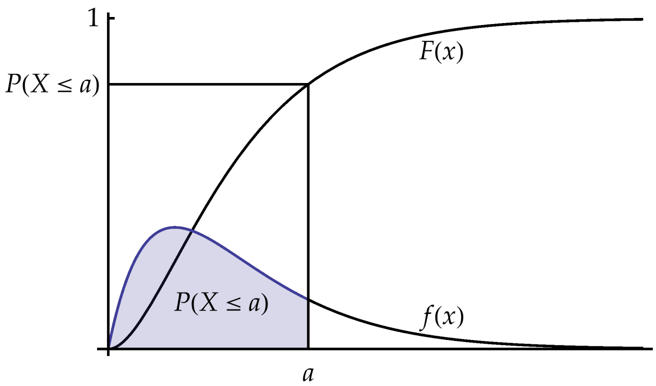
Todennäköisyyslaskenta sivuaineopiskelijoille

Heikki Ruskeepää



2012

Sisällys

1 Todennäköisyys

1.1 Klassinen todennäköisyys

1.1.1 Klassinen todennäköisyys

Määritelmä 1.1 Kokeen erilaisia tuloksia sanotaan *alkeistapauksiksi*. Kaikkien alkeistapausten joukko on *otosavaruus* Ω. Mikä hyvänsä alkeistapausten joukko on *tapaus*. Tapaus *sattuu*, jos kokeen tuloksena on alkeistapaus, joka kuuluu tapaukseen. ■

Esimerkki 1.1 Noppaa heitetään kerran.

Tapauksen A todennäköisyyttä merkitään $P(A)$. Olkoon $|A|$ tapauksen A alkeistapausten lukumäärä ja $|\Omega|$ otosavaruuden alkeistapausten lukumäärä.

Määritelmä 1.2 Jos kokeen kaikki alkeistapaukset ovat yhtä todennäköiset, niin tapauksen A *todennäköisyys* on $P(A) = \frac{|A|}{|\Omega|}$. ■

Tämä on *todennäköisyyden klassinen määritelmä*. Tapauksen A alkeistapauksia sanotaan myös *suotuisiksi* alkeistapauksiksi. Määritelmä voidaan siis kirjoittaa myös seuraavasti:

$$P(A) = \frac{\text{suotuisien alkeistapausten lukumäärä}}{\text{kaikkien alkeistapausten lukumäärä}}.$$

Esimerkki 1.2 *a)* Noppaa heitetään kerran.

b) Noppaa heitetään kaksi kertaa. Millä todennäköisyydellä silmälukujen summa on ainakin 9?

Esimerkki 1.3 On päätetty tutustua korkeintaan kolmeen puolisoehdokkaaseen. Sovelletaan seuraavaa menettelyä. Tutustutaan 1. ehdokkaaseen mutta hylätään se. Valitaan toinen ehdokas, jos se on parempi kuin 1. ehdokas; muutoin valitaan 3. ehdokas. Ehdokkaisiin tutustutaan satunnaisessa järjestyksessä. Millä todennäköisyydellä valituksi tule paras ehdokas; entä toiseksi paras ehdokas tai huonoin ehdokas?

Käytetään seuraavia merkintöjä ja nimityksiä:

- *Tyhjää joukkoa* merkitään $\emptyset$:llä.
- Tapaukset A ja B ovat *toisensa poissulkevat*, jos $A \cap B = \emptyset$.
- Tapaukset A_i, $i = 1, \ldots, n$, ovat *toisensa poissulkevat*, jos $A_i \cap A_j = \emptyset$ kaikilla $i \neq j$.
- Tapausjoukko A_i, $i = 1, \ldots, n$, on otosavaruuden Ω *partitio*, jos tapaukset A_i ovat toisensa poissulkevat ja $\bigcup_{i=1}^{n} A_i = \Omega$.
- Tapauksen A *komplementtitapaus* A^c sisältää ne alkeistapaukset, jotka eivät kuulu tapaukseen A.

Huomautus 1.1 Klassiselle todennäköisyydelle pätee

(a) $P(\emptyset) = 0, \ \ P(\Omega) = 1$,

(b) $0 \le P(A) \le 1$,

(c) $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$, jos tapaukset A_i ovat toisensa poissulkevat (*summakaava*),

(d) $\sum_{i=1}^{n} P(A_i) = 1$, jos tapaukset A_i muodostavat otosavaruuden partition (*summatesti*),

(e) $P(A) = 1 - P(A^c)$ (*komplementtikaava*). ■

Kohdat b ja d ovat erittäin tärkeitä tarkistuskeinoja:

• Jos laskettu todennäköisyys ei ole välillä [0, 1], niin lasku on väärin.

• Jos tapaukset A_i muodostavat Ω:n partition ja $\sum_{i=1}^{n} P(A_i)$ ei ole 1, niin ainakin yksi todennäköisyys $P(A_i)$ on väärin.

Jos $\sum_{i=1}^{n} P(A_i) = 1$, niin lasketut todennäköisyydet ovat luultavasti oikein, koska on epätodennäköistä, että virheellisillä todennäköisyyksillä olisi tämä ominaisuus. Esimerkissä 1.3 tapaukset V_1, V_2 ja V_3 muodostavat otosavaruuden partition. Koska $\sum_{i=1}^{n} P(V_i) = \frac{3}{6} + \frac{2}{6} + \frac{1}{6} = 1$, niin lasketut todennäköisyydet ovat luultavasti oikein.

■ 1.1.2 Geometrinen todennäköisyys

Toisinaan koetilanne on sellainen, että piste valitaan satunnaisesti annetun janan joiltain osilta. Todennäköisyyden klassista määritelmää ei voida sellaisenaan soveltaa, koska janalla on pisteitä ylinumeroituva määrä. Todennäköisyyksiä voidaan laskea janojen pituuksien suhteina:

$$P(A) = \frac{\text{janan suotuisten osien yhteispituus}}{\text{koko janan pituus}}.$$

Esimerkki 1.4 Piste valitaan satunnaisesti janalta (1, 10).

Toisinaan taas piste valitaan satunnaisesti annetun alueen joiltain osilta. Tällöin todennäköisyyksiä voidaan laskea pinta-alojen suhteina:

$$P(A) = \frac{\text{alueen suotuisten osien yhteispinta–ala}}{\text{koko alueen pinta–ala}}.$$

Esimerkki 1.5 Oletetaan, että kun tikkaa heitetään tikkatauluun, niin tikka osuu aina tauluun ja taulun kaikki pisteet ovat yhtä todennäköiset. Tällöin osumispiste on satunnainen.

Esimerkki 1.6 Poika ja tyttö saapuvat kohtaamispaikalle toisistaan riippumatta satunnaisena hetkenä aikavälillä 21.00–22.00. Poika odottaa tyttöä korkeintaan 20 min, ja tyttö odottaa poikaa korkeintaan 5 min. Kumpikin odottaa korkeintaan klo 22.00 saakka. Millä todennäköisyydellä poika ja tyttö tapaavat toisensa?

1.2 Kombinatoriikkaa

■ 1.2.1 Otanta

Kun käytetään klassista todennäköisyyden kaavaa $P(A) = \frac{|A|}{|\Omega|}$, joudutaan laskemaan tapauksien A ja Ω alkeistapauksien lukumäärät. Toisinaan nämä lukumäärät on helppo saada yksinkertaisen päättelyn avulla tai luettelemalla eri mahdollisuudet. Monesti on kuitenkin kätevämpää käyttää kombinatoriikan tuloksia.

