkki 7.6 (Mielipidekysely)**.** Erään valtion äänioikeutetuista valittiin satunnaisotannalla $n = 2000$ henkilöä ja heiltä kysyttiin, aikovatko äänestää nykyistä presidenttiä seuraavissa presidentinvaaleissa (0=“ei”, 1=“kyllä”). Vastanneista 774 vastasi kyllä. Estimoi kannatusosuus p koko populaatiossa soveltamalla binaarimallin suurimman uskottavuuden menetelmää.

Mikäli 2000 henkilön satunnaisotanta tehdään ilman palautusta N äänioikeutetun populaatiosta, jossa nykyisen presidentin kannatusosuus on p (tuntematon), on todennäköisyys havaita 774 “kyllä”-ääntä

$$L(p) \;=\; \frac{\binom{Np}{774}\binom{N-Np}{2000-774}}{\binom{N}{2000}}.$$

Ylläolevaa ylläolevaa uskottavuusfunktiota on mahdotonta maksimoida p:n suhteen tuntematta N:n arvoa. Koska tässä tilanteessa populaation koko N on kuitenkin paljon suurempi kuin satunnaisotoksen koko 2000, noudattaa mielipidemittausta vastaava datalähde likimain binaarimallia parametrina p. Tällöin

lauseen 7.4 mukaan saadaan parametrin p suurimman uskottavuuden estimaatiksi

$$p^* \;=\; \frac{774}{2000} \;=\; 38.7\%.$$

Suurimman uskottavuuden estimaatti kannatusosuudelle *kaikkien äänioikeutettujen populaatiossa* on siis sama kuin kannatusosuus mielipidemittauksessa. ■

7.4 Normaalimallin estimointi

Normaalijakauman tiheysfunktio

$$f_{\mu,\sigma}(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

on parametreja μ ja σ vaille tunnettu. Kun parametreja on kaksi, voidaan parametrit koodata vektoriksi $\theta = (\theta_1, \theta_2)$, jolloin uskottavuusfunktiosta tulee kahden muuttujan funktio $L(\theta) = L(\theta_1, \theta_2)$.

Lause 7.7. *Normaalijakauman parametrien μ ja σ suurimman uskottavuuden estimaatit datajoukolle $(x_1, \ldots, x_n)$ ovat datajoukon keskiarvo ja datajoukon keskihajonta*

$$\mu^* \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textit{ja} \qquad \sigma^* \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^*)^2\right)^{1/2}. \tag{7.4}$$

Todistus. Havaittua datajoukkoa $(x_1, \ldots, x_n)$ vastaava uskottavuusfunktio on kahden muuttujan funktio

$$L(\mu, \sigma) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \;=\; (2\pi)^{-n/2}\sigma^{-n} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}.$$

Ottamalla ylläolevan yhtälön molemmilta puolilta logaritmit saadaan logaritmiseksi uskottavuusfunktioksi

$$\ell(\mu, \sigma) \;=\; -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}.$$

Logaritmisen uskottavuusfunktion derivaatat parametrien μ ja σ suhteen ovat

$$\begin{aligned}
\frac{d}{d\mu}\ell(\mu, \sigma) \;&=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \\
\frac{d}{d\sigma}\ell(\mu, \sigma) \;&=\; -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2.
\end{aligned}$$

Asettamalla ylläolevat derivaatat nolliksi ja ratkaisemalla näistä yhtälöistä parametrit μ ja σ, saadaan derivaattojen nollakohdiksi yhtälön (7.4) mukaiset μ^* ja σ^*. Lukupari (μ^*, σ^*) on ainoa parametrikombinaatio, jolle uskottavuusfunktion molemmat derivaatat ovat nollia. Toisen kertaluvun derivaattoja tarkastelemalla voidaan varmistaa, että $L(\mu^*, \sigma^*)$ on uskottavuusfunktion globaali maksimi. □

7.5 Kaksiulotteisen lineaarisen mallin estimointi

Lineaarinen regressio on yleinen tilastollinen lähestymistapa, jossa moniulotteisen datajoukon tietyn muuttujan käyttäytymistä pyritään ennustamaan tai selittämään parametrisena funktiona muista datajoukon muuttujista. Keskeinen perustapaus on kaksiulotteinen datajoukko $(x_1, y_1), \ldots, (x_n, y_n)$, jonka y-muuttujan arvoja x-muuttujan funktiona on tarkoitus ennustaa suoran $y = \alpha x + \beta$ avulla. Yksi tapa mitata suoran sovituksen hyvyyttä on *keskineliövirhe* (engl. mean squared error)

$$\text{MSE} \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha x_i - \beta)^2. \tag{7.5}$$

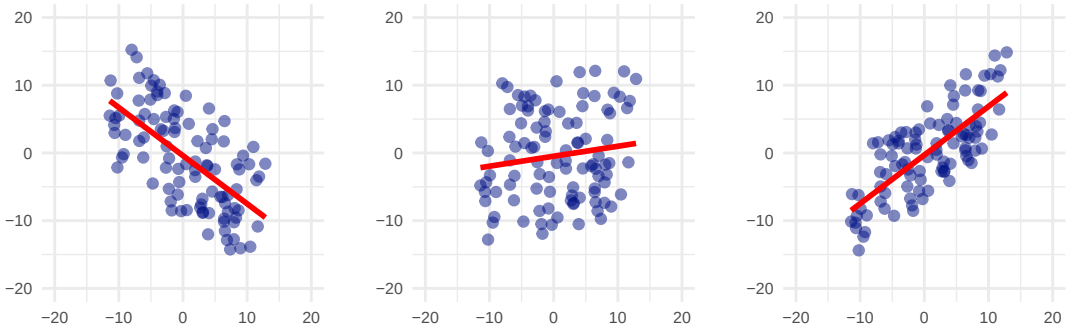
Keskineliövirhe voidaan myös kirjoittaa muodossa $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, jossa $\hat{y}_i = \alpha x_i + \beta$ on suoran avulla laskettu y-muuttujan ennuste pisteessä x_i. Tällöin paras sovitus saadaan valitsemalla suora, jonka kulmakerroin α ja vakiotermi β ovat sellaiset, että ylläoleva keskineliövirhe on pienin mahdollinen. Tätä sovitustapaa kutsutaan *pienimmän neliösumman* menetelmäksi ja määritettyä suoraa *regressiosuoraksi*.

Lause 7.8. *Keskineliövirheen näkökulmasta paras suora saadaan valitsemalla suoran kulmakerroin α^* ja vakiotermi β^* kaavoilla*[1]

$$\alpha^* \;=\; \frac{\mathrm{sd}(\vec{y})}{\mathrm{sd}(\vec{x})}\,\mathrm{cor}(\vec{x},\vec{y}), \qquad \beta^* = m(\vec{y}) - \alpha^* m(\vec{x}),$$

missä $m(\vec{x}), m(\vec{y})$ ja $\mathrm{sd}(\vec{x}), \mathrm{sd}(\vec{y})$ ovat datajoukon x- ja y-muuttujien keskiarvot ja keskihajonnat, ja $\mathrm{cor}(\vec{x},\vec{y})$ niiden välinen korrelaatio.

Alla on esitetty kolme kaksiulotteista sadan alkion datajoukkoa sekä niihin sovitetut regressiosuorat. Varoitus: Pienimmän neliösumman menetelmä sovittaa suoran sellaisiinkin kaksiulotteisiin datajoukkoihin, joissa minkäänlaista lineaarista riippuvuutta ei ole havaittavissa.



[1] Kulmakertoimen α^* kaavassa ei ole väliä, käyttääkö keskihajontaa vai otoskeskihajontaa, sillä kaavan (6.3) muuntokertoimet kumoavat toisensa osamäärässä $\frac{\mathrm{sd}(\vec{y})}{\mathrm{sd}(\vec{x})} = \frac{\mathrm{sd}_s(\vec{y})}{\mathrm{sd}_s(\vec{x})}$.

Lauseen 7.8 todistus. Keskineliövirhettä on kätevä analysoida soveltamalla kaksiulotteisen datajoukon empiirisen jakauman todennäköisyystulkintaa. Jos (X, Y) on satunnainen lukupari, joka on arvottu tasaisen satunnaisesti havaitusta datajoukosta, niin kaavan (6.15) mukaan

$$\text{MSE} \;=\; \mathbb{E}(Y - \alpha X - \beta)^2.$$

Avaamalla neliölauseke muotoon

$$(Y - \alpha X - \beta)^2 \;=\; Y^2 + \alpha^2 X^2 + \beta^2 - 2\alpha XY - 2\beta Y + 2\alpha\beta X$$

ja käyttämällä odotusarvon lineaarisuutta nähdään, että

$$\text{MSE} \;=\; \mathbb{E}Y^2 + \alpha^2 \mathbb{E}(X^2) + \beta^2 - 2\alpha\mathbb{E}(XY) - 2\beta\mathbb{E}(Y) + 2\alpha\beta\mathbb{E}(X).$$

Ylläolevan lausekkeen derivaatta parametrin β suhteen on

$$\frac{d}{d\beta}\,\text{MSE} \;=\; 2\beta - 2\mathbb{E}(Y) + 2\alpha\mathbb{E}(X).$$

Ratkaisemalla $\frac{d}{d\beta}\,\text{MSE} = 0$ saadaan

$$\beta \;=\; \mathbb{E}(Y) - \alpha\mathbb{E}(X).$$

Tätä yhtälöä soveltamalla saadaan keskineliövirheen derivaataksi parametrin α suhteen

$$\begin{aligned}
\frac{d}{d\alpha}\,\text{MSE} \;&=\; 2\alpha\mathbb{E}(X^2) - 2\mathbb{E}(XY) + 2\beta\mathbb{E}(X) \\
&=\; 2\alpha\mathbb{E}(X^2) - 2\mathbb{E}(XY) + 2(\mathbb{E}(Y) - \alpha\mathbb{E}(X))\mathbb{E}(X) \\
&=\; 2\alpha\,\text{Var}(X) - 2\,\text{Cov}(X, Y).
\end{aligned}$$

Ratkaisemalla $\frac{d}{d\alpha}\,\text{MSE} = 0$ saadaan tästä korrelaation määritelmän mukaan

$$\alpha \;=\; \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \;=\; \frac{\text{Cor}(X, Y)\,\text{SD}(X)\,\text{SD}(Y)}{\text{SD}(X)^2} \;=\; \frac{\text{SD}(Y)}{\text{SD}(X)}\,\text{Cor}(X, Y).$$

Väite seuraa tästä, sillä lauseen 6.3 mukaan $\mathbb{E}(X) = m(\vec{x})$, $\mathbb{E}(Y) = m(\vec{y})$, $\text{SD}(X) = \text{sd}(\vec{x})$ ja $\text{SD}(Y) = \text{sd}(\vec{y})$ sekä lauseen 6.5 mukaan $\text{Cor}(X, Y) = \text{cor}(\vec{x}, \vec{y})$. □

Yllä esitetty lineaarisen mallin sovitusmenetelmä voidaan tulkita myös suurimman uskottavuuden estimaattorina stokastiselle mallille, jossa datalähde tuottaa satunnaisluvut $(Y_1, \ldots, Y_n)$ muotoa

$$Y_i \;=\; \alpha x_i + \beta + \sigma Z_i, \tag{7.6}$$

missä toisistaan riippumattomat satunnaismuuttujat $Z_1, \ldots, Z_n$ noudattavat normitettua normaalijakaumaa ja luvut $x_1, \ldots, x_n$ sekä parametri $\sigma > 0$ ovat

tunnettuja. Tämä on kahden muuttujan *lineaarinen normaalimalli*. Näillä oletuksilla satunnaismuuttujien $Y_1, \ldots, Y_n$ yhteisjakauman tiheysfunktio on

$$f(y_1, \ldots, y_n \,|\, x_1, \ldots, x_n, \alpha, \beta, \sigma) \;=\; \prod_{i=1}^{n} (2\pi\sigma)^{-1/2} e^{-\frac{(y_i - \alpha x_i - \beta)^2}{2\sigma^2}}.$$

Kun lukuarvot $x_1, \ldots, x_n$ ja parametri σ sekä havainnot $y_1, \ldots, y_n$ oletetaan tunnetuiksi, voidaan ylläoleva tiheysfunktio tulkita tuntemattomien parametrien α ja β uskottavuusfunktiona $L(\alpha, \beta)$, jonka logaritmi $\ell(\alpha, \beta) = L(\alpha, \beta)$ voidaan kirjoittaa muodossa

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

Ylläoleva lauseke voidaan esittää keskineliövirheen (7.5) avulla myös muodossa

$$\ell(\alpha, \beta) \;=\; -\frac{n}{2} \log(2\pi\sigma) - \frac{n}{2\sigma^2} \,\mathrm{MSE},$$

josta nähdään että uskottavuusfunktio maksimoituu täsmälleen silloin, kun keskineliövirhe minimoituu. Näin ollen lineaarisen normaalimallin (7.6) parametrien α ja β suurimman uskottavuuden estimaattorit ovat samat kuin lauseessa 7.8.

7.6 Estimaattoreiden ominaisuuksia

Tarkastellaan datalähdettä, jonka tuottamat satunnaismuuttujat $X_1, X_2, \ldots$ noudattavat jakaumaa $f(x \,|\, \theta)$, missä parametri θ on tuntematon. Mallin parametrin θ:

- *estimaatti* on havaitun datajoukon $\vec{x} = (x_1, \ldots, x_n)$ pohjalta laskettu arvaus $\hat{\theta} = g(\vec{x})$ parametrin θ arvoksi,
- *estimaattori* on funktio $(x_1, \ldots, x_n) \mapsto g(x_1, \ldots, x_n)$, joka kuvaa datajoukon estimaatiksi[2].

Tuntemattoman parametrin estimaattoriksi voidaan periaatteessa valita mikä tahansa funktio $g(\vec{x})$. Intuitiivisesti on kuitenkin selvää, että jotkut estimaattorit ovat parempia kuin toiset.

Stokastisen mallin estimaattorin hyvyyttä voidaan luonnehtia analysoimalla, miten se käyttäytyisi saadessaan syötteekseen riippumattomia satunnaislukuja $X_1, X_2, \ldots$ mallin mukaisesta jakaumasta $f(x \,|\, \theta)$. Estimaattori $g(x_1, \ldots, x_n)$ on *tarkentuva* (engl. consistent), jos tapahtuman

$$g(X_1, \ldots, X_n) \;=\; \theta \pm \epsilon$$

[2]Estimaattoriksi kutsutaan usein myös satunnaismuuttujaa $g(X) = g(X_1, \ldots, X_n)$, joka on laskettu jakaumasta $f(x \,|\, \theta)$ generoitujen satunnaislukujen $(X_1, \ldots, X_n)$ muunnoksena.

todennäköisyys lähestyy ykköstä suurilla n:n arvoilla, oli $\epsilon > 0$ miten pieni hyvänsä. Tarkentuvuus siis tarkoittaa, että estimaattori tuottaa suurella todennäköisyydellä lähellä todellista parametria olevia arvoja, silloin kun käytössä on paljon dataa. Estimaattori $g(x_1, \ldots, x_n)$ on *harhaton* (engl. unbiased), jos

$$\mathbb{E}\Big(g(X_1, \ldots, X_n)\Big) \;=\; \theta.$$

Harhattomuus tarkoittaa, että jos samasta datalähteestä laskettaisiin suuri määrä estimaatteja $g(X_1, \ldots, X_n)$, niin estimaattien keskiarvo olisi lähellä oikeaa parametria.

Esimerkki 7.9 (Binaarimalli)**.** Binaarimallin satunnaismuuttujat noudattavat Bernoulli-jakaumaa parametrina $p = f(1)$ ja suurimman uskottavuuden estimaattori on (lause 7.4) havaitun datajoukon keskiarvo

$$m(x_1, \ldots, x_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Jos $X_1, X_2, \ldots$ ovat binaarimallin mukaisia toisistaan riippumattomia satunnaislukuja, niin $\mathbb{E}(X_i) = p$ ja odotusarvon lineaarisuuden perusteella

$$\mathbb{E}m(X_1, \ldots, X_n) \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \;=\; p.$$

Lisäksi suurten lukujen lain perusteella (lause 3.3) tapahtuman

$$\frac{1}{n}\sum_{i=1}^{n} X_i \;=\; p \pm \epsilon$$

todennäköisyys on lähellä ykköstä suurilla n:n arvoilla. Näin ollen $m(x_1, \ldots, x_n)$ on tarkentuva ja harhaton estimaattori binaarimallin parametrille p. ■

Esimerkki 7.10 (Normaalimallin odotusarvo)**.** Normaalimallin odotusarvoparametrin μ suurimman uskottavuuden estimaattori (lause 7.7) on havaitun datajoukon keskiarvo

$$m(x_1, \ldots, x_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Samoin perustein kuin esimerkissä 7.9 nähdään, että $m(x_1, \ldots, x_n)$ on tarkentuva ja harhaton estimaattori normaalimallin odotusarvoparametrille μ. ■

Esimerkki 7.11 (Normaalimallin keskihajonta ja varianssi)**.** Normaalimallin keskihajontaparametrin σ suurimman uskottavuuden estimaattori (lause 7.7) on datajoukon keskihajonta

$$\mathrm{sd}(\vec{x}) \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\vec{x}))^2\right)^{1/2}.$$

Suurten lukujen lain (lause 3.3) avulla voidaan perustella, että keskihajonta $\mathrm{sd}(\vec{x})$ on tarkentuva estimaattori normaalimallin keskihajontaparametrille σ. Myös otoskeskihajonta $\mathrm{sd}_s(\vec{x})$ on tarkentuva, sillä keskihajonnan ja otoskeskihajonnan välinen muuntokerroin (6.3) lähestyy ykköstä suurilla n:n arvoilla.

Keskihajontojen laskukaavoissa esiintyvän epälineaarisen neliöjuurioperaation johdosta $\mathrm{sd}(\vec{x})$ ja $\mathrm{sd}_s(\vec{x})$ eivät kuitenkaan ole parametrin σ estimaattoreina harhattomia. Tästä syystä varsinkin klassisessa frekventistisessä tilastotieteessä on ollut tapana etsiä harhatonta estimaattoria varianssiparametrille σ^2. Varianssiparametrin suurimman uskottavuuden estimaattori on havaitun datajoukon varianssi $\mathrm{var}(\vec{x}) = \mathrm{sd}(\vec{x})^2$. Tämäkin estimaattori on lievästi harhainen, sillä on mahdollista todistaa (lause 6.8), että

$$\mathbb{E}\Big(\,\mathrm{sd}(\vec{X})^2\Big) \;=\; \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m(\vec{X}))^2\right) \;=\; \cdots \;=\; \frac{n-1}{n}\sigma^2.$$

Datajoukon otosvarianssi $\mathrm{var}_s(\vec{x}) = \mathrm{sd}_s(\vec{x})^2$ on sen sijaan harhaton, sillä muuntokaavan (6.3) mukaan

$$\mathbb{E}\Big(\,\mathrm{sd}_s(\vec{X})^2\Big) \;=\; \frac{n}{n-1}\mathbb{E}\Big(\,\mathrm{sd}(\vec{X})^2\Big) \;=\; \sigma^2.$$

■

Luku 8

Tilastolliset luottamusvälit

8.1 Luottamusvälin käsite

Edellisessä luvussa lasketut parametriestimaatit olivat yksittäisiä reaalilukuja eli *piste-estimaatteja*. Koska käytännössä piste-estimaatti lähes aina poikkeaa jonkin verran parametrin todellisesta arvosta, tarvitaan tapa kuvailla piste-estimaatin tarkkuutta. Parametrin θ *väliestimaatti* on havaitusta datajoukosta $\vec{x} = (x_1, \ldots, x_n)$ laskettu lukuväli $[a(\vec{x}), b(\vec{x})]$, johon parametrin θ toivotaan sisältyvän. Väliestimaatin laskentamenetelmä eli *väliestimaattori* puolestaan on funktio, joka kuvaa datajoukon $\vec{x}$ väliestimaatiksi $[a(\vec{x}), b(\vec{x})]$.

Luonnollinen tapa analysoida väliestimaattorin hyvyyttä olisi laskea tapahtuman $\theta \in [a(\vec{x}), b(\vec{x})]$ todennäköisyys. Tämä on kuitenkin mahdotonta, sillä kyseiseen tapahtumaan ei liity ainuttakaan satunnaismuuttujaa[1]. Sen sijaan voidaan miettiä mitä tapahtuisi, jos väliestimaatti laskettaisiin havaitun datajoukon sijasta käyttäen mallin mukaisesta datalähteestä tuotettuja satunnaislukuja $\vec{X} = (X_1, \ldots, X_n)$. Näin muodostetun lukuvälin $[a(\vec{X}), b(\vec{X})]$ päätepisteet ovat satunnaismuuttujia, joille voidaan laskea todennäköisyys

$$\mathbb{P}\Big(\theta \in [a(\vec{X}), b(\vec{X})]\Big) \;=\; \mathbb{P}\Big(a(\vec{X}) \leq \theta \text{ ja } b(\vec{X}) \geq \theta\Big).$$

Mitä suurempi ylläoleva todennäköisyys on, sitä paremmat takeet väliestimaattorin hyvyydelle saadaan. Stokastisen mallin parametrin θ *luottamusväli* luottamustasolla α on lukuväli $[a(\vec{x}), b(\vec{x})]$, jonka päätepisteet on laskettu havaitusta datajoukosta $\vec{x} = (x_1, \ldots, x_n)$ käyttämällä väliestimaattoria, jolle

$$\mathbb{P}\Big(\theta \in [a(\vec{X}), b(\vec{X})]\Big) \;\geq\; \alpha$$

kaikilla parametrin θ arvoilla.

Esimerkki 8.1 (Jatkuva tasajakauma)**.** Datalähteen satunnaisluvut noudattavat jatkuvan välin $[0, \theta]$ tasajakaumaa, missä θ on tuntematon. Datalähteestä on havaittu arvot $x_1 = 8.1, x_2 = 2.6$ ja $x_3 = 8.8$. Määritä parametrille θ luottamusväli luottamustasolla 95%.

[1] Tämä on mahdollista silloin, kun θ tulkitaan satunnaismuuttujaksi, ks. luku 10.

Esimerkissä 7.2 johdettiin parametrin θ suurimman uskottavuuden estimaattoriksi $\hat{\theta}(\vec{x}) = \max\{x_1, \ldots, x_n\}$. Luottamusvälin määrittämiseksi voidaan tarkastella, millaisia arvoja estimaattori tuottaisi uusintamittauksesta saataville satunnaismuuttujille $\vec{X} = (X_1, \ldots, X_n)$. Yksittäisen satunnaismuuttujan X_i kertymäfunktio pisteessä $t \in [0, \theta]$ on

$$\mathbb{P}(X_i \leq t) \;=\; \int_0^t f(x \,|\, \theta)\, dx \;=\; \int_0^t \theta^{-1}\, dx \;=\; \frac{t}{\theta},$$

joten satunnaismuuttujan $\hat{\theta}(\vec{X}) = \max\{X_1, \ldots, X_n\}$ kertymäfunktio on

$$\mathbb{P}(\hat{\theta}(\vec{X}) \leq t) \;=\; \mathbb{P}(X_1 \leq t, \ldots, X_n \leq t) \;=\; \mathbb{P}(X_1 \leq t) \cdots \mathbb{P}(X_n \leq t) \;=\; \left(\frac{t}{\theta}\right)^n.$$

Näin ollen luvuille $0 \leq s \leq t \leq 1$ pätee

$$\mathbb{P}(s\theta \leq \hat{\theta}(\vec{X}) \leq t\theta) \;=\; \mathbb{P}(\hat{\theta}(\vec{X}) \leq t\theta) - \mathbb{P}(\hat{\theta}(\vec{X}) \leq s\theta) \;=\; t^n - s^n,$$

joka voidaan myös kirjoittaa muodossa

$$\mathbb{P}\left(\frac{\hat{\theta}(\vec{X})}{t} \leq \theta \leq \frac{\hat{\theta}(\vec{X})}{s}\right) \;=\; t^n - s^n.$$

Luottamustason α väliestimaattori saadaan siis kaavasta

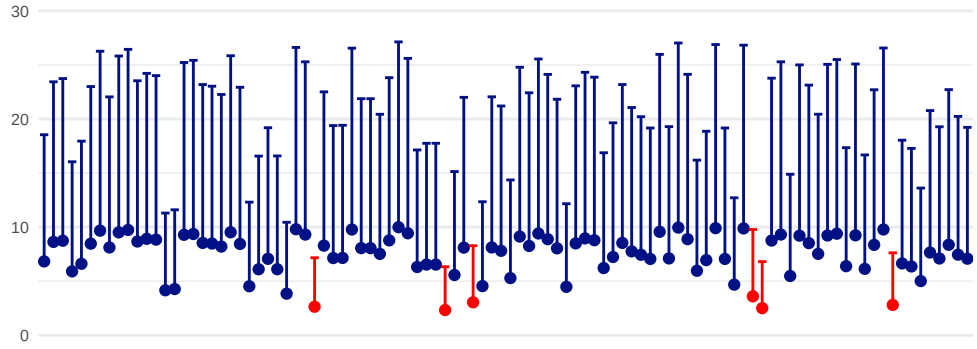
$$\left[\frac{\hat{\theta}(\vec{x})}{t}, \frac{\hat{\theta}(\vec{x})}{s}\right],$$

kun s ja t on valittu niin, että $t^n - s^n = \alpha$. Tästä nähdään, että valitun luottamustason luottamusväli voidaan määrittää monella eri tapaa. Eräs luonteva tapa on valita $s = (1-\alpha)^{1/n}$ ja $t = 1$. Tällöin datajoukolle $(8.1, 2.6, 8.8)$ saadaan luottamustason $\alpha = 0.95$ luottamusväliksi

$$\left[\hat{\theta}(\vec{x}),\, \frac{\hat{\theta}(\vec{x})}{(1-\alpha)^{1/n}}\right] \;=\; \left[8.8,\, \frac{8.8}{(1-0.95)^{1/3}}\right] \;=\; [8.8,\, 23.9]\,.$$

■

Kuvassa 8.1 on esitetty sata ylläolevalla väliestimaattorilla laskettua luottamusväliä. Jokainen luottamusväli on laskettu datajoukosta, jonka kolme alkiota on generoitu R:n pseudosatunnaislukugeneraattorilla välin $[0, 10]$ jatkuvasta tasajakaumasta. Luottamusväleistä osuus $\frac{94}{100}$ peittää estimoitavan parametrin todellisen arvon $\theta = 10$. Edellämainittu osuus lähestyisi arvoa 0.95, mikäli luottamusvälien estimoimista jatkettaisiin rajattoman monta kertaa.



Kuva 8.1: Simuloimalla tuotetuista $n = 3$ alkion datajoukoista väliestimaattorilla $\left[\hat{\theta}(\vec{x}),\ \frac{\hat{\theta}(\vec{x})}{(1-0.95)^{1/3}}\right]$ laskettuja luottamusvälejä tasajakauman parametrille θ.

8.2 Odotusarvoparametrin luottamusväli

Tarkastellaan yleistä stokastista datalähdettä, joka tuottaa riippumattomia ja samoin jakautuneita satunnaislukuja, joiden jakaumasta ei tiedetä mitään muuta kuin että jakaumalla on olemassa äärellinen odotusarvo ja keskihajonta. Havaitun datajoukon $\vec{x} = (x_1, \ldots, x_n)$ keskiarvo

$$m(\vec{x}) \ = \ \frac{1}{n}\sum_{i=1}^{n} x_i$$

on tarkentuva ja harhaton piste-estimaatti jakauman odotusarvoparametrille μ. Sitä vastaavan luottamusvälin määrittäminen vaikuttaa mahdottomalta tehtävältä ilman tarkempia tietoja datalähteen jakaumasta. Silloin kun havaittu datajoukko on suuri, luottamusväli kuitenkin voidaan likiarvoisesti määrittää keskeisen raja-arvolauseen avulla. Allaolevan symmetrisen luottamusvälin sädettä $z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$ kutsutaan luottamusvälin *virhemarginaaliksi*.

Lause 8.2. *Suurelle datajoukolle likiarvoinen luottamustason α luottamusväli odotusarvoparametrille μ saadaan kaavasta*

$$m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}, \tag{8.1}$$

missä $m(\vec{x})$ ja $\mathrm{sd}(\vec{x})$ ovat havaitun datajoukon keskiarvo ja keskihajonta[2]*, ja z on luku, jolle normitettua normaalijakaumaa noudattava satunnaismuuttuja Z toteuttaa $\mathbb{P}(-z \leq Z \leq z) = \alpha$.*

Todistus. Todetaan ensiksi, että $\mu = m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$ pätee täsmälleen silloin, kun

$$-z \ \leq \ \frac{m(\vec{x}) - \mu}{\mathrm{sd}(\vec{x})/\sqrt{n}} \ \leq \ z.$$

[2]Kaavassa (8.1) voi käyttää myös otoskeskihajontaa, sillä $\mathrm{sd}_s(\vec{x}) \approx \mathrm{sd}(\vec{x})$ suurilla n.

Tarkastellaan seuraavaksi, millä todennäköisyydellä ylläoleva lauseke pitää paikkansa, kun havaittu datajoukko $\vec{x} = (x_1, \ldots, x_n)$ korvataan mallin mukaisesta datalähteestä generoidulla satunnaismuuttujien listalla $\vec{X} = (X_1, \ldots, X_n)$. Koska suurten lukujen lain (lause 3.3) perusteella $\mathrm{sd}(\vec{X})$ on lähellä mallin keskihajontaa σ, niin keskeisen raja-arvolauseen (lause 5.9) perusteella suurilla n pätee

$$\frac{m(\vec{X}) - \mu}{\mathrm{sd}(\vec{X})/\sqrt{n}} \;\approx\; \frac{m(\vec{X}) - \mu}{\sigma/\sqrt{n}} \;\approx\; Z,$$

missä satunnaismuuttuja Z noudattaa normitettua normaalijakaumaa. Näin ollen

$$\mathbb{P}\left(\mu = m(\vec{X}) \pm z\frac{\mathrm{sd}(\vec{X})}{\sqrt{n}}\right) \;=\; \mathbb{P}\left(-z \le \frac{m(\vec{X}) - \mu}{\mathrm{sd}(\vec{x})/\sqrt{n}} \le z\right) \;\approx\; \mathbb{P}(-z \le Z \le z) \;=\; \alpha.$$

□

Likiarvoisen luottamusvälin lausekkeessa (8.1) tarvitaan luku z, jolle normitettua normaalijakaumaa noudattava Z toteuttaa $\mathbb{P}(-z \le Z \le z) = \alpha$. Todennäköisyyden peruslaskusääntöjen ja normaalijakauman symmetrian perusteella

$$\begin{aligned}
\mathbb{P}(-z \le Z \le z) &= \mathbb{P}(Z \le z) - \mathbb{P}(Z < -z) \\
&= \mathbb{P}(Z \le z) - \mathbb{P}(Z > z) \\
&= \mathbb{P}(Z \le z) - (1 - \mathbb{P}(Z \le z)) \\
&= 2F_Z(z) - 1,
\end{aligned}$$

missä $F_Z(z) = \mathbb{P}(Z \le z)$ on normitetun normaalijakauman kertymäfunktio. Tarvittu luku z saadaan siis yhtälön $2F_Z(z) - 1 = \alpha$ ratkaisuna muodossa

$$z \;=\; F_Z^{-1}\left(\frac{1+\alpha}{2}\right). \tag{8.2}$$

Kertymäfunktion käänteisfunktiolle $F_Z^{-1}(z)$ ei tunneta siistiä matemaattista lauseketta, mutta sen arvot voidaan katsoa taulukoista tai laskea numeerisilla ohjelmistoilla (R: `qnorm`, Excel: `NORM.S.INV`).

Esimerkki 8.3 (Kahviautomaatti)**.** Kahviautomaatin toimintaa testattiin valuttamalla automaatista 30 kupillista ja mittaamalla kahvin määrät kupeissa (yksikkönä cl). Mittauksesta kerättiin datajoukko

$\vec{x} =$ (10.24, 9.94, 10.55, 9.28, 10.57, 9.77, 9.50, 10.03, 10.51, 10.29, 10.92, 10.06, 9.63, 10.83, 10.36, 9.17, 10.29, 10.24, 9.73, 10.56, 9.18, 9.84, 9.91, 10.74, 9.57, 10.76, 10.53, 10.52, 10.42, 10.11),

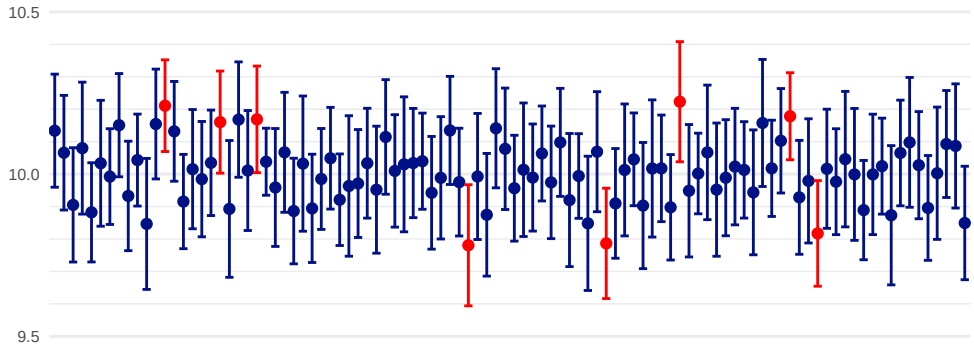
jonka keskiarvo on $m(\vec{x}) = 10.14$ ja keskihajonta $\mathrm{sd}(\vec{x}) = 0.49$. Määritä likiarvoinen luottamustason 95% luottamusväli kahviautomaatin valuttamien kahvimäärien pitkän aikavälin keskiarvolle μ.

Likiarvoiseksi luottamusväliksi saadaan kaavan (8.1) avulla

$$m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}} \;=\; 10.14 \pm 1.96 \times \frac{0.49}{\sqrt{30}} \;=\; 10.14 \pm 0.09,$$

missä kaavan (8.2) mukaan $z \;= F_Z^{-1}(0.975) = 1.96$. ■

Kuvassa 8.2 on esitetty sata ylläolevalla väliestimaattorilla laskettua luottamusväliä. Jokainen luottamusväli on laskettu datajoukosta, jonka 30 alkiota on generoitu R:n pseudosatunnaislukugeneraattorilla normaalijakaumasta odotusarvona $\mu = 10$ ja keskihajontana $\sigma = 0.5$. Luottamusväleistä osuus $\frac{92}{100}$ peittää estimoitavan parametrin todellisen arvon $\mu = 10$. Edellämainittu osuus lähestyisi arvoa 0.95, mikäli luottamusvälien estimoimista toistettaisiin rajattoman monta kertaa.



Kuva 8.2: Simuloimalla tuotetuista $n = 30$ alkion datajoukoista väliestimaattorilla $m(\vec{x}) \pm 1.96\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$ laskettuja luottamusvälejä yleisen mallin odotusarvoparametrille μ.

8.3 Binaarimallin parametrin luottamusväli

Esimerkissä 7.1 havaittiin, että binaarisen datalähteen tuottamat satunnaisluvut noudattavat Bernoulli-jakaumaa parametrina $p = f(1)$. Binaarimallin parametrin estimointi on hieman helpompaa kuin yleisen mallin, sillä $\{0, 1\}$-arvoisen satunnaismuuttujan jakauma ja kaikki siihen liittyvät tunnusluvut (odotusarvo, keskihajonta, jne.) määräytyvät täysin parametrista p.

Lause 8.4. *Suurelle datajoukolle likiarvoinen luottamustason α luottamusväli binaarimallin parametrille $p = f(1)$ saadaan kaavasta*

$$\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \tag{8.3}$$

missä $\hat{p} = \hat{p}(\vec{x})$ on ykkösten suhteellinen osuus havaitussa datajoukossa $\vec{x}$ ja z on luku, jolle normitettua normaalijakaumaa noudattava satunnaismuuttuja Z toteuttaa $\mathbb{P}(-z \leq Z \leq z) = \alpha$.

Todistus. Esimerkin 7.1 mukaan binaarimallin tuottamat $\{0,1\}$-arvoiset satunnaismuuttujat X_i noudattavat Bernoulli-jakaumaa parametrina $p = \mathbb{P}(X_i = 1)$, ja parametri p on myös jakauman odotusarvo eli pätee $p = \mathbb{E}(X_i)$. Kyseessä oleva estimointitehtävä on siis erikoistapaus yleisestä odotusarvon estimointitehtävästä. Lauseen 8.2 mukaan likiarvoinen suuren datajoukon luottamusväli saadaan näin ollen kaavasta

$$m(\vec{x}) \pm z\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}, \tag{8.4}$$

jossa $m(\vec{x})$ ja $\mathrm{sd}(\vec{x})$ ovat havaitun datajoukon keskiarvo ja keskihajonta, ja z on luku, jolle normitettua normaalijakaumaa noudattava satunnaismuuttuja Z toteuttaa $\mathbb{P}(-z \leq Z \leq z) = \alpha$.

Todetaan seuraavaksi, että havaitun datajoukon odotusarvolle pätee

$$m(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;=\; \hat{p},$$

missä $\hat{p} = \hat{p}(\vec{x})$ on ykkösten suhteellinen osuus havaitussa datajoukossa $\vec{x} = (x_1, \ldots, x_n)$. Lisäksi, koska $\{0,1\}$-arvoisille muuttujille pätee $x_i^2 = x_i$, havaitun datajoukon varianssi voidaan sieventää muotoon

$$\begin{aligned}
\mathrm{var}(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{p})^2 \;&=\; \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\hat{p} + \hat{p}^2) \\
&=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\hat{p}\frac{1}{n}\sum_{i=1}^{n} x_i + \hat{p}^2 \\
&=\; \frac{1}{n}\sum_{i=1}^{n} x_i - 2\hat{p}^2 + \hat{p}^2 \\
&=\; \hat{p} - \hat{p}^2.
\end{aligned}$$

Ottamalla neliöjuuret, havaitaan tästä että $\mathrm{sd}(\vec{x}) = \sqrt{\hat{p} - \hat{p}^2}$. Väite seuraa sijoittamalla $m(\vec{x}) = \hat{p}$ ja $\mathrm{sd}(\vec{x}) = \sqrt{\hat{p} - \hat{p}^2}$ kaavaan (8.4). □

Esimerkki 8.5 (Mielipidekysely)**.** Erään valtion äänioikeutetuista valittiin satunnaisotannalla $n = 2000$ henkilöä ja heiltä kysyttiin, aikovatko äänestää nykyistä presidenttiä seuraavissa presidentinvaaleissa (0=Ei, 1=Kyllä). Vastanneista 774 vastasi kyllä. Määritä piste-estimaatti ja luottamustason 95% luottamusväli presidentin kannatusosuudelle p kaikkien äänioikeutettujen keskuudessa.

Koska satunnaisotoksen koko on pieni suhteessa koko tutkittavaan populaatioon, ovat yksittäisten henkilöiden vastaukset toisistaan likimain riippumattomat. Tällöin mielipidekyselyn tulokset noudattavat likimain binaarimallia parametrina p. Kannatusosuuden suurimman uskottavuuden piste-estimaatti on (esimerkki 7.6) on

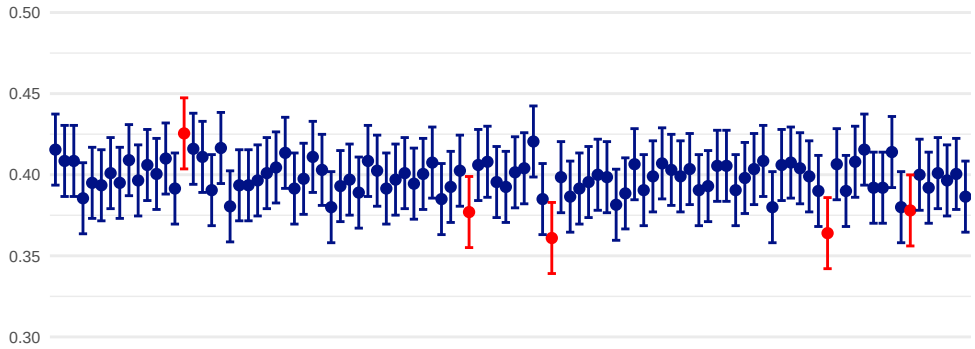
$$\hat{p} \;=\; \frac{774}{2000} \;=\; 0.387,$$

ja likiarvoinen luottamustason 95% luottamusväli saadaan kaavasta (8.3)

$$\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \;=\; 0.387 \pm 1.96 \times \frac{\sqrt{0.237}}{\sqrt{2000}} \;=\; 0.387 \pm 0.021,$$

jossa on käytetty kaavan (8.2) mukaista z-arvoa $F_Z^{-1}(0.975) = 1.96$. ■

Kuvassa 8.3 on esitetty sata binaarimallin väliestimaattorilla laskettua luottamusväliä. Jokainen luottamusväli on laskettu datajoukosta, jonka 2000 alkiota on generoitu R:n pseudosatunnaislukugeneraattorilla Bernoulli-jakaumasta parametrina $p = 0.4$. Luottamusväleistä täsmälleen osuus $\frac{95}{100}$ peittää estimoitavan parametrin todellisen arvon $p = 0.4$. Näin ei välttämättä käy aina, mutta edellä mainittu osuus lähestyisi arvoa 0.95, mikäli luottamusvälien estimoimista toistettaisiin rajattoman monta kertaa.