Joukosta voidaan ottaa alkioita eli suorittaa *otanta* kahdella tavalla. Otanta tehdään *palauttamatta*, jos joukosta otetaan pois yksi alkio kerrallaan. Otanta tehdään *palauttaen*, jos jokaisen alkion oton jälkeen alkio palautetaan joukkoon.

Joukko on *järjestetty*, jos alkioiden järjestyksen muuttuessa myös joukko muuttuu. Järjestettyä joukkoa merkitään kaarisuluilla. On siis esimerkiksi $(a, b, c) \neq (b, a, c)$. Joukko on *järjestämätön*, jos alkioiden järjestyksellä ei ole väliä. Järjestämätöntä joukkoa merkitään aaltosuluilla. On siis esimerkiksi $\{a, b, c\} = \{b, a, c\}$.

Joukon alkioiden *k-permutaatio* muodostetaan ottamalla joukosta k:n alkion otos (palauttamatta tai palauttaen) ja muodostamalla siitä järjestetty joukko. Joukon alkioiden *k-kombinaatio* muodostetaan ottamalla joukosta k:n alkion otos (palauttamatta tai palauttaen) ja muodostamalla siitä järjestämätön joukko.

■ 1.2.2 Tuloperiaate

Lause 1.1 (*Tuloperiaate*) *a)* Jos operaatio A_i voidaan tehdä n_i:llä eri tavalla, $i = 1, \ldots, k$, niin jono (operaatio A_1, …, operaatio A_k) voidaan tehdä $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ eri tavalla.

b) Jos $n_i = n$ kaikilla i, niin eri tapoja on n^k.

Esimerkki 1.7 *a)* Korttipakan kullakin kortilla on tietty maa ja tietty arvo. Mahdolliset maat ovat risti, ruutu, pata ja hertta, ja mahdolliset arvot ovat 2, 3, …, 9, 10, J, Q, K ja A.

b) Ruokalistassa on 3 erilaista keittoa, 5 alkuruokaa, 8 pääruokaa ja 4 jälkiruokaa.

■ 1.2.3 Järjestetty otanta

Lause 1.2 (*Järjestetty otanta palauttamatta*) *a)* Joukosta, jossa on n erilaista alkiota, voidaan ottaa k:n alkion järjestetty otos palauttamatta $n(n-1)(n-2)\cdots(n-k+1)$ eli $\frac{n!}{(n-k)!}$ eri tavalla. Tästä k-permutaatioiden lukumäärästä käytetään myös merkintöjä $P(n, k)$, ${}_nP_k$ ja $(n)_k$.

b) Jos joukossa on n erilaista alkiota, niin alkiot voidaan järjestää jonoon $n!$ eri tavalla.

Esimerkki 1.8 *a)* Lasketaan, kuinka monella eri tavalla loton 7 palloa voivat mennä tulosputkeen

b) Kuinka monella eri tavalla kymmenen ihmistä voi mennä jonoon?

c) Kortit asetetaan jonoon niin, että K-kortit ovat vierekkäin. Montako erilaista jonoa on olemassa?

Lause 1.3 (*Järjestetty otanta palauttaen*) Joukosta, jossa on n erilaista alkiota, voidaan ottaa k:n alkion järjestetty otos palauttaen n^k eri tavalla.

Esimerkki 1.9 *a)* Paljonko on erilaisia jokerin tuloksia?

b) Kun noppaa heitetään kaksi kertaa, niin erilaisia tuloksia on $6^2 = 36$ kappaletta:

(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).

Nämä ovat kaikki yhtä todennäköisiä. Samat tulosmahdollisuudet saadaan, jos kahta erilaista (esim. eri väristä) noppaa heitetään yhtä aikaa. ■

Esimerkki 1.10 (*Syntymäpäivätehtävä*) Oletetaan, että vuoden kaikki päivät ovat yhtä todennäköisiä syntymäpäiviä (näin ei tarkasti ottaen ole). Millä todennäköisyydellä n:stä ihmisestä ainakin kahdella on sama vuoden päivä syntymäpäivänä?

Esimerkki 1.11 Millä todennäköisyydellä n:stä ihmisestä ainakin yhdellä on sama vuoden päivä syntymäpäivänä kuin *sinulla*?

■ 1.2.4 Järjestämätön otanta

Lause 1.4 (*Järjestämätön otanta palauttamatta; binomikertoimen otantatulkinta*) Joukosta, jossa on n erilaista alkiota, voidaan ottaa k:n alkion järjestämätön otos palauttamatta $\binom{n}{k}$ eli $\frac{n!}{k!\,(n-k)!}$ eri tavalla. Tästä k-kombinaatioiden lukumäärästä käytetään myös merkintöjä $C(n,\, k)$ ja ${}_nC_k$.

Esimerkki 1.12

Lause 1.5 Jos uurnasta, jossa on N palloa, joista M on mustia ja $N - M$ valkoisia, otetaan n palloa palauttamatta, niin todennäköisyys, että saadaan k mustaa ja $n - k$ valkoista palloa, on

$$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad 0 \le k \le M, \ 0 \le n - k \le N - M.$$

Huomaa, että lauseessa 1.5 on oletettu, että mustat pallot ovat jollain tavalla erilaisia eli että ne voidaan identifioida. Samoin on oletettu, että valkoiset pallot on jollain tavalla identifioitu. Tulos kuitenkin pätee, vaikka palloja ei todellisuudessa olisi identifioitu; pallot voidaan tällöin ajatella tilapäisesti identifioidun otantaa varten. — Lause 1.5 on ns. hypergeometrisen jakauman mukainen, ks. pykälää 2.2.4.

Esimerkki 1.13 Laatikossa on 4 punaista palloa ja 6 sinistä palloa. Laatikosta otetaan 5 palloa palauttamatta.

1.3 Aksiomaattinen todennäköisyys

1.3.1 Joukko-oppia

Seuraavassa on joukko-opin perusmerkintöjä ja niiden todennäköisyystulkintoja.

• Koska joukkojen A ja B *leikkaus* $A \cap B$ sisältää ne alkeistapaukset, jotka ovat sekä A:ssa että B:ssä, niin tapaus $A \cap B$ sattuu, jos sekä A että B sattuvat. Leikkausta $A \cap B$ merkitään usein myös lyhyemmin $A\,B$.

• Koska joukkojen A ja B *unioni* $A \cup B$ sisältää ne alkeistapaukset, jotka ovat ainakin toisessa joukoista A ja B, niin tapaus $A \cup B$ sattuu, jos ainakin toinen tapauksista A ja B sattuu.

• Koska joukon A *komplementti* A^c sisältää ne alkeistapaukset, jotka eivät ole A:ssa, niin tapaus A^c sattuu, jos A ei satu.

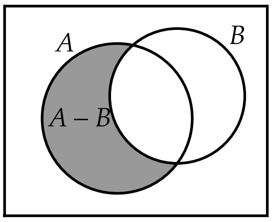
• Koska joukkojen A ja B *erotus* $A - B = A \cap B^c$ sisältää ne A:n alkeistapaukset, jotka eivät ole B:ssä, niin tapaus $A - B$ sattuu, jos A sattuu mutta B ei satu.

• Tyhjää joukkoa merkitään $\emptyset$; sanotaan, että $\emptyset$ on *mahdoton* tapaus (se ei koskaan satu).

• Koko otosavaruus Ω on *varma* tapaus (se sattuu aina).

• Oletetaan, että $A \subset B$. Jos tällöin A sattuu, niin myös B sattuu.