Kuva 8.3: Simuloimalla tuotetuista $n = 2000$ alkion datajoukoista väliestimaattorilla $\hat{p}(\vec{x}) \pm 1.96\frac{\sqrt{\hat{p}(\vec{x})(1-\hat{p}(\vec{x}))}}{\sqrt{n}}$ laskettuja luottamusvälejä binaarimallin parametrille p.

Hieman ongelmallinen puoli kaavassa (8.3) on se, että luottamusvälin leveys riippuu riippuu piste-estimaatin arvosta $\hat{p}(\vec{x})$, sillä käytännössä halutaan yleensä etukäteen määrittää luku n, joka riittää halutunlevyisen luottamusvälin aikaansaamiseen valitulla luottamustasolla. Mikäli halutaan luottamusväli, jonka leveys ei riipu havaitusta datasta, voidaan kaavan (8.3) sijaan käyttää konservatiivista väliestimaattoria

$$\hat{p} \pm z\frac{0.5}{\sqrt{n}}. \tag{8.5}$$

Ylläoleva väli on leveämpi kuin kaavan (8.3) tuottama väli, sillä derivoimalla funktiota $x \mapsto x - x^2$ nähdään, että $\sqrt{\hat{p}(1-\hat{p})} \leq 0.5$ kaikilla $\hat{p} \in [0, 1]$. Taulukossa 8.1 on muutamia esimerkkejä luottamusvälien leveyksistä. Melko vakiintunut tapa on käyttää 95% luottamustasoa, mutta tälle valinnalle ei ole sen syvällisempiä matemaattisia perusteita.

n	95%-luottamusväli	99%-luottamusväli
1000	$\hat{p} \pm 3.1\%$	$\hat{p} \pm 4.1\%$
2000	$\hat{p} \pm 2.2\%$	$\hat{p} \pm 2.9\%$
10000	$\hat{p} \pm 1.0\%$	$\hat{p} \pm 1.3\%$

Taulukko 8.1: Konservatiivisella väliestimaattorilla (8.5) laskettuja binaarimallin luottamusvälejä. Tarkemmalla kaavalla (8.3) laskettaessa saadaan kapeampia luottamusvälejä, joiden leveys vaihtelee.

8.4 Kommentteja

Esimerkissä johdettiin välin $[0, \theta]$ jatkuvan tasajakauman parametrille θ luottamustason $\alpha = 95\%$ luottamusväli

$$\left[\hat{\theta}(\vec{x}),\ \frac{\hat{\theta}(\vec{x})}{(1-\alpha)^{1/n}}\right] \ = \ [8.8,\ 23.9]$$

havaitusta datajoukosta $\vec{x} = (8.2, 2.6, 8.8)$. Hieman intuition vastaisesti on kuitenkin väärin sanoa, että näin saatu lukuväli [8.8, 23.9] peittää tuntemattoman parametriarvon θ vähintään 95% todennäköisyydellä. Luottamusvälien analysoimisessa nimittäin todennäköisyydet liittyvät tuleviin, vielä havaitsemattomiin väliestimaattorin arvoihin. Oikea tulkinnan mukaan suuresta määrästä ylläolevalla väliestimaattorilla laskettuja lukuvälejä vähintään 95% peittää tuntemattoman parametriarvon. Yksittäisestä luottamusvälistä ei luku $\alpha = 95\%$ kerro mitään. Tämä on syy, miksi lukua α kutsutaan *luottamustasoksi*, ei todennäköisyydeksi.

Kuvassa 8.2 esitettiin simuloimalla tuotetuista datajoukoista väliestimaattorilla $m(\vec{x}) \pm 1.96\frac{\mathrm{sd}(\vec{x})}{\sqrt{n}}$ laskettuja luottamusvälejä yleisen mallin odotusarvoparametrille μ. Luottamusväleistä osuus $\frac{92}{100}$ peitti estimoitavan parametrin todellisen arvon, mikä on aika paljon alle väitetyn luottamustason 95%. Tämä epätarkkuus johtuu siitä, että laskuissa käytetty $n = 30$ ei ole erityisen suuri luku suuren datajoukon likiarvoisen luottamusväliestimaattorin (8.1) näkökulmasta. Pienille datajoukoille on johdettu kehittyneempiä menetelmiä luottamusvälien estimoimiseen. Esimerkiksi normaalimallin odotusarvoparametrille voidaan johtaa pienillekin datajoukoille tarkka luottamusväliestimaattori kaavalla

$$m(\vec{x}) \pm z\frac{\mathrm{sd}_s(\vec{x})}{\sqrt{n}},$$

jossa z lasketaan normaalijakaumaa hieman paksuhäntäisemmästä *t-jakaumasta*. Kyseisen jakauman teki tunnetuksi vuonna 1908 englantilainen William S. Gosset (1876–1937), joka Guinnesin panimolla työskennellessään käytti julkaisuissaan salanimeä Student.

Luottamusvälejä voidaan odotusarvojen lisäksi laskea muillekin tunnusluvuille, esimerkiksi keskihajonnalle ja korrelaatiolle. Näiden matemaattinen ana-

lyysi on melko monimutkaista, ja normaalijakauman lisäksi tarvitaan normaalijakauman muunnosten jakaumia, esim. t-jakauma, χ^2-jakauma[3] ja F-jakauma. Luottamusvälien yleisen teorian perustukset luonnosteli puolalaislähtöinen Jerzy Neyman (1894–1981) vuonna 1937 artikkelissa *Outline of a theory of statistical estimation based on the classical theory of probability*. Monimutkaisten mallien likiarvoisia luottamusvälejä voidaan laskea tietokonesimulointiin perustuvalla *bootstrap-menetelmällä*, jonka esitteli amerikkalainen Bradley Efron (s. 1938) vuonna 1979 artikkelissa *Bootstrap methods: Another look at the jackknife*.

[3]lausutaan "khii toiseen"

Luku 9

Bayesläiset tilastolliset mallit

9.1 Priorijakauma ja posteriorijakauma

Bayesläisen tilastollisen päättelyn lähtökohtana on päivittää satunnaisilmiöön liittyvien tapahtumien todennäköisyyksiä sitä mukaa kuin ilmiöstä saadaan uutta dataa. Tämä edellyttää todennäköisyyden käsitteen subjektiivista tulkintaa. Bayesläisessä ajattelussa datalähteen käyttäytymistä kuvaava tuntematon parametri mielletään satunnaismuuttujaksi, jonka jakauma kuvaa havainnoijan uskomusta parametrin arvosta. Uskomusta päivitetään, kun havaitaan uutta dataa.

Havainnoijan uskomusta parametrin arvosta kuvaa *priorijakauma* $p(\theta)$, joka kertoo millä todennäköisyydellä havainnoija uskoo parametrin arvon olevan θ. Kun datalähteestä havaitaan uusi datapiste x, uskomus päivitetään uudeksi jakaumaksi. Päivitetty jakauma $p(\theta \mid x)$ on parametrin *posteriorijakauma*.[1] Uskomuksen päivittäminen perustuu *uskottavuusfunktioon*

$$f(x \mid \theta),$$

joka määrittää parametrin θ mukaan käyttäytyvän datalähteen tuottamien arvojen jakauman.[2] Diskreetti priorijakauma $p(\theta)$ päivitetään posteriorijakaumaksi soveltamalla *Bayesin päivityskaavaa*

$$p(\theta \mid x) \;=\; \frac{p(\theta) f(x \mid \theta)}{\sum_{\theta'} p(\theta') f(x \mid \theta')}. \tag{9.1}$$

Posteriorijakauma saadaan siis priorijakauman ja uskottavuusfunktion normitettuna tulona. Päivityskaava jatkuville priorijakaumille saadaan vaihtamalla summa integraaliksi ylläolevassa kaavassa. Päivityskaava voidaan johtaa luvun 1.7 tuloksista ja siinä esiintyvät funktiot voidaan tulkita satunnaismuuttujien yhteisjakauman reunajakaumina (luku 9.6).

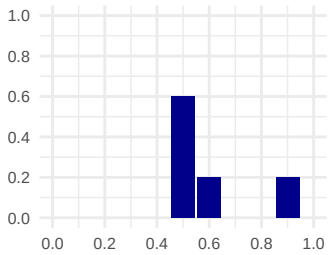
[1] lat. prior = edeltävä, posterior = seuraava

[2] Funktiota $f(x \mid \theta)$ kutsutaan tiheysfunktioksi silloin, kun se tulkitaan x:n funktiona, ja uskottavuusfunktioksi silloin, kun se tulkitaan θ:n funktiona.

Esimerkki 9.1 (Tuntematon kolikko)**.** Laatikossa tiedetään olevan kolme tasaista (kruunan tn $\theta = 0.5$), yksi lievästi vino ($\theta = 0.6$) ja yksi vahvasti vino ($\theta = 0.9$) kolikko. Satunnaisesti valittua kolikkoa heitettäessä havaitaan klaava. Millä todennäköisyydellä heitetty kolikko oli tasainen?

Ennen datan havaitsemista laatikosta satunnaisesti valittu kolikko on tasainen todennäköisyydellä $\frac{3}{5}$, lievästi vino todennäköisyydellä $\frac{1}{5}$ ja vahvasti vino todennäköisyydellä $\frac{1}{5}$. Kun tuntematonta parametria θ mallinnetaan satunnaismuuttujana Θ, on parametrin jakauma ennen datan havaitsemista ao. taulukon mukainen.

θ	0.5	0.6	0.9
$p(\theta)$	0.6	0.2	0.2



Koska θ edustaa kruunan todennäköisyyttä, on parametrin θ uskottavuusfunktio havainnon klaava suhteen

$$f(\text{klaava} \,|\, \theta) \;=\; 1 - \theta.$$

Näin ollen

$$\begin{aligned}\sum_{\theta} p(\theta) f(\text{klaava} \,|\, \theta) \;&=\; 0.6 \times (1 - 0.5) \;+\; 0.2 \times (1 - 0.6) \;+\; 0.2 \times (1 - 0.9) \\ &=\; 0.4,\end{aligned}$$

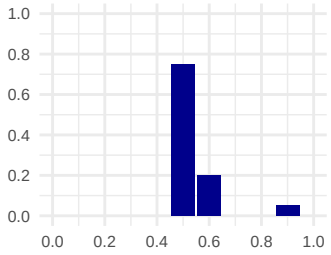
joten kaavan (9.1) mukaan parametriarvon 0.5 posterioritodennäköisyys on

$$p(0.5 \,|\, \text{klaava}) \;=\; \frac{p(0.5) f(\text{klaava} \,|\, 0.5)}{\sum_{\theta} p(\theta) f(\text{klaava} \,|\, \theta)} \;=\; \frac{0.6 \times (1 - 0.5)}{0.4} \;=\; 0.75.$$

Klaavan havaitseminen siis kasvatti kolikon tasaisuuden todennäköisyyttä arvosta 0.6 arvoon 0.75, mutta kolikolle itselleen ei heiton aikana tapahtunut mitään. Satunnaismuuttuja Θ ei siis kuvasta heitettyä kolikkoa, vaan kolikon heittäjän subjektiivista uskomusta heitetyn kolikon tyypistä.

Kaavan (9.1) avulla voidaan laskea posterioritodennäköisyydet myös parametriarvoille 0.6 ja 0.9. Tuloksena saadaan allaolevassa taulukossa esitetty posteriorijakauma.

θ	0.5	0.6	0.9
$p(\theta \,|\, \text{klaava})$	0.75	0.20	0.05



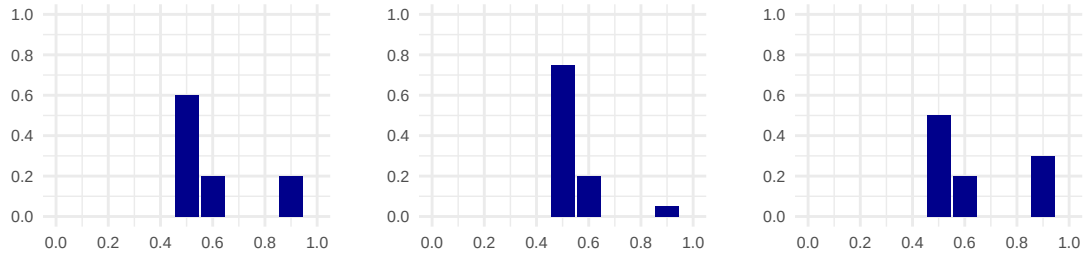
Käytännössä posteriorijakauma kannattaa laskea vaiheittain sarake sarakkeelta allaolevan taulukon avulla. Kolme ensimmäistä saraketta saadaan suoraan tehtävänannosta. Sarake 4 eli normittamaton posterioritiheys saadaan kertomalla pareittain sarakkeiden 2 ja 3 alkiot. Päivityskaavassa (9.1) esiintyvä normitusvakio saadaan summaamalla sarakkeen 4 alkiot, eli tässä tapauksessa 0.4. Sarake 5 saadaan jakamalla sarakkeen 4 alkiot normitusvakiolla 0.4.

Parametri	Prioritiheys	Uskottavuus	Normittamaton posterioritiheys	Posterioritiheys
θ	$p(\theta)$	$f(\text{klaava} \mid \theta)$	$p(\theta) f(\text{klaava} \mid \theta)$	$p(\theta \mid \text{klaava})$
0.5	0.6	0.5	0.30	0.75
0.6	0.2	0.4	0.08	0.20
0.9	0.2	0.1	0.02	0.05

Lasketaan samalla tapaa vielä posteriorijakauma havainnon $x = \text{kruuna}$ suhteen. Laskelman välivaiheet on esitetty allaolevassa taulukossa.

Parametri	Prioritiheys	Uskottavuus	Normittamaton posterioritiheys	Posterioritiheys
θ	$p(\theta)$	$f(\text{kruuna} \mid \theta)$	$p(\theta) f(\text{kruuna} \mid \theta)$	$p(\theta \mid \text{kruuna})$
0.5	0.6	0.5	0.30	0.50
0.6	0.2	0.6	0.12	0.20
0.9	0.2	0.9	0.18	0.30

Alkuperäinen priorijakauma ja tuloksena saadut kaksi posteriorijakaumaa on esitetty alla.



Priorijakauma $p(\theta)$ | Posteriorijakauma $p(\theta \mid \text{klaava})$ | Posteriorijakauma $p(\theta \mid \text{kruuna})$

9.2 Usean datapisteen posteriorijakauma

Bayesin päivityskaavaa (9.1) voidaan soveltaa myös silloin, kun havaitaan monta datapistettä. Tällöin kaavan muuttuja x tulkitaan datapisteiden listaksi $x =$

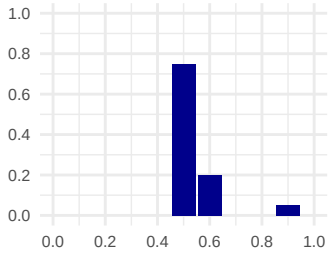
$(x_1, \ldots, x_n)$. Tarkastellaan esimerkiksi saman kolikon heittämistä monta kertaa peräkkäin. Kun havaitaan kaksi klaavaa, uskottavuusfunktio datajoukolle $(x_1, x_2) = (0, 0)$ on

$$f(0, 0 \,|\, \theta) \;=\; f(0 \,|\, \theta) f(0 \,|\, \theta) \;=\; (1 - \theta)^2.$$

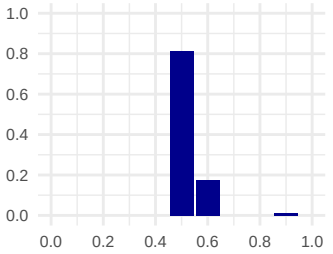
Parametrin θ posteriorijakauma kahden klaavan suhteen saadaan allaolevan taulukon mukaisilla laskuilla.

Parametri	Prioritiheys	Uskottavuus	Normittamaton posterioritiheys	Posterioritiheys
θ	$p(\theta)$	$f(0, 0 \,|\, \theta)$	$p(\theta) f(0, 0 \,|\, \theta)$	$p(\theta \,|\, 0, 0)$
0.5	0.6	0.25	0.150	0.815
0.6	0.2	0.16	0.032	0.174
0.9	0.2	0.01	0.002	0.001

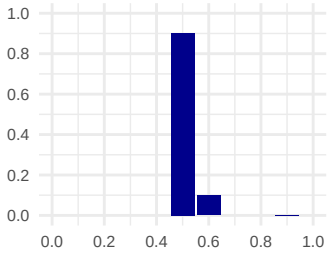
Samaan tapaan voidaan laskea posteriorijakauma useampienkin klaavojen sarjoille. Alla muutama posteriorijakauma.



Posteriorijakauma $p(\theta \,|\, 0)$



Posteriorijakauma $p(\theta \,|\, 0, 0)$



Posteriorijakauma $p(\theta \,|\, 0, 0, 0, 0, 0)$

Sadan klaavan havaitsemisen jälkeen posteriorijakauman massa keskittyy tähtitieteellisen pientä poikkeamaa vaille arvoon 0.5. Tämä tuntuu paradoksaaliselta, sillä tn saada 100 klaavaa peräkkäin tasaisella kolikolla on 2^{-100}. Paradoksi selittyy priorin valinnalla: Ylläoleva priorijakauma $p(\theta)$ kuvastaa vankkumatonta 100% ennakkovarmuutta siitä, että kolikko ei puolla klaavan suuntaan.

9.3 Uskomuksen vaiheittainen päivittäminen

Tarkastellaan datalähdettä, joka tuottaa tiheysfunktion $f(x \,|\, \theta)$ mukaan jakautuneita riippumattomia satunnaismuuttujia. Havainnoijan uskomusta tuntemattoman parametrin arvosta kuvastaa priorijakauma $p(\theta)$. Havaittuaan datapisteen x_1 hän päivittää uskomuksensa posteriorijakaumaksi $p(\theta \,|\, x_1)$. Miten uskomus tulee päivittää, kun sen jälkeen havaitaan vielä toinen uusi datapiste x_2?

Havaittuja datapisteitä x_1 ja x_2 vastaava posteriorijakauma voidaan laskea soveltamalla Bayesin päivityskaavaa (9.1) muodossa

$$p(\theta \,|\, x_1, x_2) \;=\; \frac{p(\theta) f(x_1, x_2 \,|\, \theta)}{\sum_{\theta'} p(\theta') f(x_1, x_2 \,|\, \theta')}, \tag{9.2}$$

missä $f(x_1, x_2 \mid \theta)$ on todennäköisyys havaita pistepari (x_1, x_2) parametrin θ mukaan käyttäytyvästä datalähteestä. Toinen tapa päivittää uskomusta on ensin laskea posteriorijakauma $\hat{p}(\theta) = p(\theta \mid x_1)$ datapisteen x_1 suhteen ja sen jälkeen laskea *uusi posteriorijakauma* priorijakaumasta $\hat{p}(\theta)$ datapisteen x_2 suhteen. Seuraava tärkeä tulos osoittaa, että vaiheittainen päivitys saman tuloksen kuin suorakin tapa. Tästä seuraa, että sarjamuotoista dataa havainnoivan henkilön tai tietokoneohjelman ei tarvitse pitää kirjaa menneistä datapisteitä: nykyinen uskomus jakaumaksi koodattuna sisältää ennustamisen kannalta kaiken oleellisen informaation menneestä.

Lause 9.2. *Kaavan* (9.2) *määrittämä posteriorijakauma voidaan laskea vaiheittain muodossa*

$$p(\theta \mid x_1, x_2) \;=\; \frac{\hat{p}(\theta) f(x_2 \mid \theta)}{\sum_{\theta'} \hat{p}(\theta') f(x_2 \mid \theta')},$$

missä $\hat{p}(\theta) = p(\theta \mid x_1)$.