Tapauksia in tapana havainnollistaa ns. *Venn-diagrammeina*:



Selvästi

$$(A^c)^c = A, \quad A \cup A^c = \Omega, \quad A \cap A^c = \emptyset, \quad \Omega^c = \emptyset, \quad \emptyset^c = \Omega,$$

$$A \cup \emptyset = A, \quad \emptyset \cup \emptyset = \emptyset, \quad A \cap \emptyset = \emptyset, \quad \emptyset \cap \emptyset = \emptyset.$$

Unionille ja leikkaukselle pätevät seuraavat *distributiivilait*:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \qquad A \cup \left(\bigcap_{i=1}^{n} B_i \right) = \bigcap_{i=1}^{n} (A \cup B_i),$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \qquad A \cap \left(\bigcup_{i=1}^{n} B_i \right) = \bigcup_{i=1}^{n} (A \cap B_i).$$

Komplementille pätevät *de Morganin kaavat*:

$$(A \cup B)^c = A^c \cap B^c, \qquad \left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} A_i^c,$$

$$(A \cap B)^c = A^c \cup B^c, \qquad \left(\bigcap_{i=1}^{n} A_i \right)^c = \bigcup_{i=1}^{n} A_i^c.$$

Seuraava määritelmä on todennäköisyyslaskennassa erittäin tärkeä.

Määritelmä 1.3 Tapaukset A ja B ovat *toisensa poissulkevat*, jos $A \cap B = \emptyset$. ■

Määritelmän nimitys johtuu siitä, että jos A tapahtuu, niin B ei voi tapahtua, ja jos B tapahtuu, niin A ei voi tapahtua. Yleisemmin sanotaan, että tapaukset A_i ovat toisensa poissulkevat, jos $A_i \cap A_j = \emptyset$ kaikilla $i \neq j$.

■ 1.3.2 Todennäköisyyden aksioomat

Todennäköisyyden laskusääntöjen johtamiseksi esitetään ensin joukko todennäköisyydeltä vaadittavia ominaisuuksia. Nämä ominaisuudet, joita sanotaan myös *aksioomiksi*, ovat luontevia ja sopusoinnussa todennäköisyyden intuitiivisen käsityksen kanssa. Koko todennäköisyyslaskenta voidaan sitten johtaa näistä aksioomista. Aksiomaattisen todennäköisyyslaskennan on kehittänyt A. N. Kolmogorov kirjassaan *Grundbegriffe der Wahrscheinlichkeitsrechnung* (1933).

Todennäköisyyden aksioomat

A1. $P(A) \geq 0$ kaikille tapauksille A.

A2. $P(\Omega) = 1$.

A3. Jos tapaukset A_i, $i = 1, 2, \ldots$, ovat toisensa poissulkevat, niin $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. ■

Aksiooman A1 tapaan on luontevaa olettaa, että todennäköisyys on ei-negatiivinen. Aksiooma A2 normeeraa todennäköisyydet niin, että suurin mahdollinen todennäköisyys on 1. Aksioomassa A3 esiintyvä tapaus $\bigcup_{i=1}^{\infty} A_i$ sattuu, jos ainakin yksi tapauksista A_i sattuu. Koska kuitenkin tapaukset A_i ovat toisensa poissulkevat, niin tapaus $\bigcup_{i=1}^{\infty} A_i$ sattuu, jos tarkalleen yksi tapauksista A_i sattuu. On luontevaa, että tällöin tapauksien A_i todennäköisyydet lasketaan yhteen: $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$. Ominaisuutta A3 sanotaan σ-additiivisuudeksi.

Aksioomien avulla voidaan todistaa seuraava lause.

Lause 1.6 Todennäköisyydellä on seuraavat ominaisuudet:

a) $P(\emptyset) = 0$.

b) Jos $A \cap B = \emptyset$, niin $P(A \cup B) = P(A) + P(B)$.

c) $P(A) = 1 - P(A^c)$ (*komplementtikaava*).

d) Jos $A \subset B$, niin $P(A) \leq P(B)$.

e) $0 \leq P(A) \leq 1$.

f) $P(A - B) = P(A) - P(A \cap B)$.

■ 1.3.3 Toisensa poissulkevien tapausten unioni

Seuraavan lauseen a-kohta yleistää lauseen 1.6 b-kohdan useamman tapauksen unionille.

Lause 1.7 Oletetaan, että tapaukset A_i, $i = 1, \ldots, n$, ovat toisensa poissulkevat.

a) $P\left(\bigcup\limits_{i=1}^{n} A_i\right) = \sum\limits_{i=1}^{n} P(A_i)$ (*summakaava*).

b) Jos lisäksi $\bigcup\limits_{i=1}^{n} A_i = \Omega$, niin $\sum\limits_{i=1}^{n} P(A_i) = 1$.

Huomautus 1.1 Lauseen 1.6 e-kohta $0 \le P(A) \le 1$ ja lauseen 1.7 b-kohta ovat erittäin tärkeitä todennäköisyyksien tarkistamisessa. Laskuissa on nimittäin jokin virhe,

- jos laskujen tuloksena on todennäköisyys, joka on negatiivinen tai suurempi kuin 1;
- jos tapausavaruus voidaan jakaa toisensa poissulkevien tapausten unioniksi ja tapausten todennäköisyyksien summa on eri suuri kuin 1.

Vaikka tarvittaisiin vain tietyn tapauksen todennäköisyys, niin usein on hyödyllistä laskea kaikkien tapauksien todennäköisyydet, jotta voitaisiin käyttää lauseen 1.7 b-kohdan antamaa tarkistuskeinoa.

Lauseen 1.7 b-kohtaa voidaan käyttää myös toisella tavalla: yhtälöstä $\sum_{i=1}^{n} P(A_i) = 1$ voidaan yksi todennäköisyys $P(A_i)$ ratkaista muiden todennäköisyyksien avulla. Tästä on kuitenkin se huomattava varjopuoli, että kaavan $\sum_{i=1}^{n} P(A_i) = 1$ käyttö tarkistuskeinona menetetään. ■

Esimerkki 1.14

■ 1.3.4 Yleinen unioni

Lause 1.8 (*Mukaanlukemis–poissulkemis-periaate*, principle of inclusion-exclusion)

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

b) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.

c) $P(A \cup B \cup C \cup D) = P(A) + P(B) + P(C) + P(D) -$

$$- P(A \cap B) - P(A \cap C) - P(A \cap D) - P(B \cap C) - P(B \cap D) - P(C \cap D) +$$

$$+ P(A \cap B \cap C) + P(A \cap B \cap D) + P(A \cap C \cap D) + P(B \cap C \cap D) -$$

$$- P(A \cap B \cap C \cap D).$$

Lauseen nimi tulee siitä, että siinä unionin todennäköisyyden lausekkeeseen otetaan leikkausten todennäköisyyksiä mukaan plusmerkkisinä ja niitä suljetaan pois miinusmerkkisinä.

Esimerkki 1.15

■ 1.3.5 Todennäköisyyksien määrääminen

Todennäköisyyden aksioomat ja niistä johdetut laskusäännöt antavat vain kahden tapauksen todennäköisyydelle tietyn numeroarvon — $P(\emptyset) = 0$ ja $P(\Omega) = 1$ —, mutta muuten ne vain kertovat, miten jonkin tapauksen todennäköisyys voidaan laskea joidenkin muiden tapausten todennäköisyyksien avulla; esim. $P(A) = 1 - P(A^c)$, $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$.

Laskusääntöjen avulla voidaan siis todennäköisyyksiä muokata toisiin muotoihin, mutta lopulta tullaan vaiheeseen, jossa pitää antaa joidenkin tapausten todennäköisyyksille numeroarvot, jos tarkasteltavan tapauksen todennäköisyydelle halutaan numeroarvo.

Todennäköisyyksille voidaan määrätä numeroarvoja kolmella tavalla:

- todennäköisyyksien yhtäsuuruuden avulla,
- suhteellisten frekvenssien avulla,
- subjektiivisen arvioinnin avulla.

Todennäköisyyksien yhtäsuuruus

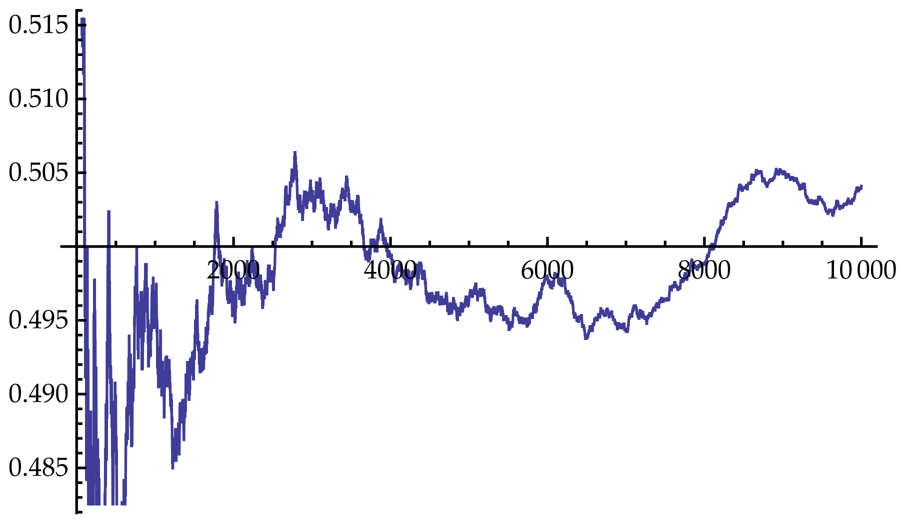
Jo kokeen kunkin tuloksen todennäköisyys on yhtä suuri, niin klassisen todennäköisyyden mukaan tapauksen todennäköisyys on suotuisien alkeistapauksien lukumäärän suhde kaikkien alkeistapauksien lukumäärään.

Suhteelliset frekvenssit

Jos tulosmahdollisuudet eivät ole yhtä todennäköiset, voidaan kokeiden tai tilastojen avulla laskea tapauksille suhteellisia frekvenssejä; niitä voidaan pitää todennäköisyyksien likiarvoina. Kun on esimerkiksi laskettu tietyllä aikavälillä syntyneiden poikien ja kaikkien syntyneiden lasten lukumäärien suhde, on pojan syntymän todennäköisyydelle saatu likiarvo 0.51.

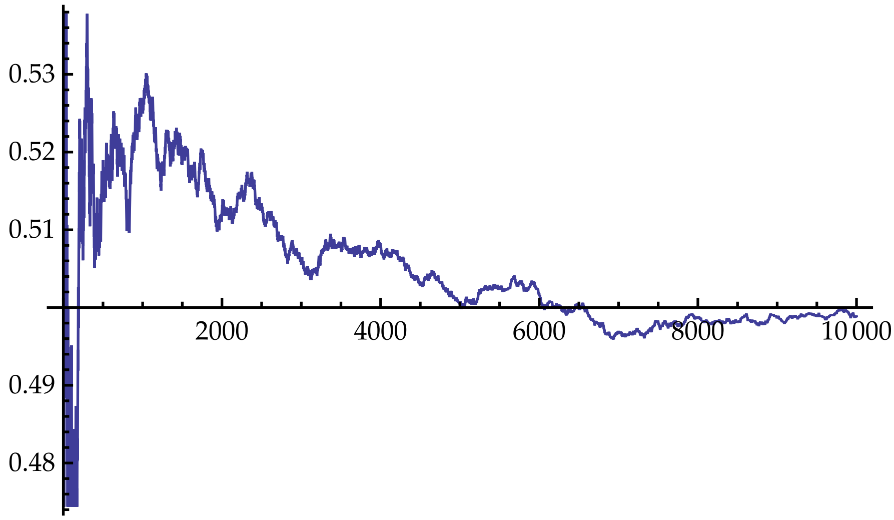
Jotta suhteellisten frekvenssien avulla saataisiin luotettava todennäköisyyden arvio, niin kokeita tai tilastoarvoja tarvitaan paljon. Seuraavassa on *Mathematica*-ohjelman avulla simuloitu rahanheittoa 10 000 kertaa ja piirretty, miten kruunien suhteellinen frekvenssi kehittyy, kun heittojen lukumäärä kasvaa:

```
ListLinePlot[Accumulate[RandomChoice[{0, 1}, 10000]] / Range[10000],
 AxesOrigin → {0, 0.5}]
```



Vielä 10 000 heiton jälkeenkin kruunan todennäköisyyden arvio on niinkin huono kuin 0.505. Eri simulontikerroilla tulos voi vaihdella paljonkin; seuraavassa simulaatiossa satutaan saamaan parempi todennäköisyyden arvio:

```
ListLinePlot[Accumulate[RandomChoice[{0, 1}, 10000]] / Range[10000],
 AxesOrigin → {0, 0.5}]
```



Mitä enemmän on käytettävissä tilastoarvoja, sen luotettavampi on suhteellinen frekvenssi todennäköisyyden arviona. Tämä voidaan myös todistaa todennäköisyyslaskennan tulosten avulla. Kurssilla Todennäköisyyslaskenta II esitetään Bernoullin lause, jonka mukaan tapauksen suhteellinen frekvenssi suppenee todennäköisyysmielessä kohti tapauksen todennäköisyyttä, kun kokeita tehdään yhä enemmän ja enemmän.

Subjektiivinen arviointi

Tiettyjen tapausten todennäköisyyksiä voidaan arvioida myös subjektiivisesti. Esimerkiksi lääkäri voi sanoa, että tietty potilas selviytyy sairaudesta todennäköisyydellä 0.8. Tämä arvio perustuu osittain lääkärin lukemiin tutkimuksiin aiheesta, osittain lääkärin omiin kokemuksiin aikaisemmista vastaavantapaisista tapauksista ja osittain lääkärin subjektiiviseen arvioon kyseisestä potilaasta. Tällainen subjektiivinen todennäköisyys ilmaisee henkilön uskomuksen siitä, että tapaus sattuu. Joku voisi sanoa esimerkiksi, että lähden lauantaina lenkille todennäköisyydellä 0.9.

Subjektiivisessakin todennäköisyydessä on usein osittain mukana frekvenssiajattelua. Kokemuksen mukaan esimerkiksi tietynlainen potilas on useimmiten selviytynyt, joten on hyvin todennäköistä, että näin tapahtuu tämänkin potilaan kohdalla, tai henkilö on lauantaisin useimmiten käynyt lenkillä, joten on hyvin todennäköistä, että näin tapahtuu seuraavanakin lauantaina.