Todistus. Bayesin päivityskaavan (9.1) mukaan

$$\hat{p}(\theta) \;=\; p(\theta \mid x_1) \;=\; c^{-1} p(\theta) f(x_1 \mid \theta),$$

missä normitusvakio $c = \sum_\theta p(\theta) f(x_1 \mid \theta)$. Koska datalähde tuottaa riippumattomia satunnaismuuttujia, pätee lisäksi

$$f(x_1, x_2 \mid \theta) \;=\; f(x_1 \mid \theta) f(x_2 \mid \theta).$$

Näin ollen

$$\begin{aligned}
\frac{\hat{p}(\theta) f(x_2 \mid \theta)}{\sum_{\theta'} \hat{p}(\theta') f(x_2 \mid \theta')} &= \frac{c^{-1} p(\theta) f(x_1 \mid \theta) f(x_2 \mid \theta)}{\sum_{\theta'} c^{-1} p(\theta') f(x_1 \mid \theta') f(x_2 \mid \theta')} \\
&= \frac{p(\theta) f(x_1, x_2 \mid \theta)}{\sum_{\theta'} p(\theta') f(x_1, x_2 \mid \theta')} \\
&= p(\theta \mid x_1, x_2).
\end{aligned}$$

□

9.4 Bayesläinen binaarimalli

Binaarinen datalähde tuottaa riippumattomia $\{0, 1\}$-arvoisia satunnaislukuja siten, että arvon 1 todennäköisyys on θ. Lähtökohtaisesti parametrista θ ei tiedetä mitään, joten sen uskotaan noudattavan jatkuvan välin $[0, 1]$ tasajakaumaa tiheysfunktiona

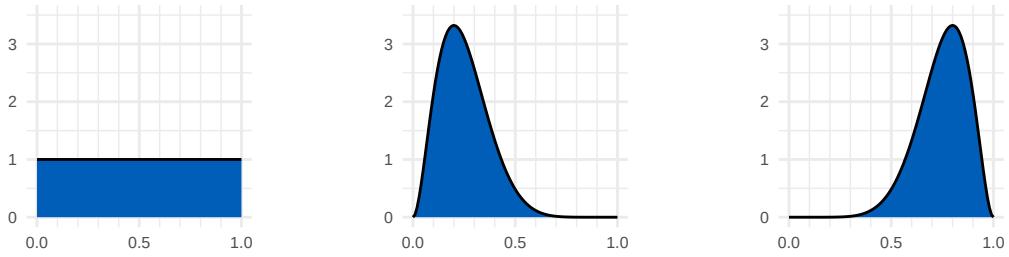
$$p(\theta) \;=\; \begin{cases} 1, & \theta \in [0, 1], \\ 0, & \text{muuten.} \end{cases}$$

Kun datalähteestä havaitaan arvot $\vec{x} = (x_1, \ldots, x_n)$, halutaan priorijakauma päivittää posteriorijakaumaksi. Bayesläisen binaarisen mallin kannalta keskeisiä jatkuvia jakaumia ovat betajakaumat.

Yleinen *betajakauma* parametreina $a > 0$ ja $b > 0$ on jatkuva jakauma, jonka tiheysfunktio on

$$f(\theta) \;=\; \begin{cases} c_{a,b}^{-1}\,\theta^{a-1}(1-\theta)^{b-1}, & \text{kun } \theta \in [0,1], \\ 0, & \text{muuten,} \end{cases}$$

missä normitusvakio[3] $c_{a,b} = \frac{(a-1)!(b-1)!}{(a+b-1)!}$. Betajakaumien tiheysfunktioita on piirretty kuvaan 9.1. Yksikkövälin $[0,1]$ tasajakauma on erikoistapaus betajakaumasta parametreina $a = 1$ ja $b = 1$.



$a = 1,\ b = 1$ $\qquad$ $a = 3,\ b = 9$ $\qquad$ $a = 9,\ b = 3$

Kuva 9.1: Betajakaumien tiheysfunktioita.

Bayesläisen binaarimallin erityispiirre on, että päivityskaavassa datajoukosta $\vec{x}$ riittää tietää ykkösten lukumäärä $||\vec{x}||$ ja nollien lukumäärä $n - ||\vec{x}||$. Betajakauma parametreina a ja b päivittyy betajakaumaksi parametreina $a + ||\vec{x}||$ ja $b + n - ||\vec{x}||$. Erityistapauksessa $a = 1$ ja $b = 1$ allaolevasta tuloksesta saadaan päivityskaava tasaiselle priorijakaumalle.

Lause 9.3. *Jos binaarisen datalähteen parametrin priorijakauma on betajakauma parametreina a ja b, niin posteriorijakauma havainnon $\vec{x} = (x_1, \ldots, x_n)$ suhteen on betajakauma parametreina $a + ||\vec{x}||$ ja $b + n - ||\vec{x}||$, missä $||\vec{x}|| = \sum_{i=1}^{n} x_i$ on ykkösten lukumäärä datajoukossa $\vec{x}$.*

Todistus. Yksittäisen datapisteen x_i uskottavuusfunktio on

$$f(x_i \,|\, \theta) \;=\; \begin{cases} 1-\theta, & x_i = 0, \\ \theta, & x_i = 1, \end{cases}$$

joten riippumattomuuden perusteella koko datajoukon uskottavuusfunktioksi saadaan

$$f(\vec{x} \,|\, \theta) \;=\; \prod_{i=1}^{n} f(x_i \,|\, \theta) \;=\; \theta^{||\vec{x}||}(1-\theta)^{n-||\vec{x}||}.$$

[3]Betajakauma voidaan määritellä myös silloin, kun a ja b eivät ole kokonaislukuja. Tällöin normitusvakiossa esiintyvät kertomat korvataan gammafunktiolla.

Normittamaton posteriorijakauma saadaan priorijakauman ja uskottavuusfunktion tulona muotoon

$$\begin{aligned} p(\theta)f(x \mid \theta) &= \left(c_{a,b}^{-1}\,\theta^{a-1}(1-\theta)^{b-1}\right)\theta^{||\vec{x}||}(1-\theta)^{n-||\vec{x}||} \\ &= c_{a,b}^{-1}\,\theta^{a+||\vec{x}||-1}(1-\theta)^{b+n-||\vec{x}||-1}. \end{aligned}$$

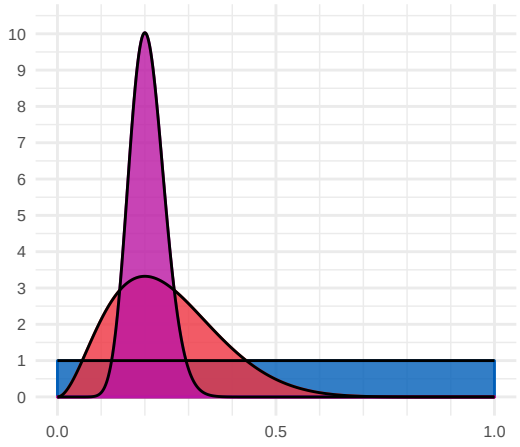
Bayesin päivityskaavan (9.1) mukaan posteriorijakauma on näin ollen

$$p(\theta \mid \vec{x}) = \frac{p(\theta)f(\vec{x} \mid \theta)}{\int p(\eta)f(\vec{x} \mid \eta)d\eta} = \frac{\theta^{a'-1}(1-\theta)^{b'-1}}{\int \eta^{a'-1}(1-\eta)^{b'-1}d\eta},$$

missä $a' = a + ||\vec{x}||$ ja $b' = b + n - ||\vec{x}||$. Ylläoleva lauseke on betajakauman tiheysfunktio parametreina a' ja b'. □

Esimerkki 9.4 (Tuntematon kolikko)**.** Tuntematonta kolikkoa heitettäessä havaittiin 2 kruunaa ja 8 klaavaa. Kolikosta ei ole ennalta mitään taustatietoja. Määritä kruunan todennäköisyyttä kuvaavan parametrin posteriorijakauma. Kuvaile posteriorijakauma myös tapauksessa, kun havaitaan 20 kruunaa ja 80 klaavaa.

Koska kolikosta ei ennalta tiedetä mitään, valitaan kruunan todennäköisyyttä kuvaavan parametrin θ priorijakaumaksi yksikkövälin $[0, 1]$ tasajakauma, joka on erityistapaus betajakaumasta parametreina $a = 1$ ja $b = 1$. Posteriorijakauma on lauseen 9.3 mukaan betajakauma parametreina $a + 2 = 3$ ja $b + 8 = 9$. Jos havaittaisiinkin 20 kruunaa ja 80 klaavaa, saataisiin posteriorijakaumaksi betajakauma parametreina $a + 20 = 21$ ja $b + 80 = 81$. Molempien tapauksien posteriorijakaumat on esitetty alla. ■



Kuva 9.2: Binaarimallin posteriorijakauma havainnon 2 kruunaa ja 8 klaavaa suhteen (punainen) sekä havainnon 20 kruunaa ja 80 klaavaa suhteen (pinkki). Priorijakauma on esitetty sinisellä.

9.5 Bayesläinen normaalimalli

Tarkastellaan datalähdettä, joka tuottaa riippumattomia normaalijakautuneita satunnaislukuja $X_1, X_2, \ldots$, joiden odotusarvo θ on tuntematon ja keskihajonta σ tunnettu. Mallin tuntematon odotusarvoparametri tulkitaan satunnaismuuttujaksi, joka noudattaa normaalijakaumaa odotusarvona μ_0 ja keskihajontana σ_0. Priorijakauman parametreja μ_0 ja σ_0 kutsutaan *hyperparametreiksi* erotuksena datalähteen käyttäytymistä suoraan kuvaaville parametreille. Datalähteen varsinaiset parametrit (θ, σ) kuvaavat siis datalähteen toimintaa, ja hyperparametrit (μ_0, σ_0) havainnoijan uskomusta mallin tuntemattomasta parametrista θ.

Normaalimallissa datalähteen odotusarvoparametrin priorijakauma on normaalijakauma tiheysfunktiona

$$p(\theta) \;=\; (2\pi\sigma_0^2)^{-1/2} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}} \tag{9.3}$$

ja uskottavuusfunktio havaitun datajoukon $\vec{x} = (x_1, \ldots, x_n)$ suhteen on

$$f(\vec{x}\,|\,\theta) \;=\; \prod_{i=1}^{n} f(x_i\,|\,\theta) \;=\; (2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}}. \tag{9.4}$$

Normaalimallin tekee erityisen käyttökelpoiseksi se, että normaalijakautunutta prioria vastaa normaalijakautunut posteriori. Lisäksi posteriorijakauman odotusarvon ja keskihajonnan laskemiseksi riittää tietää havaitun datajoukon keskiarvo $m(\vec{x})$ ja koko n. Allaolevassa normaalimallin päivityskaavassa posteriorijakauman odotusarvo on painotettu keskiarvo priorijakauman odotusarvosta ja havaitun datajoukon keskiarvosta.

Lause 9.5. *Bayesläisen normaalimallin posteriorijakauma on normaalijakauma, jonka odotusarvo ja keskihajonta saadaan kaavoista*

$$\mu_1 \;=\; \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2} m(\vec{x})}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \quad \textit{ja} \quad \sigma_1 \;=\; \frac{1}{\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}}, \tag{9.5}$$

missä $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$ *on havaitun datajoukon keskiarvo.*

Todistus. Posterijakauman normittamaton tiheysfunktio saadaan lausekkeiden (9.3) ja (9.4) tulona kirjoitettua muotoon $p(\theta) f(\vec{x}\,|\,\theta) = c_0 e^{-h(\theta)}$, missä

$$h(\theta) \;=\; \frac{(\theta-\mu_0)^2}{2\sigma_0^2} + \frac{\sum_i (x_i-\theta)^2}{2\sigma^2}.$$

ja $c_0 = (2\pi\sigma_0^2)^{-1/2}(2\pi\sigma^2)^{-n/2}$. Ylläoleva toisen asteen polynomi voidaan sieventää muotoon

$$h(\theta) \;=\; a\theta^2 + b\theta + c,$$

missä $a = \frac{1}{2}(\sigma_0^{-2}+n\sigma^{-2})$, $b = -\left(\sigma_0^{-2}\mu_0 + \sigma^{-2}\sum_i x_i\right)$ ja $c = \frac{1}{2}(\sigma_0^{-2}\mu_0^2+\sigma^{-2}\sum_i x_i^2)$. Yleistä neliöksi täydentämisen kaavaa

$$a\theta^2 + b\theta + c \;=\; a\left(\theta + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$$

soveltamalla saadaan funktiolle $h(\theta)$ esitys

$$h(\theta) \;=\; \frac{(\theta-\mu_1)^2}{2\sigma_1^2} + c - \frac{b^2}{4a},$$

missä

$$\mu_1 \;=\; -\frac{b}{2a} \;=\; \frac{\sigma_0^{-2}\mu_0 + n\sigma^{-2}m(\vec{x})}{\sigma_0^{-2}+n\sigma^{-2}}.$$

ja

$$\sigma_1 \;=\; (2a)^{-1/2} \;=\; (\sigma_0^{-2}+n\sigma^{-2})^{-1/2}.$$

Posterijakauman normittamaton tiheysfunktio saadaan siis muotoon

$$p(\theta)f(\vec{x}\,|\,\theta) \;=\; c_0 e^{-h(\theta)} \;=\; c_0 e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}+c-\frac{b^2}{4a}}.$$

Posterijakauman tiheysfunktio on siis

$$p(\theta\,|\,\vec{x}) \;=\; \frac{p(\theta)f(\vec{x}\,|\,\theta)}{\int p(\theta')f(\vec{x}\,|\,\theta')d\theta'} \;=\; \frac{e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}}}{\int e^{-\frac{(\theta'-\mu_1)^2}{2\sigma_1^2}}d\theta'} \;=\; \frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}},$$

missä viimeinen yhtälö seuraa siitä, että normaalijakauman tiheysfunktio integroituu ykköseksi. □

Esimerkki 9.6 (Kohinainen kanava)**.** Lukuarvoisia signaaleja lähetetään kohinaisen tiedonsiirtokanavan välityksellä. Kohinan seurauksena pisteestä A lähetetyn signaalin arvo θ vastaanotetaan pisteessä B satunnaismuuttujana, joka noudattaa normaalijakaumaa odotusarvona θ ja keskihajontana $\sigma = 2$. Tiedonsiirtovirheiden kompensoimiseksi sama signaali lähetetään kolme kertaa peräkkäin. Vastaanottaja arvelee ennalta, että lähetetyn signaalin arvo on peräisin normaalijakaumasta, jonka odotusarvo on $\mu_0 = 5$ ja keskihajonta $\sigma_0 = 1$. Mikä on lähetetyn signaalin posteriorijakauma vastaanotettujen signaaliarvojen $\vec{x} = (3.1, 7.9, 7.0)$ suhteen?

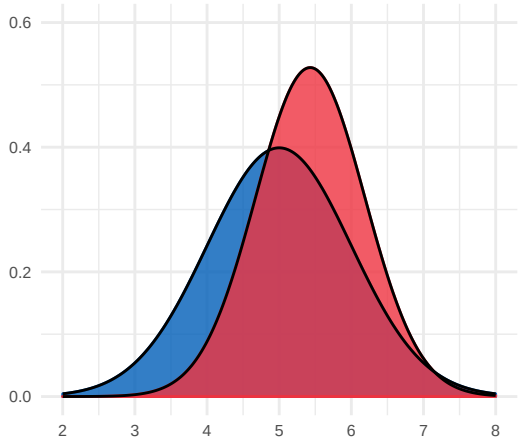
Lähetetyn signaalin posteriorijakauma on normaalijakauma, jonka parametrit saadaan sijoittamalla normaalimallin päivityskaavoihin (9.5) havaitun datajoukon keskiarvo $m(\vec{x}) = 6.0$ ja koko $n = 3$. Posteriorijakauman odotusarvo on

$$\mu_1 \;=\; \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}m(\vec{x})}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}} \;=\; \frac{\frac{1}{1^2}\times 5 + \frac{3}{2^2}\times 6}{\frac{1}{1^2}+\frac{3}{2^2}} \;\approx\; 5.43$$

ja keskihajonta

$$\sigma_1 \;=\; \frac{1}{\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}} \;=\; \frac{1}{\sqrt{\frac{1}{1^2}+\frac{3}{2^2}}} \;\approx\; 0.76.$$

Vastaanotettu datajoukko $\vec{x} = (3.1, 7.9, 7.0)$ päivittää vastaanottajan uskomusta lähetetyn signaalin arvosta niin, että odotusarvo siirtyy tasolta 5 tasolle 5.43 ja keskihajonta pienenee tasolta 1 tasolle 0.76. Priori- ja posterijakauma on esitetty kuvassa 9.3



Kuva 9.3: Lähetetyn signaalin priorijakauma (sininen) ja posteriorijakauma (punainen).

9.6 Kommentteja

Diskreetti bayesläinen malli voidaan tulkita satunnaismuuttujien parina (Θ, X), jonka yhteisjakaumalla on tiheysfunktio

$$f_{\Theta,X}(\theta, x) \;=\; p(\theta) f(x \,|\, \theta).$$

Satunnaismuuttujien Θ ja X jakaumien tiheysfunktiot voidaan kirjoittaa yhteisjakauman reunajakaumina (luku 2.4) muodossa

$$f_\Theta(\theta) \;=\; \sum_x p(\theta) f(x \,|\, \theta) \;=\; p(\theta)$$

ja

$$f_X(x) \;=\; \sum_\theta p(\theta) f(x \,|\, \theta) \;=\; f(x).$$

Lisäksi uskottavuusfunktio $f(x \mid \theta) = f_{X \mid \Theta}(x \mid \theta)$ voidaan mieltää satunnaismuuttujan X ehdolliseksi tiheysfunktioksi satunnaismuuttujan Θ suhteen (luku 2.5). Vastaavat kaavat ovat voimassa jatkuville jakaumille, kun summat vaihdetaan integraaleiksi.

Jos havaintoa x vastaava Bayesin päivityskaava (9.1) tulkitaan todennäköisyysjakaumien avaruudessa operoivana kuvauksena

$$U_x : \text{priorijakauma} \mapsto \text{posteriorijakauma},$$

voidaan lause 9.2 ilmaista ytimekkäästi muodossa $U_{x_2}(U_{x_1}(p)) = U_{(x_1,x_2)}(p)$. Tai vielä ytimekkäämmin muodossa $U_{x_2} \circ U_{x_1} = U_{x_1,x_2}$. Tämä kaava yleistyy induktiolla muotoon $U_{x_n} \circ \cdots \circ U_{x_1} = U_{x_1,x_2,\cdots,x_n}$, josta saadaan rekursiokaavat

$$\begin{aligned} U_{x_1,x_2,\cdots,x_n} &= U_{x_n} \circ U_{x_1,x_2,\cdots,x_{n-1}}, \\ U_{x_1,x_2,\cdots,x_n} &= U_{x_2,x_3,\ldots,x_n} \circ U_{x_1}. \end{aligned}$$

Bayesläinen binaarimalli oli Thomas Bayesin 1763 artikkelin keskeinen tutkimuskohde.

Luku 10

Bayes-estimaattorit

10.1 Bayesläiset piste-estimaatit

Tarkastellaan datalähdettä, joka tuottaa tiheysfunktion $f(x \mid \theta)$ mukaan jakautuneita riippumattomia satunnaismuuttujia. Havainnoijan uskomusta tuntemattoman parametrin arvosta kuvastaa priorijakauma $p(\theta)$. Havaittuaan datajoukon $\vec{x}$ hän päivittää uskomuksensa posteriorijakaumaksi $p(\theta \mid \vec{x})$. Mikä on havainnoijan paras estimaatti tuntemattoman parametrin arvolle?

Koska posteriorijakauma $p(\theta \mid \vec{x})$ sisältää kaiken oleellisen informaation havainnoijan uskomuksesta ja havaitusta datajoukosta $\vec{x}$, voidaan sitä pitää estimointitehtävän täydellisenä ratkaisuna. Käytännön tilanteissa on kuitenkin usein tarpeen raportoida yksi luku, joka edustaa jossain mielessä havainnoijan parasta arvausta tuntemattoman parametrin θ arvosta. Parametrin piste-estimaatiksi voidaan valita jokin posteriorijakauman tunnusluku. Yksi vaihtoehto on posteriorijakauman odotusarvo[1]

$$\theta_{\mathrm{E}}(\vec{x}) \;=\; \int_{-\infty}^{\infty} \theta \, p(\theta \mid \vec{x}) \, d\theta,$$

joka kuvaa millaisia arvoja posteriorijakaumasta keskimäärin saadaan simuloimalla suuri määrä otoksia. Toinen luonteva vaihtoehto on posteriorijakauman moodi eli sellainen parametrin arvo

$$\theta_{\mathrm{M}}(\vec{x}),$$

jossa posteriorijakauman tiheysfunktio $p(\theta \mid \vec{x})$ saavuttaa suurimman arvonsa.

Esimerkki 10.1 (Tuntematon kolikko)**.** Kolikkoa, josta ei ennalta tiedetä mitään, heittämällä havaitaan 4 kruunaa ja 1 klaava. Määritä kruunan todennäköisyyttä mallintavan parametrin θ piste-estimaatti käyttäen (a) posteriorijakauman odotusarvoa (b) posteriorijakauman moodia.

Koska kolikon luonteesta ei ennalta tiedetä mitään, valitaan parametrin priorijakaumaksi yksikkövälin tasajakauma. Parametrin posterijakauma on tällöin

[1] diskreetin jakauman tapauksessa integraali korvataan summalla

lauseen 9.3 mukaan betajakauma parametreina $a = 5$ ja $b = 2$. Posteriorijakauman tiheysfunktio voidaan kirjoittaa muodossa

$$p(\theta \,|\, \vec{x}) \;=\; \begin{cases} 30\,\theta^4(1-\theta), & \theta \in (0,1), \\ 0, & \text{muuten.} \end{cases}$$

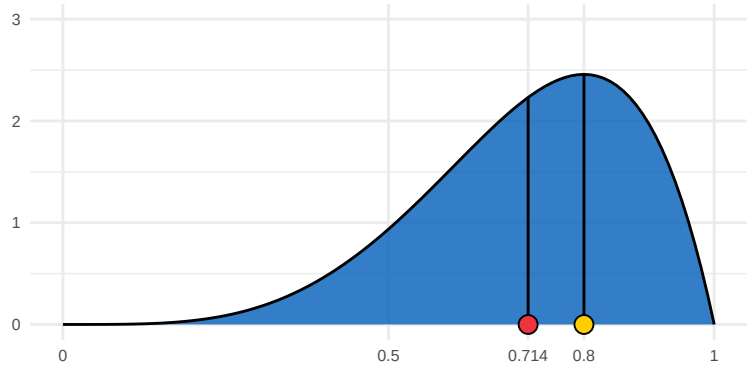
(a) Posteriorijakauman odotusarvo saadaan integraalina

$$\begin{aligned} \theta_{\mathrm{E}}(\vec{x}) \;&=\; \int_{-\infty}^{\infty} \theta\, p(\theta \,|\, \vec{x})\, d\theta \;=\; \int_0^1 \theta \times 30\,\theta^4(1-\theta)\, d\theta \\ &=\; 30\left(\int_0^1 \theta^5 d\theta - \int_0^1 \theta^6 d\theta\right) \\ &=\; 30 \times \left(\frac{1}{6} - \frac{1}{7}\right) \;=\; \frac{5}{7}. \end{aligned}$$

(b) Posteriorijakauman moodin määrittämiseksi etsitään funktion $p(\theta \,|\, \vec{x})$ maksimikohta. Posteriotiheyden derivaatta pisteessä $\theta \in (0,1)$ on

$$p'(\theta \,|\, \vec{x}) \;=\; 30\frac{d}{d\theta}(\theta^4 - \theta^5) \;=\; 30(4\theta^3 - 5\theta^4) \;=\; 30\,\theta^3(4 - 5\theta),$$

joten piste $\theta = \frac{4}{5}$ on derivaatan ainoa nollakohta välillä $(0,1)$. Derivoimalla toisen kerran voidaan tarkistaa, että $p''(\frac{4}{5}) < 0$, joten kyseinen piste on funktion $p(\theta \,|\, x)$ globaali maksimikohta. Näin ollen posterijakauman moodi on $\theta_{\mathrm{M}}(\vec{x}) = \frac{4}{5}$. Lasketut piste-estimaatit on esitetty kuvassa 10.1. ■

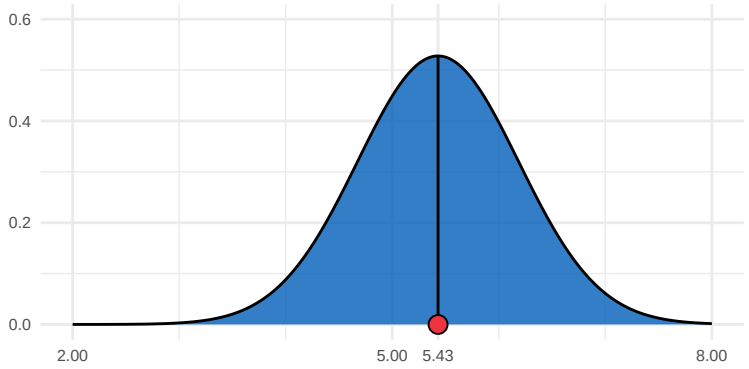


Kuva 10.1: Havaintoa {4 kruunaa ja 1 klaava} vastaavasta posteriorijakaumasta lasketut piste-estimaatit: odotusarvo $\theta_{\mathrm{E}}(\vec{x}) \approx 0.714$ ja moodi $\theta_{\mathrm{M}}(\vec{x}) = 0.8$.

Esimerkki 10.2 (Kohinainen kanava)**.** Esimerkissä 9.6 lähetetyn signaalin priorijakauma oli normaalijakauma odotusarvona $\mu_0 = 5$ ja keskihajontana $\sigma_0 = 1$. Posterijakaumaksi vastaanotettujen signaaliarvojen $\vec{x} = (3.1, 7.9, 7.0)$ suhteen saatiin normaalijakauma odotusarvona $\mu_1 = 5.43$ ja keskihajontana $\sigma_1 \approx 0.76$. Määritä lähetetyn signaalin piste-estimaatti käyttäen (a) posteriorijakauman odotusarvoa (b) posteriorijakauman moodia.

Normaalijakauman tiheysfunktio on symmetrinen odotusarvonsa suhteen ja saavuttaa maksiminsa odotusarvon kohdalla (kuva 10.2). Näin ollen molempien piste-estimaattien arvoiksi saadaan

$$\theta_{\mathrm{E}}(\vec{x}) \;=\; \theta_{\mathrm{M}}(\vec{x}) \;=\; 5.43.$$



Kuva 10.2: Lähetetyn signaalin posteriorijakaumasta lasketut piste-estimaatit: odotusarvo = moodi = 5.43.

■

Esimerkki 10.3 (Tasajakauman ylärajan estimointi)**.** Tuntemattoman välin $\{1, 2, \ldots, \theta\}$ tasajakaumaa noudattavasta datalähteestä on havaittu $x_1 = 21$, $x_2 = 7$ ja $x_3 = 22$. Ennakkoon on aihetta uskoa, että tuntemattoman parametrin arvo on todennäköisesti lähellä arvoa 30 ja uskomukseen liittyvä epävarmuus noudattaa Poisson-jakaumaa. Määritä posterijakauman moodia vastaava piste-estimaatti tuntemattoman parametrin arvolle.

Poisson-jakauman parametrina $\lambda = 30$ tiheysfunktio on

$$p(\theta) \;=\; e^{-\lambda}\frac{\lambda^{\theta}}{\theta!}, \quad \theta = 0, 1, 2, \ldots$$

Yksittäisen datapisteen x_i uskottavuusfunktio on

$$f(x_i \,|\, \theta) \;=\; \begin{cases} \frac{1}{\theta}, & 1 \le x_i \le \theta, \\ 0, & \text{muuten}, \end{cases}$$

joten datajoukon $\vec{x} = (x_1, x_2, x_3)$ uskottavuusfunktio on

$$f(\vec{x} \,|\, \theta) \;=\; \prod_{i=1}^{3} f(x_i \,|\, \theta) \;=\; \begin{cases} \frac{1}{\theta^3}, & 1 \le x_1, x_2, x_3 \le \theta, \\ 0, & \text{muuten}. \end{cases}$$

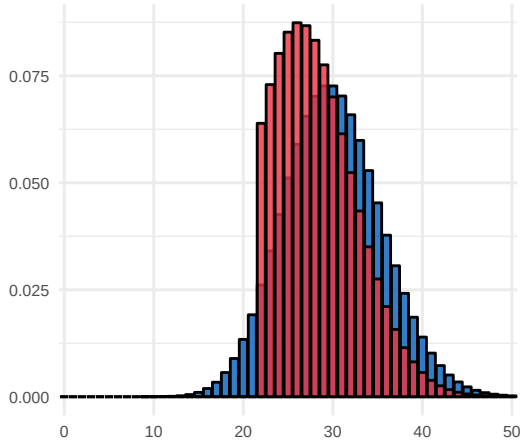
Posteriorijakauma lasketaan päivityskaavasta

$$p(\theta \,|\, \vec{x}) \;=\; \frac{p(\theta) f(\vec{x} \,|\, \theta)}{\sum_{\theta'} p(\theta') f(\vec{x} \,|\, \theta')} \;=\; \begin{cases} c^{-1} e^{-30} \frac{30^{\theta}}{\theta!} \frac{1}{\theta^3}, & \theta \ge 22, \\ 0, & \text{muuten}, \end{cases}$$

missä normitusvakio on $c = \sum_{\theta'} p(\theta') f(\vec{x} \,|\, \theta')$. Normitusvakion tarkka laskeminen on hieman hankalaa, mutta posteriorijakauman moodin määrittämiseksi normitusvakion arvoa ei tarvitse tietää. Riittää etsiä normittamattoman posterioritiheysfunktion $\theta \mapsto \frac{30^\theta}{\theta!} \frac{1}{\theta^3}$ maksimi joukossa $\theta \geq 22$. Kokeilemalla eri lukuarvoja havaitaan, että maksimi saavutetaan pisteessä $\theta = 26$, joten posteriorijakauman moodi on

$$\theta_{\mathrm{M}}(\vec{x}) \;=\; 26.$$

■



Kuva 10.3: Lukumäärän priorijakauma (sininen) ja posteriorijakauma (punainen).

10.2 Bayesläiset väliestimaatit

Todennäköisyysvälit ovat käyttökelpoinen tapa kuvailla ja raportoida subjektiivisia uskomuksia, sillä priori- tai posteriorijakaumaa on yleensä melko vaikea kuvailla sanallisesti. Satunnaiselle vastaantulijalle on luultavasti ymmärrettävämpää kertoa, että

> "parametri sisältyy 50% todennäköisyydellä välille $[0.48, 0.66]$"

kuin

> "parametri noudattaa betajakaumaa parametreina 8 ja 6".

Jakauman *todennäköisyysväli* tasolla α on lukuväli $[a, b]$, jolle kyseistä jakaumaa noudattava satunnaismuuttuja X toteuttaa

$$\mathbb{P}(a \leq X \leq b) \;=\; \alpha.$$

Todennäköisyysväli on symmetrinen, mikäli lisäksi pätee

$$\mathbb{P}(X < a) \;=\; \mathbb{P}(X > b).$$

Jatkuvien jakaumien todennäköisyysvälejä on käytännössä helpointa määrittää kvantiilien avulla. Jos q_s ja q_t ovat jakauman kvantiileja tasoilla s ja t, niin tällöin

$$\int_{q_s}^{q_t} f(x)\,dx \;=\; \int_{-\infty}^{q_t} f(x)\,dx - \int_{-\infty}^{q_s} f(x)\,dx \;=\; t - s,$$

joten väli $[q_s, q_t]$ on jakauman todennäköisyysväli tasolla $t-s$. Tason 50% todennäköisyysvälejä ovat siis esimerkiksi $[q_{0.25}, q_{0.75}]$ ja $[q_{0.49}, q_{0.99}]$. Näistä $[q_{0.25}, q_{0.75}]$ on symmetrinen.

Bayesläisessä tilastollisessa päättelyssä väliestimaatteja voidaan laatia posteriorijakauman todennäköisyysvälejä käyttämällä. Luonnollinen vaihtoehto on määrittää symmetrinen todennäköisyysväli jollain riittävän suurella tasolla, esimerkiksi $\alpha = 0.95$.

Esimerkki 10.4 (Tuntematon kolikko)**.** Kolikkoa, josta ei ennalta tiedetä mitään, heittämällä havaitaan 4 kruunaa ja 1 klaava. Määritä 95% todennäköisyysväli kruunan todennäköisyyttä mallintavalle parametrille θ.

Kun priorijakaumaksi valitaan yksikkövälin tasajakauma, saadaan posteriorijakaumaksi (esimerkki 10.1) betajakauma parametreina $a = 5$ ja $b = 2$, tiheysfunktiona

$$p(\theta \mid \vec{x}) \;=\; \begin{cases} 30\,\theta^4(1-\theta), & \theta \in (0,1), \\ 0, & \text{muuten.} \end{cases}$$

Posteriorijakauman odotusarvoksi saatiin kolmen desimaalin tarkkuudella $\theta_{\mathrm{E}}(\vec{x}) = 0.714$ ja moodiksi $\theta_{\mathrm{M}}(\vec{x}) = 0.800$. Posteriorijakauman symmetrinen 95% todennäköisyysväli on väli $[q_{0.025}, q_{0.975}]$, jonka päätepisteet ovat tasojen 0.025 ja 0.975 kvantiilit betajakaumalle parametreina $a = 5$ ja $b = 2$. Betajakauman taulukoista tai numeerisilla ohjelmistoilla kvantiileiksi saadaan $q_{0.025} = 0.359$ ja $q_{0.975} = 0.957$.

■

10.3 Binaarimallin Bayes-estimointi

Aiempien mielipidemittausten perusteella uskotaan, että nykyisen presidentin kannatusosuus tulevissa vaaleissa on noin 0.4 ja sisältyy välille $[0.3, 0.5]$ noin 95% todennäköisyydellä. Uudessa 200 henkilön satunnaisotokseen pohjautuvassa mielipidemittauksessa havaittiin ehdokkaan kannatusosuudeksi 0.35. Määritä kannatusosuuden piste-estimaatti, joka ottaa huomioon ennakkouskomuksen ja havaitun mielipidemittauksen.

Bayesläisen piste-estimaatin laskemiseksi tulee määrittää tehtävässä kuvattua ennakkouskomusta kuvaava priorijakauma $p(\theta)$, jonka odotusarvo on

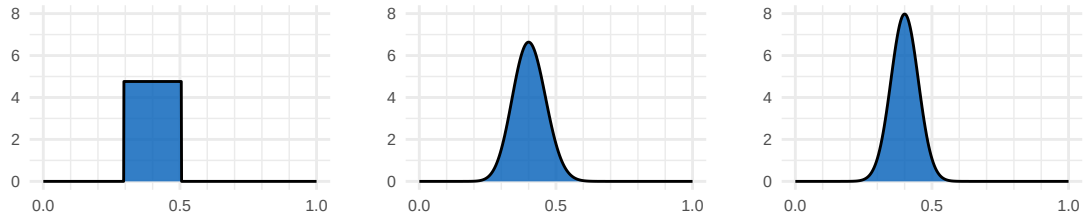
$$\int_{-\infty}^{\infty} \theta\, p(\theta) d\theta \;\approx\; 0.4, \tag{10.1}$$

ja välin $[0.3, 0.5]$ todennäköisyys on

$$\int_{0.3}^{0.5} p(\theta)d\theta \;\approx\; 0.95 \tag{10.2}$$

Ylläolevat ehdot eivät määrää priorijakaumaa yksikäsitteisesti. Esimerkiksi seuraavat jakaumat toteuttavat yo. ehdot:

1. Lukuvälin $[0.295, 0.505]$ tasajakauma.
2. Normaalijakauma odotusarvona $\mu = 0.4$ ja keskihajontana $\sigma = 0.05$.
3. Betajakauma parametreina $a = 27$ ja $b = 40$.



Tasajakauma ($a = 0.295$, $b = 0.505$) | Normaalijakauma ($\mu = 0.4$, $\sigma = 0.05$) | Betajakauma ($a = 27$ ja $b = 40$)

Ylläolevista jakaumista tasajakauman käyttö on arveluttavaa, sillä siinä priorijakauma antaa todennäköisyyden nolla välin $[0.295, 0.505]$ ulkopuolelle, mikä vastaa absoluuttista ennakkouskomusta, että parametri 100% varmuudella sisältyy välille $[0.295, 0.505]$. Normaalijakauma periaatteessa sallii parametrin arvoksi lukuja välin $[0, 1]$ ulkopuolelta, mutta käytännössä tällaiset kokoluokkaa 8σ tai suuremmat poikkeamat ovat hyvin epätodennäköisiä. Betajakauma on ylläolevista jakaumista luontevin valinta priorijakaumaksi, sillä sen arvojoukko on lukuväli $[0, 1]$ ja se antaa positiivisen todennäköisyyden kaikille yksikkövälin avoimille epätyhjille osaväleille. Betajakauma on myös käytännön laskennan kannalta mukava, sillä binaarimallissa se päivittyy betajakaumaksi (lause 9.3).

Kun havaitaan $n = 200$ alkion datajoukko, jossa 70 ykköstä ja 130 nollaa (ykkösten osuus = 35%), tällöin betajakauma ($a = 27$ ja $b = 40$) päivittyy betajakaumaksi parametreina $a + 70 = 97$ ja $b + 130 = 170$. Posteriorijakauman odotusarvoksi saadaan

$$\theta_{\mathrm{E}}(\vec{x}) \;=\; \frac{97}{97 + 170} \;\approx\; 0.363.$$

Etsitään piste-estimaatin 0.363 ympäriltä väli, johon posteriorijakaumaa noudattava parametri sisältyy tn:llä 95%. Voidaan valita esimerkiksi symmetrinen väli $[q_{0.025}, q_{0.975}]$. Betajakauman taulukoista $q_{0.025} = 0.307$ ja $q_{0.975} = 0.422$. Johtopäätöksenä voidaan ilmoittaa, että ennakkouskomuksen ja havaitun datan valossa sisältyy kannatusosuus välille $[0.307, 0.422]$ todennäköisyydellä 95%.

Luku 11

Tilastolliset testit

11.1 Nollahypoteesi ja p-arvo

Aiemmissa luvuissa opittiin määrittämään piste-estimaatteja ja väliestimaatteja tilastollisen mallin tuntemattomalle parametrille. Monissa käytännön tilanteissa pelkkä estimaatin määrittäminen ei riitä, vaan havaintojen pohjalta täytyy tehdä johtopäätös, pitääkö jokin mallia tai sen parametria koskeva asia paikkaansa vai ei. *Tilastollinen testi* on systemaattinen menetelmä laatia tällaisia johtopäätöksiä. Tilastollisen testin lähtökohtana on *nollahypoteesi* H_0, joka kuvastaa datalähteen oletusarvoista käyttäytymistä. Mikäli datalähteestä havaitut arvot poikkeavat paljon nollahypoteesin mukaisesta ennusteesta, antaa tämä aihetta nollahypoteesin hylkäämiseen. Tilastollisessa testissä on myös tapana määritellä *vastahypoteesi* H_1, joka yleensä vastaa nollahypoteesin vastakohtaa.

Havaintojen poikkeuksellisuutta analysoidaan laskemalla havaitusta datajoukosta $\vec{x} = (x_1, \ldots, x_n)$ tunnusluku

$$t(\vec{x}) \;=\; t(x_1, \ldots, x_n)$$

ja määrittämällä sitä vastaava p-arvo. Tilastollisen testin *p-arvo* on todennäköisyys, jolla nollahypoteesin mukaisen datalähteen generoimat satunnaisluvut $\vec{X} = (X_1, \ldots, X_n)$ tuottavat havaittua tunnuslukua $t(\vec{x})$ poikkeavamman tai yhtä poikkeavan arvon $t(\vec{X})$. Mikäli p-arvo on lähellä nollaa, johtuu havaittu poikkeama hyvin epätodennäköisesti satunnaisvaihtelusta ja antaa aiheen hylätä nollahypoteesi. Tällaista poikkeamaa kutsutaan *tilastollisesti merkitseväksi*. Useimmissa sovelluskonteksteissa on p-arvojen kokoa tapana luonnehtia allaolevan taulukon nyrkkisääntöjen avulla.

p-arvo	**Tulkinta**
> 0.10	Havainto ei ole ristiriidassa H_0:n kanssa
≈ 0.05	Havainto todistaa jonkun verran H_0:aa vastaan
< 0.01	Havainto todistaa vahvasti H_0:aa vastaan

Esimerkki 11.1 (Kolikko)**.** Tasaiseksi väitettyä kolikkoa 50 kertaa heitettäessä saadaan heittosarja, joka sisältää 42 kruunaa ja 8 klaavaa. Kuuluuko havaittu

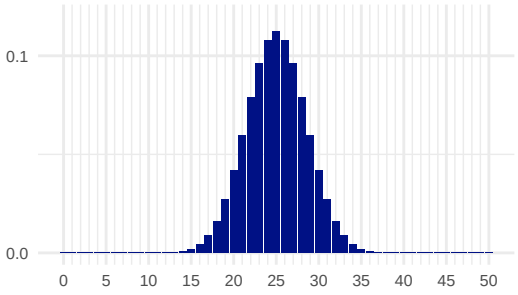
tulos tyypillisen satunnaisvaihtelun piiriin vai onko syytä epäillä kolikon tasaisuutta?