1.4 Ehdollinen todennäköisyys

■ 1.4.1 Ehdollinen todennäköisyys

Esimerkki 1.16

Millä todennäköisyydellä tapaus A sattuu, kun tiedetään, että tapaus B on sattunut? Tämä todennäköisyys on ns. *ehdollinen todennäköisyys*, ja sitä merkitään $P(A \mid B)$. Tämän merkinnän voi lukea esim. seuraavilla tavoilla: ”P A ehdolla B”, ”A:n ehdollinen todennäköisyys ehdolla B”, ”todennäköisyys, että A tapahtuu, kun B on tapahtunut”. Edellisen esimerkin mukaisesti voidaan kirjoittaa:

Määritelmä 1.4a Jos $P(B) > 0$, niin e*hdollinen todennäköisyys* $P(A \mid B)$ on tapauksen A todennäköisyys otosavaruudessa B. ■

Koska tässä siis alkuperäinen otosavaruus Ω korvataan suppeammalla otosavaruudella B, niin tätä ehdollisen todennäköisyyden laskentamenetelmää sanotaan *supistetun otosavaruuden menetelmäksi* (määritelmässä 1.4b tullaan esittämään kaava, jossa käytetään alkuperäisen otosavaruuden Ω todennäköisyyksiä).

Esimerkki 1.17 Heitetään kahta noppaa. Millä todennäköisyydellä silmälukujen summa on 6, kun tiedetään, että ensimmäisen nopan tulos on 3?

Määritelmä 1.4b Jos $P(B) > 0$, niin *ehdollinen todennäköisyys* $P(A \mid B)$ on $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. ■

Ehdollinen todennäköisyys $P(A \mid B)$ voidaan siis määritellä kahdella tavalla:

- A:n todennäköisyytenä *supistetussa otosavaruudessa* B;
- *alkuperäisen otosavaruuden* Ω todennäköisyyksien avulla: $\frac{P(A \cap B)}{P(B)}$.

Kumpikin määritelmä on tärkeä käytännön laskuissa. Edellinen määritelmä on lisäksi tärkeä auttaessaan ymmärtämään, mistä ehdollisessa todennäköisyydessä on kyse. Jälkimmäinen määritelmä taas on tärkeä myös siksi, että sen avulla voidaan johtaa ehdolliseen todennäköisyyteen liittyviä tuloksia (mm. kokonaistodennäköisyyskaava ja Bayesin kaava).

Esimerkki 1.18 Kun tiettyä ohjelmajoukkoa tutkittiin, niin havaittiin, että 20 prosentissa ohjelmia oli syntaksivirheitä ja 6 prosentissa sekä syntaksi- että I/O-virheitä. Jos tietyssä ohjelmassa on syntaksivirheitä, niin millä todennäköisyydellä siinä on myös I/O-virheitä?

Esimerkki 1.19 Oletetaan, että lapsi on poika todennäköisyydellä 1/2. Tiedetään, että perheen kahdesta lapsesta ainakin yksi on poika. Millä todennäköisyydellä toinenkin lapsi on poika? Pintapuolisesti ajattelemalla voisi päätellä, että toinenkin lapsi on poika todennäköisyydellä 1/2.

Ehdolliselle todennäköisyydelle pätevät kaikki tavalliselle todennäköisyydelle johdetut tulokset. Esimerkiksi

$$0 \le P(A \mid B) \le 1,$$

$$P(A \mid B) = 1 - P(A^c \mid B).$$

■ 1.4.2 Kokonaistodennäköisyyskaava

Ratkaisemalla kaavoista $P(A \mid B) = P(A \cap B) / P(B)$ ja $P(B \mid A) = P(A \cap B) / P(A)$ todennäköisyys $P(A \cap B)$, saadaan seuraava lause. Leikkauksen todennäköisyyttä tarkastellaan lähemmin pykälässä 1.5.

Lause 1.9 Tapausten leikkauksen todennäköisyys toteuttaa kaavat

$$P(A \cap B) = P(A \mid B)\, P(B), \text{ jos } P(B) > 0;$$
$$P(A \cap B) = P(A)\, P(B \mid A), \text{ jos } P(A) > 0.$$

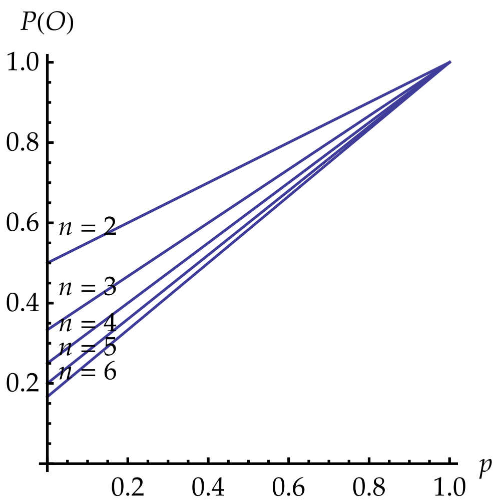
Määritelmä 1.5 Tapaukset $A_1, \ldots, A_n$ muodostavat otosavaruuden Ω *partition*, jos tapaukset ovat toisensa poissulkevat, $\Omega = \bigcup_{i=1}^{n} A_i$ ja $P(A_i) > 0$ kaikilla i. ■

Lause 1.10 (*Kokonaistodennäköisyyslause*) Jos tapaukset $A_1, \ldots, A_n$ muodostavat otosavaruuden partition, niin

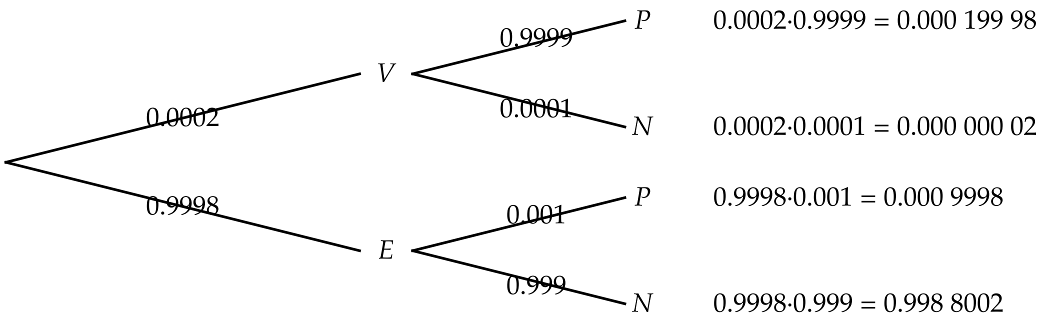
$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)\, P(A_i).$$

Lauseen 1.10 kokonaistodennäköisyyskaavan avulla "kokonais"-todennäköisyys $P(B)$ voidaan koota ehdollisista todennäköisyyksistä $P(B \mid A_i)$ painottamalla kutakin ehdollista todennäköisyyttä ehdon todennäköisyydellä $P(A_i)$. Kokonaistodennäköisyyskaava on hyödyllinen monissa tapauksissa, joissa todennäköisyyttä $P(B)$ on vaikea laskea suoraan, mutta laskenta on helppoa ehdollisten todennäköisyyksien $P(B \mid A_i)$ avulla.

Esimerkki 1.20 Monivalintakokeen kussakin tehtävässä on n vastausvaihtoehtoa, joista yksi on oikea. Opiskelija tietää kuhunkin kysymykseen oikean vastauksen todennäköisyydellä p. Jos opiskelija ei tiedä vastausta, hän arvaa. Millä todennäköisyydellä vastaus yhteen kysymykseen on oikea?