Heittotulosten (0=klaava, 1=kruuna) datalähde vastaa binaarimallia, jonka parametri θ on kruunan todennäköisyys. Ylläoleva kysymyksenasettelu vastaa tilastollista testiä, jonka nollahypoteesi ja vastahypoteesi ovat

$$H_0 : \theta = 0.5, \qquad H_1 : \theta \neq 0.5.$$

Kun testin tunnusluvuksi valitaan kruunien lukumäärä $t(\vec{x}) = \sum_{i=1}^{50} x_i$, noudattaa mallin mukaisista satunnaisluvuista $\vec{X} = (X_1, \ldots, X_n)$ laskettu tunnusluvun arvo $T = t(\vec{X})$ binomijakaumaa tiheysfunktiona

$$f_\theta(x) \;=\; \binom{50}{x} \theta^x (1-\theta)^{50-x}.$$



$H_0 : \theta = 0.5$

Nollahypoteesin vallitessa tunnusluvun T odotusarvo on 25, josta havaittu tunnusluvun arvo $t(\vec{x}) = 42$ poikkeaa 17 yksikköä. Testin p-arvo eli näin suuren tai vielä suuremman poikkeaman todennäköisyys nollahypoteesin vallitessa on

$$\text{p-arvo} \;=\; \mathbb{P}_{H_0}(|T - 25| \geq 17) \;=\; \sum_{x=0}^{8} f_{0.5}(x) + \sum_{x=42}^{50} f_{0.5}(x) \;=\; 1.2 \times 10^{-6}.$$

Näin lähellä nollaa oleva p-arvo antaa vahvan perusteen epäillä kolikon tasaisuutta ja hylätä nollahypoteesi. ■

Esimerkki 11.2 (Tiedonsiirtokanava)**.** Tiedonsiirtokanavan kohinan takia lähetetty lukuarvoinen signaali μ vastaanotetaan normaalijakautuneena satunnaismuuttujana, jonka odotusarvo on μ ja keskihajonta $\sigma = 3$. Eräänä päivänä lähettäjän uskotaan lähettävän arvon $\mu = 8$. Kun vastaanotettu arvo on $x_1 = 12.8$, onko syytä epäillä, että lähetetyn signaalin arvo ei ollutkaan 8?

Kysymyksenasettelua vastaa tilastollinen testi, jonka nollahypoteesi ja vastahypoteesi ovat

$$H_0 : \mu = 8, \qquad H_1 : \mu \neq 8.$$

Valitaan testisuureeksi keskihajonnalla normitettu poikkeama

$$t(\vec{x}) \;=\; \frac{x_1 - 8}{3} \;=\; 1.6.$$

Tällöin testisuureen arvoja mallintava satunnaismuuttuja $T = t(\vec{X})$ noudattaa nollahypoteesin vallitessa normitettua normaalijakaumaa kertymäfunktiona

$F_Z(x)$. Testin p-arvo eli todennäköisyys, että nollahypoteesin vallitessa T poikkeaa odotusarvostaan 1.6 yksikköä tai enemmän, saadaan symmetrian perusteella normitetun normaalijakauman taulukoista

$$\text{p-arvo} \;=\; \mathbb{P}_{H_0}(|T| \geq 1.6) \;=\; 2\mathbb{P}_{H_0}(T \geq 1.6) \;=\; 2(1 - F_Z(1.6)) \;=\; 0.11.$$

Koska p-arvo $> 10\%$, on havaittu poikkeama selitettävissä melkon hyvin tavanomaisella satunnaisvaihtelulla. Vastaanotettu signaalin arvo ei tarjoa syytä epäillä H_0:n paikkansapitävyyttä. ■

11.2 Yhdistetty nollahypoteesi

Seuraava esimerkki edustaa yhdistettyä nollahypoteesia, joka vaatii huolellisempaa analyysiä, sillä siinä nollahypoteesi ei yksiselitteisesti määritä datalähteen stokastisen mallin jakaumaa. Poikkeavien arvojen määrittämisessä pitää lisäksi ottaa huomioon, poikkeaako havaittu testisuureen arvo ylöspäin vai alaspäin tyypillisestä arvosta.

Esimerkki 11.3 (Laadunvalvonta)**.** Tukkukauppias väittää, että sen toimittamista tomaateista enintään 5% on huonolaatuisia. Suuresta tilauserästä poimittiin satunnaisesti 50 tomaattia ja niistä 4 todettiin huonolaatuisiksi. Kuuluuko tehty havainto tyypillisen satunnaisvaihtelun piiriin vai onko syytä epäillä tukkukauppiaan väitettä?

Tarkastettujen tomaattien laatuarvioita (0=Hyvä, 1=Huono) voidaan mallintaa Bernolli-jakaumalla, jonka (tuntematon) parametri θ vastaa huonolaatuisten tomaattien osuutta koko tilauserässä. Kysymyksenasettelua vastaa tilastollinen testi, jonka nollahypoteesi ja vastahypoteesi ovat

$$H_0 : \theta \leq 0.05, \qquad H_1 : \theta > 0.05.$$

Valitaan testin tunnusluvuksi huonolaatuisten tomaattien lukumäärä $t(\vec{x}) = \sum_{i=1}^{50} x_i$. Koska tarkastetut 50 tomaattia on poimittu satunnaisotannalla suuresta populaatiosta, noudattaa tunnusluvun arvoja mallintava satunnaismuuttuja $T = t(\vec{X})$ suurella tarkkuudella binomijakaumaa tiheysfunktiona

$$f_\theta(x) \;=\; \binom{50}{x} \theta^x \, (1-\theta)^{50-x}, \quad x = 0, 1, \ldots, 50.$$

Ylläolevan yksisuuntaisen nollahypoteesin näkökulmasta poikkeavat havainnot ovat niitä, joiden tunnusluku on *suuri*. Todennäköisyys, että nollahypoteesin mukaiset satunnaisluvut $\vec{X} = (X_1, \ldots, X_n)$ tuottavat havaittua tunnusluvun arvoa $t(\vec{x}) = 4$ poikkeavampia tai yhtä poikkeavia testisuureen arvoja on siis

$$\mathbb{P}_\theta \left(T \geq t(\vec{x}) \right) \;=\; \sum_{x=4}^{50} f_\theta(x).$$

Ongelmana on, että tässä tapauksessa nollahypoteesi ei suoraan määritä datalähteen satunnaislukujen eikä niitä vastaavan tunnusluvun jakaumaa. Tällaisessa tilanteessa p-arvo määritellään kaavalla

$$\text{p-arvo} \;=\; \max_{\theta \le 0.05} \mathbb{P}_\theta(T \ge t(\vec{x})).$$

Koska ylläoleva todennäköisyys maksimoituu[1] arvolla $\theta = 0.05$, saadaan p-arvoksi

$$\text{p-arvo} \;=\; \sum_{x=4}^{50} f_{0.05}(x) \;=\; \sum_{x=4}^{50} \binom{50}{x} 0.05^x \, (1-0.05)^{50-x} \;=\; 0.24,$$

mikä ei anna aihetta hylätä nollahypoteesiä. ■

11.3 Testausvirheet

Tietyissä tilanteissa testaajalta vaaditaan yksiselitteistä johtopäätöstä: testin pohjalta H_0 joko hyväksytään tai hylätään. Johtopäätöksen pohjaksi valitaan testin *merkitsevyystaso* $\alpha \in (0, 1)$ ja johtopäätös muodostetaan seuraavasti:

- Jos p-arvo $\ge \alpha$, nollahypoteesi hyväksytään (jätetään voimaan),
- Jos p-arvo $< \alpha$, nollahypoteesi hylätään.

Näin menetellessä mikään ei takaa, että tehty johtopäätös olisi oikea.

- Esimerkin 11.1 kolikonheitossa havaittiin 42 kruunaa p-arvona $\approx 10^{-6}$, joten nollahypoteesi (tasainen kolikko) hylätään merkitsevyystasolla $\alpha = 0.01$. On kuitenkin periaatteessa mahdollista, että tasaisella kolikolla satuttiin tekemään äärimmäisen epätodennäköinen 42 kruunan heittosarja. Tällöin tehdään *hylkäysvirhe*.
- Esimerkin 11.2 tiedonsiirtokanavassa saatiin p-arvoksi 11%, joten nollahypoteesi (lähetetty signaalin arvo on 8) hyväksytään merkitsevyystasolla $\alpha = 0.01$. On kuitenkin täysin mahdollista, että havaittu virhe onkin mitattu kanavasta, jossa kohinan aiheuttamat virheet noudattavat normaalijakaumaa esim. odotusarvona 2 ja keskihajontana 3. Tällöin tehdään *hyväksymisvirhe*.

Eri tavat tehdä oikea tai virheellinen johtopäätös voidaan taulukoida seuraavasti:

[1] Yläraja voidaan perustella niin, että tulkitaan T kruunien lukumääräksi 50 kolikon heittosarjassa, jossa kruunan todennäköisyys on θ. Kun kruunan todennäköisyys kasvaa, kasvaa myös todennäköisyys että havaittujen kruunien lukumäärä ylittää kynnysarvon 4.

	Johtopäätös	
Totuus	H_0 hyväksytään	H_0 hylätään
H_0 tosi	Oikea päätös	Hylkäysvirhe
H_0 epätosi	Hyväksymisvirhe	Oikea päätös

Seuraava keskeinen tulos takaa, että hylkäysvirheen todennäköisyys voidaan säätää pieneksi asettamalla riittävän pieni merkitsevyystaso. Tuloksen todistus on esitetty luvussa 11.5.

Lause 11.4. *Tilastollisen testin hylkäysvirheen todennäköisyys on korkeintaan testin merkitsevyystaso α.*

Hyväksymisvirheen todennäköisyyttä sen sijaan on yleisesti vaikea kontrolloida. Tästä syystä nollahypoteesi on syytä muotoilla niin, että se vastaa yleistä vallalla olevaa käsitystä tutkittavasta asiasta, tai niin että virheellisen nollahypoteesin hyväksymisellä ei ole vakavia seuraamuksia.

Esimerkki 11.5 (Rikosoikeus)**.** Yksityishenkilön epäiltyä rikosta koskevassa oikeudenkäynnissä tulee päättää, onko syytetty henkilö syytön vai syyllinen saatavilla olevan datan perusteella. Yleensä nollahypoteesi asetetaan muodossa

$$H_0\text{: Epäilty henkilö on syytön.}$$

Nollahypoteesiä H_0 vastaava hyväksymisvirhe vastaa syyllisen henkilön tuomitsematta jättämistä ja hylkäysvirhe syyttömän henkilön tuomitsemista. Hylkäysvirheen todennäköisyys saadaan lauseen 11.4 perusteella pieneksi valitsemalla testille riittävän pieni merkitsevyystaso. Hyväksymisvirheen todennäköisyyttä on sen sijaan vaikeampi kontrolloida.

Toinen tapa valita nollahypoteesi olisi asettaa

$$H_0'\text{: Epäilty henkilö on syyllinen.}$$

Tällöin tilastollisessa testissä ei syyttömien henkilöiden tuomitsemisen riskiä (H_0':n hyväksymisvirhe) voisi kunnolla kontrolloida, mikä olisi ristiriidassa yleisen oikeusperiaatteen kanssa, jonka mukaan epäilty on syytön kunnes toisin todistetaan. ■

Esimerkki 11.6 (Tupakkatuote)**.** Jos halutaan tutkia uudentyyppisen tupakkatuotteen vaikutusta terveyteen, voidaan nollahypoteesi esittää joko muodossa

$$H_0\text{: Uudella tupakkatuotteella ei ole terveyttä edistäviä vaikutuksia}$$

tai muodossa

$$H_0'\text{: Uudella tupakkatuotteella on terveyttä edistäviä vaikutuksia}$$

Nollahypoteesiä H_0 vastaavan hyväksymisvirheen tekeminen olisi tupakkatuotteen valmistajan kannalta haitallista, mutta yhteiskunnalliset vaikutukset eivät luultavasti olisi suuria. Nollahypoteesin H_0 hylkäysvirheellä saattaisi olla

yhteiskunnalle hyvinkin haitalliset seuraukset. Hylkäysvirheen todennäköisyys saadaan lauseen 11.4 perusteella pieneksi valitsemalla testille riittävän pieni merkitsevyystaso.

Jos nollahypoteesiksi valittaisiinkin H_0', niin merkitsevyystasoa pieneksi säätämällä voitaisiin kontrolloida tupakkavalmistajan epäreilun kohtelun (H_0':n hylkäysvirhe) riskiä, mutta ei yhteiskunnallisten terveyshaittojen (H_0':n hyväksymisvirhe) riskiä. ■

Merkitsevyystason valinta vaikuttaa testausvirheiden todennäköisyyteen seuraavasti:

- Henkilö A, joka elämänsä aikana tekee suuren määrän hypoteesitestejä merkitsevyystasolla $\alpha = 5\%$ on henkisesti varautunut siihen, että tietty osuus testien johtopäätöksistä on virheellisiä. Hän myös tietää, että pitkällä tähtäyksellä hänen hylkäämistään nollahypoteeseista enintään 5% on virheellisesti hylätty. (Hän ei kuitenkaan tiedä mitkä niistä.)

- Henkilö B, joka tekee elämänsä aikana suuren määrän hypoteesitestejä merkitsevyystasolla $\alpha = 1\%$ on myös varautunut siihen, että tietty osuus testipäätelmistä on virheellisiä. Hän tietää, että hänen hylkäämistään nollahypoteeseista enintään 1% on virheellisesti hylätty. Lisäksi hän tietää, että hänen hyväksymistään nollahypoteeseista on suurempi osuus virheellisiä kuin henkilöllä A.

Henkilö B on taipuvaisempi harvemmin hylkäämään nollahypoteeseja, minkä johdosta hän tekee pitkällä tähtäyksellä harvemmin hylkäysvirheitä mutta useammin hyväksymisvirheitä.

Esimerkki 11.7 (Tuntematon kolikko)**.** Tasaiseksi väitettyä kolikkoa 10 kertaa heitettäessä (0=klaava, 1=kruuna) havaitaan data $\vec{x} = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0)$. Testaa väitteen paikkansapitävyyttä 5% merkitsevyystasolla ja analysoi hylkäys- ja hyväksymisvirheiden todennäköisyyksiä.

Testin nollahypoteesi ja vastahypoteesi ovat

$$H_0 : \theta = 0.5, \qquad H_1 : \theta \neq 0.5,$$

missä θ on kruunan todennäköisyys. Valitaan testin tunnusluvuksi kruunien lukumäärä $t(\vec{x}) = x_1 + \cdots + x_{10}$. Datalähteen tuottamia tunnusluvun arvoja mallintava satunnaismuuttuja $T = t(\vec{X})$ noudattaa binomijakaumaa tiheysfunktiona

$$f_\theta(x) \;=\; \binom{10}{x}\theta^x(1-\theta)^{10-x}, \quad x = 0, 1, \ldots, 10.$$

Nollahypoteesin vallitessa havaittu tunnusluvun arvo $t(\vec{x}) = 1$ poikkeaa 4 yksikköä binomijakauman odotusarvosta $10\theta = 5$. Näin ollen p-arvoksi saadaan

$$\text{p-arvo} \;=\; \mathbb{P}_{H_0}(\,|T-5| \geq 4) \;=\; \sum_{x=0}^{1} f_{0.5}(x) + \sum_{x=9}^{10} f_{0.5}(x) \;\approx\; 2.1\%.$$

Koska p-arvo alittaa valitun merkitsevyystason 5%, nollahypoteesi hylätään.

Testausvirheiden analysoimiseksi määritetään ensiksi testin *hylkäysalue* eli nollahypoteesin hylkäämiseen johtavien tunnusluvun arvojen joukko. Toistamalla ylläolevat laskelmat eri testisuureen arvoille voidaan testin mahdolliset p-arvot taulukoida seuraavasti:

Kruunien lkm	0	1	2	3	4	5	6	7	8	9	10
p-arvo (%)	0.2	2.1	10.9	34.4	75.4	100	75.4	34.4	10.9	2.1	0.2

Taulukon mukaan testin hylkäysalue 5% merkitsevyystasolla on joukko $\{0, 1, 9, 10\}$. Hylkäysvirheen todennäköisyys on siis

$$\mathbb{P}_{H_0}(T \in \{0, 1, 9, 10\}) \;=\; \sum_{x=0}^{1} f_{0.5}(x) + \sum_{x=9}^{10} f_{0.5}(x) \;\approx\; 2.1\%.$$

Hyväksymisvirheen todennäköisyyttä on sen sijaan käytännössä mahdotonta analysoida, sillä vastahypoteesi $\theta \neq 0.5$ ei kerro juuri mitään datalähteen jakaumasta. Esimerkiksi jos parametrin todellinen arvo olisi $\theta = 0.499$, saataisiin hyväksymisvirheen todennäköisyydeksi

$$\mathbb{P}_{H_1}(T \in \{2, \ldots, 8\}) \;=\; \sum_{x=2}^{8} f_{0.499}(x) \;\approx\; 0.979.$$

Jos taas parametrin todellinen arvo olisi $\theta = 0.001$, saataisiin hyväksymisvirheen todennäköisyydeksi

$$\mathbb{P}_{H_1}(T \in \{2, \ldots, 8\}) \;=\; \sum_{x=2}^{8} f_{0.001}(x) \;\approx\; .00004.$$

■

11.4 Odotusarvon testi suurelle datajoukolle

Tarkastellaan datalähdettä, jonka oletetaan tuottavan toisistaan riippumattomia ja samoin jakautuneita satunnaislukuja $X_1, X_2, \ldots$ odotusarvona μ ja keskihajontana σ. Satunnaislukujen jakauma on tuntematon, jolloin myös μ ja σ ovat tuntemattomia. Halutaan testata nollahypoteesia

$$H_0\colon \mu = \mu_0,$$

missä μ_0 on oletettu odotusarvoparametrin μ arvo. Mikäli datalähteen arvojen jakaumasta ei tiedetä mitään, vaikuttaa mahdottomalta laatia toimivaa testiä ylläolevalle hypoteesille. Jos datalähteestä on saatu kerättyä suuri määrä dataa, voidaan toimiva testi kuitenkin muodostaa.

Havaitusta suuresta datajoukosta $\vec{x} = (x_1, \ldots, x_n)$ lasketaan ensiksi keskiarvo ja keskihajonta kaavoilla

$$m(\vec{x}) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{ja} \qquad \mathrm{sd}(\vec{x}) \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - m(\vec{x}))^2\right)^{1/2},$$

ja sen jälkeen testisuureeksi määritellään datajoukon keskiarvon normitettu poikkeama väitetystä odotusarvosta

$$t(\vec{x}) \;=\; \frac{m(\vec{x}) - \mu_0}{\mathrm{sd}(\vec{x})/\sqrt{n}}.$$

Allaolevan tuloksen perusteella testin p-arvo suurilla n:n arvoilla on likimain

$$\mathbb{P}_{H_0}(|t(\vec{X})| \geq |t(\vec{x})|) \;\approx\; \mathbb{P}(|Z| \geq |t(\vec{x})|),$$

missä Z noudattaa normitettua normaalijakaumaa. Testisuureen likiarvoiseksi p-arvoksi saadaan siis

$$p(\vec{x}) \;=\; 2(1 - F_Z(|t(\vec{x})|),$$

missä $F_Z(t)$ on normitetun normaalijakauman kertymäfunktio.

Lause 11.8. *Yllä kuvatun datalähteen tuottamia testisuureen arvoja mallintavan satunnaismuuttujan $t(\vec{X})$ jakauma suppenee nollahypoteesin vallitessa kohti normitettua normaalijakaumaa, kun $n \to \infty$.*

Todistus. Keskeisen raja-arvolauseen (lause 5.9) seurauksena todellisella keskihajontaparametrilla normitettu keskiarvon poikkeama

$$\frac{m(\vec{X}) - \mu_0}{\sigma/\sqrt{n}}$$

on jakaumaltaan lähellä normitettua normaalijakaumaa. Soveltamalla suurten lukujen lakia (lause 3.3) satunnaislukuihin $X_1^2, X_2^2, \ldots$ voidaan lisäksi perustella, että nollahypoteesin vallitessa

$$\frac{\sigma}{\mathrm{sd}(\vec{X})} \;\approx\; 1$$

suurella todennäköisyydellä, kun $n \to \infty$. Väite seuraa näistä havainnosta käyttämällä nk. Slutskyn lemmaa [vdV00]. □

Esimerkki 11.9 (Kahviautomaatti)**.** Kahviautomaatin on tarkoitus laskea jokaiseen kuppiin keskimäärin 10.0 cl kahvia. Laitteen toimintaa testattiin valuttamalla automaatista 30 kupillista ja mittaamalla kahvin määrät kupeissa. Mittauksessa havaittiin arvot (cl):

11.05 9.65 10.93 9.46 10.27 10.02 10.07 10.74 11.15 10.40 10.12 11.20 10.07 10.27 9.99
9.80 10.83 10.21 11.26 10.11 10.49 10.10 10.15 11.02 10.00 11.68 10.51 11.20 11.29 10.15

Onko kahviautomaatti oikein kalibroitu?