Esimerkki 1.21 Tietty virus on kahdella ihmisellä 10 000:sta. Jos ihmisellä on tämä virus, niin tietty testikin ilmoittaa 99.99 prosentin varmuudella, että ihmisellä on tämä virus (testin tulos on positiivinen). Jos ihmisellä ei ole tätä virusta, niin testikin ilmoittaa 99.9 prosentin varmuudella, että ihmisellä ei ole tätä virusta (testin tulos on negatiivinen). Millä todennäköisyydellä testin tulos on positiivinen; entä negatiivinen?



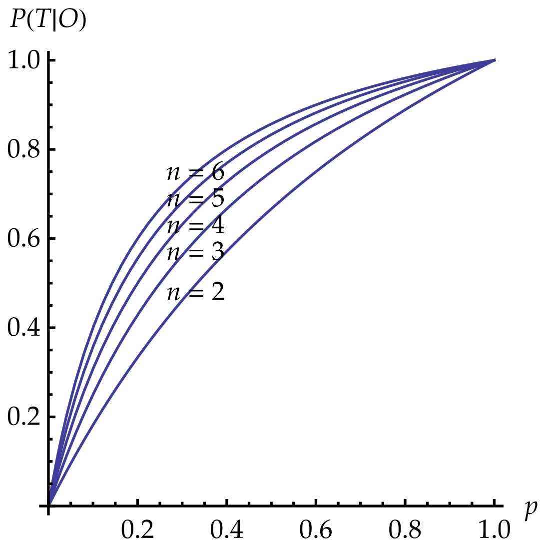
■ 1.4.3 Bayesin kaava

Bayesin kaava on usein hyödyllinen, kun halutaan laskea ehdollisia todennäköisyyksiä.

Lause 1.11 (*Bayesin lause*; Thomas Bayes, 1701–1761) Jos tapaukset $A_1, \ldots, A_n$ muodostavat otosavaruuden partition ja $P(B) > 0$, niin

$$P\,(A_k \mid B) = \frac{P(B|A_k)\,P(A_k)}{P(B)} = \frac{P(B|A_k)\,P(A_k)}{\sum_{i=1}^{n} P(B|A_i)\,P(A_i)}, \quad k = 1, \ldots, n.$$

Esimerkki 1.22 Jatketaan esimerkkiä 1.20. Monivalintakokeen kussakin tehtävässä on n vastausvaihtoehtoa, joista yksi on oikea. Opiskelija tietää kuhunkin kysymykseen oikean vastauksen todennäköisyydellä p. Jos opiskelija ei tiedä vastausta, hän arvaa. Jos vastaus yhteen kysymykseen on oikea, niin millä todennäköisyydellä opiskelija tiesi vastauksen eikä arvannut?



Esimerkki 1.23 Jatketaan esimerkkiä 1.21. Tietty virus on kahdella ihmisellä 10 000:sta. Jos ihmisellä on tämä virus, niin tietty testikin ilmoittaa 99.99 prosentin varmuudella, että ihmisellä on tämä virus (testin tulos on positiivinen). Jos ihmisellä ei ole tätä virusta, niin testikin ilmoittaa 99.9 prosentin varmuudella, että ihmisellä ei ole tätä virusta (testin tulos on negatiivinen). Jos testin tulos on positiivinen, niin millä todennäköisyydellä ihmisellä todella on virus?

Mistä testin huono luotettavuus johtuu? Jos tarkastellaan esimerkiksi 10 000 ihmisen joukkoa, niin on odotettavissa, että siinä kahdella ihmisellä on virus. Testi on näiden kahden ihmisen joukossa positiivinen keskimäärin $0.9999 \cdot 2 = 1.9998$:lle ihmiselle. Testi on positiivinen 9998 terveen ihmisen joukossa keskimäärin $0.001 \cdot 9998 = 9.998$:lle ihmiselle. Keskimäärin saadaan siis yhteensä 11.9978 positiivista testitulosta. Tästä sairaiden osuus 1.9998 on todellakin vain 16.7 prosenttia. Huono tulos johtuu siis virheellisten positiivisten tulosten suuresta määrästä. Se taas johtuu siitä, että vaikka virheellisen positiivisen tuloksen todennäköisyys onkin pieni (0.001), niin virheellisiä positiivisia tuloksia kuitenkin syntyy melko paljon, koska terveiden ihmisten osuus oli hyvin suuri (0.9998).

Huomautus 1.2 Bayesin kaava voidaan tulkita kahdellakin tavalla.

a) Käänteiset ehdolliset todennäköisyydet. Ehdollisille todennäköisyyksille $P(B \mid A_i)$, $i = 1, \ldots, n$, voidaan Bayesin kaavan avulla laskea *"käänteiset" ehdolliset todennäköisyydet* $P(A_k \mid B)$, $k = 1, \ldots, n$. Jos A_i on tapauksen B mahdollinen syy, niin $P(B \mid A_i)$ on seurauksen todennäköisyys, kun syy tiedetään, mutta $P(A_i \mid B)$ on *syyn todennäköisyys, kun seuraus tiedetään*.

b) Posterioriset todennäköisyydet. Alun perin tiedetään todennäköisyydet $P(A_i)$, $i = 1, \ldots, n$; nämä ovat ns. *prioriset todennäköisyydet*. Sitten saadaan se informaatio, että tapaus B on sattunut. Bayesin kaavan avulla voidaan nyt laskea ns. *posterioriset todennäköisyydet* $P(A_k \mid B)$, $k = 1, \ldots, n$. Tällä tavalla voidaan todennäköisyyksiä *päivittää*, kun saadaan uutta informaatiota.

Prioriset tn:t	Posterioriset tn:t	
$P(V) = 0.0002$	$P(V \mid P) = 0.167,$	$P(V \mid N) = 0.000\,000\,02$
$P(E) = 0.9998$	$P(E \mid P) = 0.833,$	$P(E \mid N) = 0.999\,999\,98$

1.4.4 Laskentaohjeita

Kun ratkaistaan todennäköisyyslaskennan tehtäviä, niin vaarana on, että lasku etenee niin kuin laskijasta vain *tuntuu* järkevältä ja laskun eri vaiheet jäävät perustelematta tai vain hämärän intuition varaan. Tästä on kaksi varjopuolta:

- *Lasku menee usein väärin*, koska perusteluja ei mietitä.
- Vaikka laskun lopputulos olisi oikeakin, niin *ulkopuolisen on vaikeaa saada laskusta selvää*: todennäköisyydet jäävät tarkemmin määrittelemättä ja laskennan vaiheet perustelematta.

Jotta laskusta saataisiin varmemmin oikea tulos ja lasku olisi myös muiden ymmärrettävissä, niin kannattaa noudattaa seuraavia ohjeita. Monet luentojen esimerkit noudattavat näitä ohjeita. ***Ohjeita on noudatettava myös tenttitehtävissä.***

Ohjeita todennäköisyyksien laskemiseen:
1) Anna tapauksille helposti muistettavat *symbolit*.
2) Kirjoita *annetut todennäköisyydet* symbolien avulla.
3) Kirjoita ja laske *kysytty todennäköisyys* symbolien avulla.
4) Käytä laskemisessa vain todennäköisyyslaskennan *tunnettuja tuloksia*.
5) *Sijoita* saatuun symboliseen lausekkeeseen annetut todennäköisyydet ja sievennä.
6) Usein on mielenkiintoista tietää todennäköisyydelle sekä *tarkka arvo* että *desimaaliarvo*.

1) *Symbolien määrittely* tapauksille on välttämätöntä, jotta lasku voitaisiin ensin ratkaista symbolisesti. Voidaan esimerkiksi määritellä V = ihmisellä on virus, E = ihmisellä ei ole virusta, P = testi on positiivinen.

2) *Annettujen todennäköisyyksien kirjoittaminen symbolien avulla* selkeyttää lähtötilanteen ja helpottaa myöhempää laskemista. Voidaan esimerkiksi kirjoittaa $P(V) = 0.0002$, $P(E) = 0.9998$, $P(P \mid V) = 0.9999$, $P(P \mid E) = 0.001$.

3) Kun *kysytty todennäköisyys kirjoitetaan symbolien avulla*, tiedetään tarkalleen, mitä pitää laskea. Kun *tämä todennäköisyys sitten lasketaan symbolien avulla*, tulee samalla mietityksi, millä perusteella ja millä kaavalla todennäköisyys lasketaan, ja lukijan on helppo seurata laskennan etenemistä. Voidan esimerkiksi kirjoittaa kokonaistodennäköisyyskaavan avulla

$$P(P) = P(P \mid V)\,P(V) + P(P \mid E)\,P(E).$$

4) Laskennassa on tärkeää, että *käytetään vain todennäköisyyslaskennan tunnettuja tuloksia*, koska se varmistaa, että ratkaisu on oikea. Vältä sellaisten ad hoc -päätelyjen käyttöä, jotka vain tuntuvat järkeviltä. Alla on lueteltu todennäköisyyksiä koskevat tärkeimmät kaavat.

5) Vasta kun todennäköisyyden symbolinen lauseke on saatu selville, siihen *sijoitetaan annetut lähtötiedot* (jotka kirjoitettiin kohdassa 2) ja lauseke sievennetään. Kun symbolinen lauseke on selvillä, niin sijoitusvaihe on usein yksinkertaista numeerista laskentaa. Voidaan esimerkiksi kirjoittaa

$$P(P) = P(P \mid V)\,P(V) + P(P \mid E)\,P(E) = 0.9999 \cdot 0.0002 + 0.001 \cdot 0.9998 = 0.001\,199\,78.$$

6) Todennäköisyyden *tarkka arvo* on usein mielenkiintoinen ja arvokas, varsinkin, jos se on verraten yksinkertainen murtoluku tai muu tarkka lauseke. *Desimaaliarvo* on usein myös havainnollinen, koska siitä näkyy helposti todennäköisyyden suuruusluokka. Näytä desimaaliluvuissa riittävän monta nollasta eroavaa desimaalia (esim. 4), jotta todennäköisyyksien oikeellisuus tulisi selväksi myös lukijalle ja jotta todennäköisyydet voisi tarkistaa kaavan $\sum_{i=1}^{n} P(A_i) = 1$ avulla riittävän tarkasti (todennäköisyyksien tarkistusta tarkastellaan hetken kuluttua).

Kohdassa 4 kehotetaan käyttämään vain todennäköisyyslaskennan tunnettuja tuloksia. Seuraavassa on lueteltu tällaisia tuloksia (näistä kaavat 6 ja 7 tulevat esille vasta pykälässä 1.5).

Todennäköisyyksiä koskevat tärkeimmät kaavat:

Peruskaavoja:

1) Klassinen todennäköisyys: $P(A) = \frac{|A|}{|\Omega|}$, jos alkeistapaukset ovat yhtä todennäköiset

2) Komplementtikaava: $P(A) = 1 - P(A^c)$

Unioniin liittyviä kaavoja:

3) Kahden tapauksen unioni: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4) Toisensa poissulkevien tapausten unioni: $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ (summakaava)

Leikkaukseen liittyviä kaavoja:

5) Kahden tapauksen leikkaus: $P(A \cap B) = P(A)\,P(B \mid A) = P(A \mid B)\,P(B)$

6) Yleinen leikkaus: $P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1 \cap A_2) \cdots$

7) Riippumattomien tapausten leikkaus: $P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i)$ (tulokaava)

Ehdolliseen todennäköisyyteen liittyviä kaavoja:

8) Ehdollinen todennäköisyys: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

9) Kokonaistodennäköisyyskaava: $P(B) = \sum_{i=1}^{n} P(B \mid A_i)\,P(A_i)$

10) Bayesin kaava: $P(A_k \mid B) = \frac{P(B|A_k)\,P(A_k)}{P(B)}$

Ole tarkka todennäköisyyksien yhteenlaskussa ja kertomisessa:

Todennäköisyyksien yhteenlasku ja kertominen:

- Jos lasket todennäköisyyksiä yhteen, niin tapausten on oltava *toisensa poissulkevat* (kaava 4).
- Jos kerrot todennäköisyyksiä, niin
  - joko tapausten on oltava *riippumattomat* (kaava 7)
  - tai tulon tekijöiden on oltava *ehdollisia todennäköisyyksiä* (kaavat 5, 6, 9 ja 10).

Aikaisemmin mainitut kuusi ohjetta auttavat pääsemään oikeaan lopputulokseen. Kun todennäköisyys on laskettu, on kuitenkin syytä vielä kiinnittää huomiota tuloksen oikeellisuuteen. Todennäköisyyksiä voi tarkistaa seuraavilla tavoilla:

Todennäköisyyksien tarkistaminen:

- $0 \le P(A) \le 1$.
- $\sum_{i=1}^{n} P(A_i) = 1$, jos $\{A_1, \dots, A_n\}$ on Ω:n partitio. Jos siis tapaukset muodostavat Ω:n partition ja olet laskenut kaikkien tapausten todennäköisyydet, niin todennäköisyyksien summan täytyy olla tasan 1.
- Vaikka kysytään vain yhden tai muutaman tapauksen todennäköisyyttä, niin laske *kaikkien* toisensa poissulkevien tapausten todennäköisyydet, jotta voit käyttää testiä $\sum_{i=1}^{n} P(A_i) = 1$.
- Laske yleisen todennäköisyyden arvo joissakin *erikoistapauksissa*, joissa todennäköisyyden pystyy varmasti laskemaan oikein.
- Mieti, tuntuuko tulos *järkevältä* (mutta muista kuitenkin, että todennäköisyyslaskennassa on yllättäviäkin tuloksia).
- Arvioi todennäköisyyttä *simuloimalla* tehtävän tilannetta tietokoneen avulla ja vertaa arviota todennäköisyyden laskettuun arvoon.
- Tutki kirjallisuutta ja kysy neuvoa.

1.5 Riippumattomuus

1.5.1 Leikkauksen todennäköisyys

Lause 1.12 Seuraavat kaavat pätevät, jos niissä esiintyvät ehdolliset todennäköisyydet ovat olemassa.

a) $P(A \cap B) = P(A)\ P(B \mid A)$,
$P(A \cap B) = P(A \mid B)\ P(B)$;

b) $P(A \cap B \cap C) = P(A)\ P(B \mid A)\ P(C \mid A \cap B)$;

c) $P(A \cap B \cap C \cap D) = P(A)\ P(B \mid A)\ P(C \mid A \cap B)\ P(D \mid A \cap B \cap C)$;

d) $P(\bigcap_{i=1}^{n} A_i) = P(A_1)\ P(A_2 \mid A_1)\ P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \ldots \cap A_{n-1})$

Esimerkki 1.24 Noppaa heitetään kerran. Olkoon E = "silmäluku on parillinen" ja V = "silmäluku on korkeintaan 5". Lasketaan tapauksen $E \cap V$ todennäköisyys kolmella tavalla

Esimerkki 1.25 Luokassa on 7 tyttöä ja 5 poikaa. Satunnaiset 3 oppilasta asettuvat jonoon. Millä todennäköisyydellä tytöt ja pojat vuorottelevat jonossa?