Merkitään kahviautomaatin tuottamien kahvikupillisten keskiarvoa parametrilla μ (tuntematon) ja väitettyä keskiarvoa μ_0. Kysymys voidaan tulkita tilastollisena testinä, jossa nollahypoteesi ja vastahypoteesi ovat

$$H_0\colon \mu = 10.0 \qquad H_1\colon \mu \neq 10.0$$

Mittausdatan $\vec{x}$ keskiarvo on $m(\vec{x}) = 10.473$ ja keskihajonta $\mathrm{sd}(\vec{x}) = 0.554$. Havaitun datajoukon normitettu poikkeama on näin ollen

$$t(\vec{x}) \;=\; \frac{m(\vec{x}) - \mu_0}{\mathrm{sd}(\vec{x})/\sqrt{n}} \;=\; \frac{10.473 - 10.0}{0.554/\sqrt{30}} \;=\; 4.68.$$

Koska $n = 30$ on kohtalaisen suuri luku, sovelletaan suuren datajoukon likiarvoista testiä, jolloin p-arvoksi saadaan

$$\text{p-arvo} \;\approx\; \mathbb{P}_{H_0}(|t(\vec{X})| \geq |t(\vec{x})|) \;\approx\; \mathbb{P}(|Z| \geq 4.68) \;\approx\; 2.9 \times 10^{-6}.$$

Näin pieni p-arvo johtaa H_0:n hylkäämiseen kaikilla yleisesti käytetyillä merkitsevyystasoilla. ■

11.5 Hylkäysvirheen todennäköisyyden analyysi

Tässä kappaleessa esitetään lauseen 11.4 todistus. Todistus vaatii reaalianalyysin yksityiskohtia, jotka eivät ole tilastollisen päättelyn kannalta keskeisiä. Näin ollen tämän todistuksen voi kiireinen lukija sivuuttaa. Todistuksen pohjaksi perustellaan ensiksi seuraava aputulos.

Lemma 11.10. *Jos $S_\alpha = \{s \in \mathbb{R} : \mathbb{P}(Z \geq s) \leq \alpha\}$ on niiden arvojen joukko, joita suurempia tai yhtä suuria arvoja reaaliarvoinen satunnaismuuttuja Z saa enintään todennäköisyydellä $\alpha \in (0, 1)$, niin*

$$\mathbb{P}(Z \in S_\alpha) \;\leq\; \alpha.$$

Todistus. Koska todennäköisyyden monotonisuuden perusteella funktio $s \mapsto \mathbb{P}(Z \geq s)$ on vähenevä, voidaan päätellä, että joukko S_α on lukuväli muotoa $[s_\alpha, \infty)$ tai (s_α, ∞), missä luku s_α on joukon S_α suurin alaraja.

(i) Tapauksessa $S_\alpha = [s_\alpha, \infty)$ luku s_α sisältyy joukkoon S_α, joten joukon S_α määritelmän mukaan

$$\mathbb{P}(Z \in S_\alpha) \;=\; \mathbb{P}(Z \geq s_\alpha) \;\leq\; \alpha.$$

(ii) Tapauksessa $S_\alpha = (s_\alpha, \infty)$ tehdään vastaoletus, että $\mathbb{P}(Z \in S_\alpha) > \alpha$. Tällöin $\mathbb{P}(Z > s_\alpha) > \alpha$. Todennäköisyyden jatkuvuuden perusteella tiedetään, että $\mathbb{P}(Z > s_\alpha) = \lim_{s \downarrow s_\alpha} \mathbb{P}(Z \geq s)$. Näin ollen $\mathbb{P}(Z \geq s) > \alpha$ jollain $s > s_\alpha$. Tästä seuraa että luku s ei kuulu joukkoon S_α, joten s on kyseisen joukon alaraja. Tämä on looginen ristiriita, sillä s_α määriteltiin joukon S_α suurimmaksi alarajaksi. Vastaoletus on siis epätosi, ja pätee $\mathbb{P}(Z \in S_\alpha) \leq \alpha$. □

Lauseen 11.4 todistus. Merkitään satunnaismuuttujalla $T = t(\vec{X})$ nollahypoteesin mukaisen datalähteen tuottamasta datajoukosta laskettua testisuureen arvoa ennen datan havaitsemista. Olkoon $t_0 = \mathbb{E}_{H_0}(T)$ kyseisen testisuureen odotusarvo. Tällöin nollahypoteesi havaitulle datajoukolle $\vec{x}$ hylätään täsmälleen silloin, kun sitä vastaavan testisuureen poikkeama $|t(\vec{x}) - t_0|$ sisältyy joukkoon

$$S_\alpha \;=\; \Big\{ s \in \mathbb{R} : \mathbb{P}_{H_0}(|T - t_0| \geq s) < \alpha \Big\}.$$

Merkitsemällä joukon S_α sulkeumaa

$$\bar{S}_\alpha \;=\; \Big\{ s \in \mathbb{R} : \mathbb{P}_{H_0}(|T - t_0| \geq s) \leq \alpha \Big\}$$

saadaan hylkäysvirheelle yläraja

$$\mathbb{P}_{H_0}(H_0 \text{ hylätään}) \;=\; \mathbb{P}_{H_0}(|T - t_0| \in S_\alpha) \;\leq\; \mathbb{P}_{H_0}(|T - t_0| \in \bar{S}_\alpha).$$

Soveltamalla lemmaa 11.10 satunnaismuuttujaan $|T - t_0|$ havaitaan, että yllä-olevan epäyhtälön oikea puoli on enintään α. □

Liite A

Todennäköisyysjakaumia

Tähän liitteeseen on koostettu lista tärkeimmistä todennäköisyysjakaumista.

A.1 Yksiulotteisia diskreettejä jakaumia

A.1.1 Dirac-jakauma

Dirac-jakauma on diskreetti jakauma, joka kuvaa satunnaismuuttujaa, jossa ei ole lainkaan satunnaisvaihtelua eli sen keskihajonta on nolla. Dirac-jakaumalla on yksi parametri $\theta \in (-\infty, \infty)$ ja sen tiheysfunktio on

$$f(x) \;=\; \begin{cases} 1, & x = \theta, \\ 0, & \text{muuten.} \end{cases}$$

Dirac-jakaumaa merkitään usein δ_θ.

A.1.2 Bernoullijakauma

Tämä on joukon $\{0, 1\}$ yleinen diskreetti jakauma. Sen parametrit voidaan kirjoittaa muodossa (p_0, p_1), missä $p_0 + p_1 = 1$. Käytännössä ilmoitetaan vain $p = p_1$. Kuvaa yksittäisen tapahtuman indikaattorimuuttujaa. Arvojoukko $\{0, 1\}$, parametrit $p \in [0, 1]$. Tiheysfunktio on

$$f(x) \;=\; (1-p)^{1-x} p^x \;=\; \begin{cases} 1 - p, & x = 0, \\ p, & x = 1. \end{cases}$$

Jokainen $\{0, 1\}$-arvoinen satunnaismuuttuja X noudattaa Bernoulli-jakaumaa parametrina $p = \mathbb{P}(X = 1)$.

A.1.3 Multinoullijakauma

Tämä on joukon $\{1, \ldots, k\}$ yleinen diskreetti jakauma, sen parametrit ovat luvut $p_1, \ldots, p_k \geq 0$, joille $p_1 + \cdots + p_k = 1$. Tiheysfunktio on

$$f(x) \;=\; p_x, \quad x \in \{1, \ldots, k\}.$$

Erikoistapaus: $k = 1$ vastaa pisteen 1 Dirac-jakaumaa.

Erikoistapaus: $k = 2$ vastaa Bernoullijakaumaa, kunhan arvojoukko numeroidaan sopivasti.

A.1.4 Diskreetti tasajakauma

Äärellisen joukon A tasajakauman tiheysfunktio on

$$f(x) \;=\; \frac{1}{\#A}, \quad x \in A.$$

Erikoistapaus: $A = \{1, \ldots, k\}$ vastaa multinoullijakaumaa $p_j = \frac{1}{k}$ kaikilla $j = 1, \ldots, k$.

A.1.5 Binomijakauma

Binomijakauma on diskreetti jakauma joukossa $\{0, 1, \ldots, n\}$. Se kuvaa binaarisesta datalähteestä tuotettujen satunnaismuuttujien summan jakaumaa. Yleisesti se kuvaa kruunien lukumäärä n:llä kolikonheitolla, kun kruunan todennäköisyys on p. Jakauman parametrit ovat (n, p), missä $n \geq 1$ on kokonaisluku ja $p \in [0, 1]$ reaaliluku. Tiheysfunktio on

$$f(x) \;=\; \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \ldots, n\}.$$

Erikoistapaus: $n = 1$ on Bernoullijakauma parametrina p.

A.1.6 Geometrinen jakauma

Geometrisia jakaumia on kahdenlaisia. Molemmat liittyvät sarjaan yrityksiä, joissa jokaisella yrityksellä muista riippumattomasti onnistutaan todennäköisyydellä p.

Lukujoukon $\{0, 1, \ldots\}$ geometrinen jakauma kuvaa, kuinka monta epäonnistumista tapahtuu ennen ensimmäistä onnistumista. Tiheysfunktio

$$f(x) \;=\; (1-p)^x p, \quad x \in \{0, 1, \ldots, \}.$$

Lukujoukon $\{1, 2, \ldots\}$ geometrinen jakauma kuvaa, kuinka monta yritystä tulee suorittaa, kunnes onnistutaan ensimmäisen kerran. Tiheysfunktio

$$f(x) \;=\; (1-p)^{x-1} p, \quad x \in \{1, 2, \ldots, \}.$$

Yleistys on tilanne, jossa katsotaan yritysten tai epäonnistumisten lukumäärää, kunnes saadaan $r \geq 1$ onnistumista. Geometrisen jakauman yleistyksiä tähän tilanteeseen kutsutaan *negatiivisiksi binomijakaumiksi*, joista myös on useita eri versioita riippuen siitä, lasketaanko yrityksiä vai epäonnistumisia.

A.1.7 Hypergeometrinen jakauma

Populaatiossa on N alkiota, joista K on positiivisia. Kun poimitaan palauttamatta n alkion otos tasaisen satunnaisesti, niin montako positiivista saadaan? Positiivisten lukumäärän jakauma on joukon $\{0, \ldots, n\}$ diskreetti jakauma, tiheysfunktiona

$$f(x) \;=\; \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \quad x \in \{0, 1, \ldots, n\}.$$

Tämä on hypergeometrinen jakauma parametreina (N, K, n).

Jos sama satunnaisotanta suoritettaisiin palauttaen, niin positiivisten lukumäärän jakauma olisi binomijakauma parametreina n ja $p = K/N$. Silloin kun N on suuri suhteessa otoskokoon K, voidaan hypergeometrista jakaumaa arvioida edellä mainitulla binomijakaumalla.

A.1.8 Poisson-jakauma

Poisson-jakauma kertoo likiarvon onnistumisten lukumäärälle pitkässä sarjassa yrityksiä, joissa onnistumisia yhteensä sattuu odotusarvoisesti λ kappaletta. Sen voi myös tulkita kruunien lukumääräksi pitkässä $n \gg 1$ kolikonheiton sarjassa, jossa kruunan todennäköisyys on $\approx \lambda n^{-1}$. Tiheysfunktio

$$f(x) \;=\; e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, \ldots, \}.$$

A.2 Moniulotteisia diskreettejä jakaumia

A.2.1 Multinomijakauma

Multinomijakauma on k-ulotteinen diskreetti jakauma, joka kuvaa diskreetistä datalähteestä tuotettujen satunnaismuuttujien esiintyvyyksien yhteisjakaumaa. Multinomijakaumalla on parametrina kokonaisluvut $n, k \geq 1$ ja reaaliluvut $p_1, \ldots, p_k \geq 0$, joille $p_1 + \cdots + p_k = 1$. Tarkastellaan datalähdettä, joka tuottaa n riippumatonta satunnaislukua $(X_1, \ldots, X_n)$ jakaumasta $(p_1, \ldots, p_k)$. Merkitään arvon j esiintyvyyttä tulossarjassa symbolilla

$$N_j \;=\; \sum_{i=1}^{n} 1(X_i = j).$$

Tällöin esiintyvyyksien $(N_1, \ldots, N_k)$ yhteisjakauma on multinomijakauma tiheysfunktiona

$$f(x_1, \ldots, x_k) \;=\; \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, \quad x_1 + \cdots + x_k = n.$$

Erikoistapaus $k = 2$ vastaa binomijakaumaa parametreina n ja p_1.

Erikoistapaus $n = 1$ vastaa multinoullijakaumaa eli yleistä joukon $\{1, \ldots, k\}$ diskreettiä todennäköisyysjakaumaa.

Tämä jakauma voidaan myös tulkita niin, että n palloa heitetään k:hon koriin. Jokainen pallo, muista riippumatta osuu koriin j todennäköisyydellä p_j. Korien sisältämien pallojen lukumäärien yhteisjakauma on tällöin multinomijakauma.

Voidaan myös tulkita niin, että populaatiossa on k:n eri tyypin alkioita ja tyypin j alkioiden suhteellinen osuus on p_j. Tehdään n satunnaisotantaa palauttaen. Tällöin havaittujen tyyppien esiintyvyyksien (eri frekvenssien) yhteisjakauma on multinomijakauma.

A.2.2 Hypergeometrinen jakauma

Tämä on d-ulotteinen diskreetti jakauma, jonka parametrit ovat kokonaisluvut n ja $K_1, \ldots, K_d$. Jakauma kuvaa eri tyyppien esiintyvyyksien yhteisjakaumaa n:n alkion otoksessa, joka on poimittu satunnaisotannalla palauttamatta äärellisestä populaatiosta, jossa on K_i tyypin $i = 1, \ldots, d$ alkiota. Merkitään tyypin i esiintyvyyttä n:n alkion otoksessa symbolilla N_i. Tällöin satunnaisvektori $(N_1, \ldots, N_d)$ noudattaa hypergeometrista jakaumaa tiheysfunktiona

$$f(x_1, \ldots, x_d) \;=\; \frac{\binom{K_1}{x_1} \cdots \binom{K_d}{x_d}}{\binom{K}{n}}, \quad x_1 + \cdots + x_d = n,$$

missä $K = K_1 + \cdots + K_d$.

Huom. Jos sama otanta tehtäisiin palauttaen, niin jakaumana olisi multinomijakauma parametreina n ja $p_1, \ldots, p_d$, missä $p_i = K_i/K$.

A.3 Yksiulotteisia jatkuvia jakaumia

A.3.1 Jatkuva tasajakauma

Välin $[a, b]$ jatkuvalla tasajakaumalla on tiheysfunktio

$$f(x) \;=\; \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{muuten.} \end{cases}$$

A.3.2 Eksponenttijakauma

Eksponenttijakauman parametrina $\lambda > 0$ tiheysfunktio on

$$f(x) \;=\; \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & \text{muuten.} \end{cases}$$

A.3.3 Normaalijakauma

Normaalijakaumalla on kaksi parametria: odotusarvo $\mu \in (-\infty, \infty)$ ja keskihajonta $\sigma \in (0, \infty)$. Normaalijakauman tiheysfunktio on

$$f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Normitetun normaalijakauman odotusarvo on $\mu = 0$ ja keskihajonta $\sigma = 1$. Sen kertymäfunktion lukuarvoja on taulukoitu liitteeseen B.

Liite B

Normaalijakauman lukuarvoja

Allaolevaan taulukkoon on koottu lukuarvoja normitetun normaalijakauman kertymäfunktiolle

$$F_Z(x) \;=\; \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

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

Liite C

Merkintöjä

Merkintä	Nimitys	R	Python	Excel
$m(\vec{x})$	Keskiarvo	`mean()`	`np.mean()`	`AVERAGE()`
$\mathrm{sd}(\vec{x})$	Keskihajonta	`sqrt(1-1/n)*sd()`	`np.std()`	`STDEV.P()`
$\mathrm{sd}_s(\vec{x})$	Otoskeskihajonta	`sd()`	`np.std( ,ddof=1)`	`STDEV.S()`
$\mathrm{var}(\vec{x})$	Varianssi	`(1-1/n)*var()`	`np.var()`	`VAR.P()`
$\mathrm{var}_s(\vec{x})$	Otosvarianssi	`var()`	`np.var( ,ddof=1)`	`VAR.S()`
$\mathrm{cov}(\vec{x},\vec{y})$	Kovarianssi	`(1-1/n)*cov()`	`np.cov( ,ddof=0)[0][1]`	`COVARIANCE.P()`
$\mathrm{cov}_s(\vec{x},\vec{y})$	Otoskovarianssi	`cov()`	`np.cov( ,ddof=1)[0][1]`	`COVARIANCE.S()`
$\mathrm{cor}(\vec{x},\vec{y})$	Korrelaatio	`cor()`	`np.corrcoef()[0][1]`	`CORREL()`
$q_{0.5}(\vec{x})$	Mediaani	`median()`	`np.median()`	`MEDIAN()`
$q_{0.25}(\vec{x})$	Alakvartiili	`quantile( ,.25)`	`np.quantile( ,.25)`	`PERCENTILE.INC( ,.25)`
$q_{0.75}(\vec{x})$	Yläkvartiili	`quantile( ,.75)`	`np.quantile( ,.75)`	`PERCENTILE.INC( ,.75)`

Taulukko C.1: Datajoukkojen $\vec{x} = (x_1, \ldots, x_n)$ ja $\vec{y} = (y_1, \ldots, y_n)$ tunnuslukuja.

Merkintä	Nimitys
$\mathbb{E}(X), \mu_X$	Satunnaismuuttujan X jakauman odotusarvo
$\mathrm{SD}(X), \sigma_X$	Satunnaismuuttujan X jakauman keskihajonta
$\mathrm{Cov}(X,Y), \sigma_{X,Y}$	Satunnaismuuttujien X ja Y jakauman kovarianssi
$\mathrm{Cor}(X,Y), \rho_{X,Y}$	Satunnaismuuttujien X ja Y jakauman korrelaatio
$\mathrm{Var}(X), \sigma_X^2$	Satunnaismuuttujan X jakauman varianssi

Taulukko C.2: Todennäköisyysjakaumien tunnuslukuja.

Liite D

Suomi–englanti-sanasto

Alla tässä monisteessa esiintynyttä sanastoa englanniksi käännettynä. Monet tähän aihepiiriin liittyvät termit eivät kuitenkaan ole täysin vakiintuneita kummassakaan kielessä.

suomi	**englanti**
aksiooma	axiom
alakvartiili	lower quartile
alkio	element
Bayesin kaava	Bayes formula
bayesläinen malli	Bayesian model
Bernoulli-jakauma	Bernoulli distribution
binaarimalli	binary model
binomijakauma	binomial distribution
binomikerroin	binomial coefficient
Chebyshevin epäyhtälö	Chebyshev's inequality
datajoukko	data set
datakehikko	data frame
diskreetti jakauma	discrete distribution
diskreetti satunnaismuuttuja	discrete random variable
ehdollinen jakauma	conditional distribution
ehdollinen odotusarvo	conditional expectation
ehdollinen tiheysfunktio	conditional density function
ehdollinen todennäköisyys	conditional probability
eksponenttijakauma	exponential distribution
empiirinen jakauma	empirical distribution
(empiirinen) keskihajonta	empirical/population standard deviation
(empiirinen) kovarianssi	empirical/population covariance
empiirinen tiheysfunktio	empirical density function
empiirinen yhteisjakauma	empirical joint distribution
entropia	entropy
ergodinen	ergodic
esiintyvyys	frequency
esiintyvyysharha	base rate fallacy

esiintyvyystaulukko	contingency table
estimaatti	estimate
estimaattori	estimator
harha	bias
harhaton	unbiased
havainto	observation
havaittu	observed
histogrammi	histogram
hylkäysvirhe	rejection error, type I error, false positive
hyperparametri	hyperparameter
hyväksymisvirhe	acceptance error, type II error, false negative
indikaattorifunktio	indicator function
jakauma	distribution
jatkuva jakauma	continuous distribution
joukko	set, space
järjestetty lista	ordered list
järjestystunnusluku	order statistic
järjestämätön joukko	unordered set
kertoma	factorial
kertymäfunktio	cumulative distribution function, distribution function
keskeinen raja-arvolause	central limit theorem
keskiarvo	mean, average
keskihajonta	standard deviation
keskineliövirhe	mean squared error
komplementti	complement
konvoluutio	convolution
korrelaatio	correlation
korreloimaton	uncorrelated
korreloitu	correlated
kovarianssi	covariance
kvartiili	quartile
leikkaus	intersection
lineaarinen regressio	linear regression
lineaarinen riippuvuus	linear dependence
logaritminen uskottavuusfunktio	logarithmic likelihood function
luottamustaso	confidence level
luottamusväli	confidence interval
mediaani	median
merkitsevyystaso	significance level
merkitsevä	significant
mielipidekysely	opinion poll
mitallinen	measurable
momentti	moment
moniulotteinen	multidimensional, multivariate
moodi	mode
muuttuja	variable
nollahypoteesi	null hypothesis

normaaliapproksimaatio	normal approximation
normaalijakauma	normal distribution, Gaussian distribution
normitettu	normalised
normitettu normaalijakauma	standard normal distribution
odotusarvo	expectation, mean
osajoukko	subset
ositus	partition
otanta palauttaen	sampling with replacement
otanta palauttamatta	sampling without replacement
otos	sample
otoskeskiarvo	sample mean/average
otoskeskihajonta	sample standard deviation
otoskovarianssi	sample covariance
otosvarianssi	sample variance
p-arvo	p-value
parametrinen jakauma	parametric distribution
perusjoukko	sample space
pienimmän neliösumman menetelmä	least squares method
piste-estimaatti	point estimate
Poisson-approksimaatio	Poisson approximation
Poisson-jakauma	Poisson distribution
posteriorijakauma	posterior distribution
posterioritiheys	posterior density
priorijakauma	prior distribution
prioritiheys	prior density
prosentiili	percentile
puukaavio	tree diagram
regressio	regression
regressiosuora	regression line
reunajakauma	marginal distribution
reunatiheysfunktio	marginal density function
riippumattomuus	independence
riippuvuus	dependence
ristiintaulukointi	cross tabulation
ristitaulukko	contingency table
satunnaisilmiö	random phenomenon
satunnaiskenttä	random field
satunnaisluku	random number
satunnaismatriisi	random matrix
satunnaismuuttuja	random variable
satunnaismuuttujan muunnos	transformation of a random variable
satunnaisotanta	random sampling
satunnaisotos	random sample
satunnaisvektori	random vector
satunnaisverkko	random graph
stokastiikka	stochastics
stokastinen prosessi	stochastic process

stokastinen riippuvuus	stochastic dependence
suhteellinen esiintyvyys	relative frequency
suhteellinen osuus	relative proportion
supeta stokastisesti	converge in probability
suurimman uskottavuuden estimaatti	maximum likelihood estimate
suurten lukujen laki	law of large numbers
tapahtuma	event
tarkentuva	consistent
tasajakauma	uniform distribution
taulukko	table
testi	test
testisuure	test statistic
tiheysfunktio (diskreetin jakauman)	density function, probability mass function
tiheysfunktio (jatkuvan jakauman)	density function, probability density function
tilasto	statistics
tilastotiede	statistics
tilastollinen päättely	statistical inference
todennäköinen	probable, likely
todennäköisyys	probability
todennäköisyysjakauma	probability distribution
todennäköisyyslaskenta	probability
todennäköisyysmitta	probability measure
todennäköisyysteoria	probability theory
todennäköisyysväli	probability interval, credible interval
toteuma	realisation, outcome
tulojoukko	product set, product space
tunnusluku	statistic
uskomus	belief
uskottavuus	likelihood
uskottavuusfunktio	likelihood function
vaiheittainen päivittäminen	sequential updating
varianssi	variance
vastahypoteesi	alternative hypothesis
virhemarginaali	margin of error
väliestimaatti	interval estimate
väliestimaattori	interval estimator
yhdiste	union
yhdistetty hypoteesi	composite hypothesis
yhteisjakauma	joint distribution
yksinkertainen hypoteesi	simple hypothesis
yksiulotteinen	one-dimensional, univariate
yläkvartiili	upper quartile

Liite E

Lisälukemista

Stokastiikan ja tilastotieteen perusteet

[CM12] Elementary Decision Theory
Herman Chernoff and Lincoln E Moses
Dover, 364 sivua, 2. painos, 2012
Tämä ajaton klassikko alunperin vuodelta 1959 on uudelleen painettu Doverin kustantamana 2012. Kirjan luvut 1–4 ja 7–10 vastaavat hyvin läheisesti tämän luentomonisteen kokonaisuutta. Kirjan selkeä esitys sisältää minimaalisen määrän matemaattisia kaavoja, mutta paljon havainnollistavia esimerkkejä. Kirjan lukemiseen esitiedoiksi riittää lukion matematiikka. Lisäksi luvut 5–6 sisältävät hyödyllisiä asioita stokastisesta optimoinnista ja tilastollisesta päätösteoriasta.

[Ros14] Introduction to Probability and Statistics for Engineers and Scientists
Sheldon M Ross
Academic Press, 686 sivua, 5. painos, 2014
Tämä kirja vastaa perinteisen todennäisyyslaskennan ja tilastotieteen täyden lukukauden mittaista kurssia. Luvut 1–8 vastaavat läheisesti tämän monisteen asiakokonaisuutta. Luvut 9–15 puolestaan vastaavat melko tarkalleen Aalto-yliopiston tilastollisen päättelyn kandikurssin (MS-C1620) sisältöä.

Matemaattinen tilastotiede

[HMC18] Introduction to Mathematical Statistics
Robert V Hogg, Joseph W McKean, Allen T Craig
Pearson, 694 sivua, 8. painos, 2018
Tämä matemaattisen tilastotieteen perusteos on edennyt jo kahdeksanteen painokseensa. Kirja sisältää kattavan ja selkeän esityksen klassisen tilastollisen päättelyn ydinasioista, mm. tärkeimmät jakaumat (t, χ^2, F), estimaattoreiden tehokkuus, tyhjentävät tunnusluvut, regressioanalyysi, sekä robustit ja parametrittomat menetelmät.

[Was10] All of Statistics
Larry Wasserman
Springer, 442 sivua, 2010

Tämä kirja kattaa laajan valikoiman tilastotieteen faktoja, kaavoja ja menetelmiä tiiviissä paketissa. Tuloksia ja menetelmiä ei sen koommin selitetä, perustella eikä todisteta, minkä johdosta kirjaa ei voi suositella itseopiskeluun. Kirja on kuitenkin hyvin jäsennetty ja siitä löytää etsimänsä helposti. Näin ollen tämä on mainio käsikirja, josta voi nopeasti katsoa ja tarkistaa tärkeiden kaavojen ja menetelmien tarkan muodon.

Todennäköisyysteoria

[JP04] Probability Essentials
Jean Jacod, Philip Protter
Springer, 254 sivua, 2. painos, 2004
Tämä oppikirja kattaa stokastiikan kannalta keskeisen mitta- ja integraaliteorian perusteet (luvut 1–16), satunnaisjonojen ja todennäköisyysmittojen raja-arvoja koskevat päätulokset (luvut 17–21) sekä diskreettiaikaisten martingaalien teorian (luvut 22–28). Kirjan luvut 1–21 vastaavat melko tarkalleen Aalto-yliopiston todennäköisyysteorian kurssin (MS-E1600) sisältöä.

[Wil91] Probability with Martingales
David Williams
Cambridge University Press, 275 sivua, 1. painos, 1991
Tämä on todennäköisyysteorian oppikirja vastaa sisällöltään ja paksuudeltaan Jacodin ja Protterin kirjaa, mutta on selvästi nopeampitahtisempi. Tekstiä ja kaavoja on vähemmän, sillä monet yksityiskohdat esimerkiksi jatkuvien jakaumien tiheysfunktioista on sivuutettu ja jätetty lukijan itse mietittäviksi. Kirjassa on kolme osaa. Ensimmäinen osa (luvut 1–8) kattaa stokastiikan kannalta keskeisen mitta- ja integraaliteorian perusteet. Toinen osa (luvut 9–15) kattaa diskreettiaikaisten martingaalien teorian. Kolmas osa (luvut 16–18) käsittelee satunnaisjonojen ja todennäköisyysmittojen raja-arvoja. Kirjan ensimmäinen ja kolmas osa vastaavat melko tarkalleen Aalto-yliopiston todennäköisyysteorian maisterikurssin (MS-E1600) sisältöä.

[Kal02] Foundations of Modern Probability
Olav Kallenberg
Springer, 638 sivua, 2. painos, 2002
Tämä todennäköisyysteorian klassikko sisältää tiiviin ja laajan esityksen todennäköisyysteorian tärkeimmistä tuloksista. Kaikille kirjan tuloksille on esitetty lyhyet ja elegantit todistukset. Tiiviin ja abstraktin esitystavan johdosta kirjaa voi olla vaikea lukea, jollei ennalta tiedä mitä asiaa kirjasta haluaa opiskella. Toisaalta yksiin kansiin on tällä tavoin saatu valtava määrä sisältöä kattaen mm. diskreettiaikaiset ja jatkuva-aikaiset martingaalit, Brownin liikkeen, Lévy-prosessit, stokastisen analyysin ja Itō-laskennan perusteorian, semimartingaalit, Poisson-satunnaismitat jne. Kirja soveltuu käsikirjaksi niille, jotka ovat kiinnostuneet opiskelemaan stokastiikkaa kanditasoa syvällisemmin. Esitietoina vaaditaan differentiaali- ja integraalilaskennan peruskurssit sekä topologian peruskäsitteet.

Liite F

Satunnaislukujen generoiminen

F.1 Kvantiilifunktion avulla

Tarkastellaan kertymäfunktion F määrittämää jakaumaa. Oletetaan, että Q on lukujoukossa $(0,1)$ määritelty funktio, jolle kaikilla $x \in \mathbb{R}$ ja $u \in (0,1)$ pätee nk. Galois'n yhteys

$$u \leq F(x) \quad \text{jos ja vain jos} \quad Q(u) \leq x. \tag{F.1}$$

Voidaan todistaa, että jokaista (diskreetin tai jatkuvan jakauman) kertymäfunktiota F kohden on olemassa täsmälleen yksi ylläolevan ehdon toteuttava funktio Q. Tätä funktiota kutsutaan jakauman vasemmalta jatkuvaksi kvantiilifunktioksi.

Lause F.1. *Jos funktio Q toteuttaa ehdon (F.1) ja satunnaismuuttuja U noudattaa välin $(0,1)$ jatkuvaa tasajakaumaa, niin tällöin satunnaismuuttuja $Q(U)$ noudattaa jakaumaa F.*

Todistus. Koska jatkuvan tasajakaumaa noudattavan satunnaismuuttujan U arvot kuuluvat joukkoon $(0,1)$ todennäköisyydellä yksi, ja lisäksi $\mathbb{P}(U \leq t) = t$ kaikilla $0 \leq t \leq 1$, havaitaan ehdon (F.1) avulla, että kaikilla x pätee

$$\mathbb{P}(Q(U) \leq x) \;=\; \mathbb{P}(U \leq F(x)) \;=\; F(x).$$

Näin ollen satunnaismuuttujan $Q(U)$ kertymäfunktio on F. Koska kertymäfunktio määrittää jakauman yksikäsitteisesti, väite seuraa. □

Jos kertymäfunktiolla on olemassa käänteisfunktio F^{-1}, niin tällöin $Q = F^{-1}$. Seuraavissa esimerkeissä käänteisfunktiota ei ole olemassa, sillä kertymäfunktio ei ole injektio. Vasemmalta jatkuva kvantiilifunktio on kuitenkin helppo määrittää.

Esimerkki F.2 (Eksponenttijakauma)**.** Eksponenttijakauma parametrina $\lambda > 0$ on jatkuva jakauma, jonka tiheysfunktio on

$$f(x) \;=\; \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{muuten}, \end{cases}$$

ja kertymäfunktio

$$F(x) \;=\; \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & \text{muuten.} \end{cases}$$

Määritetään vasemmalta jatkuva kvantiilifunktio suoraan Galois'n ehdosta (F.1). Koska jokaisella $0 < u < 1$ pätee

$$\begin{aligned} u &\leq 1 - e^{-\lambda x} && \iff \\ e^{-\lambda x} &\leq 1 - u && \iff \\ -\lambda x &\leq \log(1-u) && \iff \\ x &\geq -\frac{1}{\lambda}\log(1-u), \end{aligned}$$

todetaan että $u \leq F(x)$ jos ja vain jos $-\frac{1}{\lambda}\log(1-u) \leq x$. Näin ollen eksponenttijakauman vasemmalta jatkuva kvantiilifunktio on

$$Q(u) \;=\; -\frac{1}{\lambda}\log(1-u).$$

■

Esimerkki F.3 (Paretojakauma)**.** Paretojakauma tiheyseksponenttina $\beta > 1$ on jatkuva jakauma, jonka tiheysfunktio on

$$f(x) \;=\; \begin{cases} (\beta - 1)x^{-\beta}, & x \geq 1, \\ 0, & \text{muuten.} \end{cases}$$

ja kertymäfunktio

$$F(x) \;=\; \begin{cases} 1 - x^{-\beta+1}, & x \geq 1, \\ 0, & \text{muuten.} \end{cases}$$

Määritetään vasemmalta jatkuva kvantiilifunktio suoraan Galois'n ehdosta (F.1). Koska jokaisella $0 < u < 1$ pätee

$$\begin{aligned} u &\leq 1 - x^{-\beta+1} && \iff \\ x^{-\beta+1} &\leq 1 - u && \iff \\ x &\geq (1-u)^{-1/(\beta-1)} \end{aligned}$$

todetaan että $u \leq F(x)$ jos ja vain jos $(1-u)^{-1/(\beta-1)} \leq x$. Näin ollen Paretojakauman vasemmalta jatkuva kvantiilifunktio on

$$Q(u) \;=\; (1-u)^{-1/(\beta-1)}.$$

■

Esimerkki F.4 (Yleinen äärellisen arvojoukon jakauma)**.** Tarkastellaan diskreettiä jakaumaa, jonka mahdolliset arvot ovat $x_1 < x_2 < \cdots < x_n$ ja niiden todennäköisyydet $p_1, p_2, \ldots, p_n > 0$. Tätä jakaumaa vastaava diskreetti tiheysfunktio on

$$f(x) = \begin{cases} p_i, & x = x_i, \\ 0, & \text{muuten.} \end{cases}$$

Diskreeteille jakaumille ei yleensä määritetä kertymäfunktiota, sillä diskreettejä jakaumia sisältävät laskut on usein helpoin suorittaa suoraan tiheysfunktiota summamalla. Satunnaislukujen generointia varten kertymäfunktio voidaan kuitenkin määrittää muodossa

$$F(x) = \begin{cases} 0, & x < x_1, \\ p_1, & x_1 \leq x < x_2, \\ p_1 + p_2, & x_2 \leq x < x_3, \\ p_1 + p_2 + p_3, & x_3 \leq x < x_4, \\ \vdots & \vdots \\ p_1 + p_2 + \cdots + p_{n-1}, & x_{n-1} \leq x < x_n, \\ 1, & x \geq x_n. \end{cases}$$

Sitä vastaava vasemmalta jatkuva kvantiilifunktio on

$$Q(u) = \begin{cases} x_1, & 0 \leq u < p_1, \\ x_2, & p_1 \leq u < p_1 + p_2, \\ x_3, & p_1 + p_2 \leq u < p_1 + p_2 + p_3, \\ \vdots & \vdots \\ x_n, & p_1 + p_2 + \cdots + p_{n-1} \leq u < 1. \end{cases}$$

Tästä havaitaan, että satunnaismuuttuja $Q(U)$ saa arvon x_1 täsmälleen silloin kun $U \in [0, p_1)$, arvon x_2 täsmälleen silloin kun $U \in [p_1, p_1 + p_2)$ jne. Näiden välien pituudet ovat $p_1, p_2, \ldots$, niin kuin pitääkin. ■

F.2 Hylkäysotanta

Useille jakaumille kvantiilifunktion määrittäminen analyyttisesti tai laskennallisesti käyttökelpoisessa muodossa on hankalaa. Hylkäysotanta on yleinen algoritmi tiheysfunktion f mukaan jakautuneiden satunnaislukujen generoimiseen käyttämällä apuna vaihtoehtoisen tiheysfunktion g mukaan jakautuneita satunnaislukuja. Algoritmin käyttäminen edellyttää, että ehdotusjakauman g mukaisia satunnaislukuja voidaan tehokkaasti tuottaa, ja että on olemassa vakio M, jolle pätee

$$\frac{f(x)}{g(x)} \leq M \quad \text{aina kun } g(x) > 0.$$

Lause F.5. *Jos Y noudattaa jakaumaa g ja U on Y:stä riippumaton yksikkövälin jatkuvaa tasajakaumaa noudattava satunnaisluku, niin tällöin satunnaismuuttujan Y ehdollinen jakauma tapahtuman $U \leq \frac{f(Y)}{Mg(Y)}$ sattuessa on f.*

Todistus. Oletetaan, että $\frac{f(x)}{g(x)} \leq M$ kaikilla x. Tällöin

$$\mathbb{P}\left(Y = x \,|\, U \leq \frac{f(Y)}{Mg(Y)}\right) \;=\; \mathbb{P}(X = x).$$

sillä

$$\begin{aligned}
\mathbb{P}\left(Y \in A,\, U \leq \frac{f(Y)}{Mg(Y)}\right) &= \int\int 1(y \in A) 1\left(u \leq \frac{f(y)}{Mg(y)}\right) f_U(u) du g(y) dy \\
&= \int_A \left(\int_0^1 1\left(u \leq \frac{f(y)}{Mg(y)}\right) du\right) g(y) dy \\
&= \int_A \frac{f(y)}{Mg(y)} g(y) dy \\
&= \frac{1}{M}\int_A f(y) dy \\
&= \frac{1}{M}\mathbb{P}(X \in A).
\end{aligned}$$

Valitsemalla $A = \mathbb{R}$, tästä seuraa

$$\mathbb{P}\left(U \leq \frac{f(Y)}{Mg(Y)}\right) \;=\; \frac{1}{M}.$$

Väite seuraa. □

Esimerkki F.6 (Zipf-jakauma)**.** Zipf-jakauma tiheyseksponenttina $\beta > 1$ äärettömän lukujoukon $\{1, 2, \dots\}$ diskreetti jakauma, jonka tiheysfunktio on

$$f(x) \;=\; \zeta(\beta)^{-1} x^{-\beta}, \quad x = 1, 2, 3, \dots,$$

missä normitusvakio $\zeta(\beta) = \sum_{x=1}^{\infty} x^{-\beta}$ on Riemannin zeta-funktio pisteessä β. Zipf-jakaumasta on vaikea tuottaa satunnaislukuja kvantiilifunktion käytännössä. Sitä vastoin jatkuvasta paretojakauma tiheysfunktiona

$$f_{\mathrm{Par}}(t) \;=\; (\beta - 1) t^{-\beta}, \quad t \geq 1,$$

on helppo tuottaa satunnaislukuja hyödyntämällä kvantiilifunktiota $Q_{\mathrm{Par}}(u) = (1-u)^{-1/(\beta-1)}$ (esimerkki F.3). Kun syötteenä on yksikkövälin tasajakautunut satunnaisluku U, niin vasteena saadaan Pareto-jakautunut $Q_{\mathrm{Par}}(U)$, joka voidaan muuntaa kokonaislukuarvoisiksi pyöristämällä alaspäin. Tällöin saadaan diskreetti satunnaisluku

$$Y \;=\; \lfloor Q_{\mathrm{Par}}(U) \rfloor,$$

jonka arvojoukko on sama kuin Zipf-jakaumalla. Tämän diskreetin satunnaisluvun tiheysfunktio on

$$\begin{aligned} g(x) &= \mathbb{P}(\lfloor Q_{\mathrm{Par}}(U) \rfloor = x) \\ &= \mathbb{P}(x \leq Q_{\mathrm{Par}}(U) < x+1) \\ &= \int_x^{x+1} (\beta - 1) t^{-\beta}\, dt \\ &= x^{-\beta+1} - (x+1)^{-\beta+1}. \end{aligned}$$

Ylläoleva laskelma osoittaa, että $f(x) \neq g(x)$, joten satunnaismuuttujan $\lfloor Q_{\mathrm{Par}}(U) \rfloor$ jakauma ei ole Zipf-jakauma. Jakauma g on kuitenkin muodoltaan hyvin lähellä kohdejakaumaa f, joten g:tä voidaan käyttää ehdotusjakaumana Zipf-jakautuneiden satunnaismuuttujien generoimisessa. Derivoimalla voidaan tarkastaa, että

$$M := \max_{x \geq 1} \frac{f(x)}{g(x)} = \frac{f(1)}{g(1)} = \frac{\zeta(\beta)^{-1}}{1 - 2^{-\beta+1}}$$

Tällöin

$$\frac{f(x)}{M g(x)} = \frac{1 - 2^{-\beta+1}}{x^{\beta} g(x)} = \frac{1 - 2^{-\beta+1}}{x - (x+1)(1 + 1/x)^{-\beta}}.$$

Tätä menetelmää käytetään mm. Pythonin NumPy-ohjelmistokirjastossa.

```
rzipf <- function(beta) {
  repeat {
   X <- floor((1-runif(1))^(-1/(beta-1)))
   thres <- (1-2^(-beta+1))/(X-(X+1)*(1+1/X)^(-beta))
   U <- runif(1)
   if ( U <= thres )
     break;
  }
return(X)
}
```

Kirjallisuutta

[CM12] Herman Chernoff and Lincoln E. Moses. *Elementary Decision Theory.* Dover, 2012.

[HMC18] Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to Mathematical Statistics.* Pearson, eighth edition, 2018.

[JP04] Jean Jacod and Philip Protter. *Probability Essentials.* Springer, second edition, 2004.

[Kal02] Olav Kallenberg. *Foundations of Modern Probability.* Springer, second edition, 2002.

[Ros14] Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists.* Academic Press, 2014.

[vdV00] Aad W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 2000.

[Was10] Larry Wasserman. *All of Statistics.* Springer, 2010.

[Wil91] David Williams. *Probability with Martingales.* Cambridge University Press, 1991.