Esimerkki 1.26 *a)* Avainnipussa on n avainta. Avaimia kokeillaan peräjälkeen, kunnes oikea avain löytyy; kokeillut avaimet pidetään erillään kokeilemattomista. Millä todennäköisyydellä vasta k:s avain on oikea?

1.5.2 Kahden tapauksen riippumattomuus

Ehdollinen todennäköisyys $P(A \mid B)$ riippuu yleisesti B:stä, ts. $P(A \mid B)$ on eri kuin $P(A)$. Toisinaan on kuitenkin $P(A \mid B) = P(A)$. Tällöin on luontevaa sanoa, että tapaus A on *riippumaton* tapauksesta B, sillä tieto tapauksen B sattumisesta ei vaikuta mitenkään tapauksen A todennäköisyyteen. Jos on $P(A \mid B) = P(A)$, niin on myös $P(B \mid A) = P(B)$.
Näin ollen myöskin tapaus B on riippumaton tapauksesta A. Voidaan siis sanoa, että jos $P(A \mid B) = P(A)$, niin tapaukset A ja B ovat *riippumattomat*.

Kaava $P(A \mid B) = P(A)$ voitaisiinkin ottaa tapausten riippumattomuuden määritelmäksi, mutta yleisesti käytetty määritelmä on kuitenkin seuraava:

$$P(A \cap B) = P(A)\, P(B).$$

Kaavat $P(A \mid B) = P(A)$ ja $P(A \cap B) = P(A)\, P(B)$ ovat nimittäin ekvivalentit.

Määritelmä 1.6 Tapaukset A ja B ovat *riippumattomat*, jos $P(A \cap B) = P(A)\, P(B)$. Jos tapaukset eivät ole riippumattomat, ne ovat *riippuvat*. ■

Tapausten A ja B riippumattomuutta voidaan merkitä $A \perp B$.

Esimerkki 1.27 *a)* Noppaa heitetään kaksi kertaa. Olkoon E = “1. tulos on neljä” ja S = “silmälukujen summa on 6”.

b) Noppaa heitetään edelleen kaksi kertaa. Olkoon nyt E = “1. tulos on neljä” ja S = “silmälukujen summa on 7”.

Lause 1.13 Jos tapaukset A ja B ovat riippumattomat, niin samoin ovat A ja B^c, A^c ja B sekä A^c ja B^c.

Huomautus 1.3 Tapausten A ja B riippumattomuus eli $P(A \cap B) = P(A)\, P(B)$ ja toisensa poissulkevuus eli $A \cap B = \emptyset$ ovat aivan eri asioita. Kummastakaan ominaisuudesta ei seuraa toinen ominaisuus.

Jos esimerkiksi tapaukset ovat toisensa poissulkevat, niin silloin tapaukset ovat selvästi riippuvat: jos toinen sattuu, niin toinen ei voi sattua.

Jos taas tapaukset ovat riippumattomat, niin eivät ne välttämättä sulje toisiaan pois.

1.5.3 Useamman tapauksen riippumattomuus

Sanotaan, että tapaukset A, B ja C ovat *parittain riippumattomat*, jos

$$P(A \cap B) = P(A)\, P(B), \quad P(A \cap C) = P(A)\, P(C), \quad P(B \cap C) = P(B)\, P(C).$$

Kolmen tapauksen varsinaiseen riippumattomuuteen vaaditaan parittainen riippumattomuus ja myös kolmittainen riippumattomuus:

Määritelmä 1.7 Tapaukset A, B ja C ovat *riippumattomat*, jos seuraavat ehdot toteutuvat:

$$P(A \cap B) = P(A)\, P(B),$$

$$P(A \cap C) = P(A)\, P(C),$$

$$P(B \cap C) = P(B)\, P(C),$$

$$P(A \cap B \cap C) = P(A)\, P(B)\, P(C).$$ ■

Lause 1.14 Jos tapaukset A, B ja C ovat riippumattomat, niin A on riippumaton kaikista tapauksista, jotka on muodostettu tapauksista B ja C.

Määritelmä 1.8 Tapaukset A_i, $i = 1, \ldots, n$, ovat *riippumattomat*, jos seuraavat ehdot toteutuvat:

$$P(A_i \cap A_j) = P(A_i)\,P(A_j) \text{ kaikilla } i < j,$$

$$P(A_i \cap A_j \cap A_k) = P(A_i)\,P(A_j)\,P(A_k) \text{ kaikilla } i < j < k,$$

$$\vdots$$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\,P(A_2) \cdots P(A_n). \ \blacksquare$$

Riippumattomuuden määritelmiä 1.6, 1.7 ja 1.8 voidaan käyttää kahdella tavalla:

- Määritelmien avulla voidaan testata, ovatko annetut tapaukset riippumattomat.
- Jos annettujen tapausten riippumattomuus on selvää, niin määritelmien kaavoja voidaan käyttää leikkausten todennäköisyyksien laskemiseen.

Näistä jälkimmäinen käyttötapa on paljon yleisempi ja tärkeämpi. Seuraavaan lauseeseen on kirjoitettu leikkauksen todennäköisyyden kaava siinä erikoistapauksessa, että tapaukset ovat riippumattomat.

Lause 1.15 Jos tapaukset A_i, $i = 1, \ldots, n$, ovat riippumattomat, niin

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) \quad (\textit{tulokaava}).$$

Esimerkki 1.28 Tiettyä koetta toistetaan riippumatomasti. Kukin koe onnistuu todennäköisyydellä p; olkoon $q = 1 - p$. Merkitään O_i = ”i:s koe onnistuu” ja E_i = ”i:s koe epäonnistuu”.

Lause 1.16 Oletetaan, että koetta toistetaan riippumattomasti ja kukin koe onnistuu todennäköisyydellä p. Olkoon $q = 1 - p$.

a) Jos koe toistetaan n kertaa ja X onnistuneiden kokeiden lukumäärä, niin $P(X = k) = \binom{n}{k} p^k q^{n-k}$, $k = 0, 1, \ldots, n$.

b) Jos koe toistetaan, kunnes koe onnistuu ensimmäisen kerran, ja X = tarvittavien kokeiden lukumäärä, niin $P(X = k) = q^{k-1} p$, $k = 1, 2, \ldots$

c) Jos koe toistetaan, kunnes koe onnistuu n:nnen kerran, ja X = tarvittavien kokeiden lukumäärä, niin $P(X = k) = \binom{k-1}{n-1} p^n q^{k-n}$, $k = n, n+1, \ldots$

Esimerkki 1.29 *a)* Noppaa heitetään n kertaa.

b) Noppaa heitetään, kunnes saadaan kruuna.

c) Noppaa heitetään, kunnes saadaan n:s kruuna.

Esimerkki 1.30 Kolme metsästäjää ampuu samaa jänistä täsmälleen samanaikaisesti. Oletetaan, että osumiset ovat riippumattomat. Metsästäjien osumistodennäköisyydet ovat 0.01, 0.05 ja 0.08. Käytetään sellaista merkintätapaa, että esimerkiksi $1 \cap 2 \cap 3^c$ tarkoittaa tapausta, että 1. ja 2. metsästäjä osuvat mutta 3. ei.

2 Satunnaismuuttujat

2.1 Diskreetti satunnaismuuttuja

2.1.1 Jakauma- ja todennäköisyysfunktio

Esimerkki 2.1 *a)* Noppaa heitetään; otosavaruus on $\Omega = \{$ykkönen, …, kuutonen$\}$ ja alkeistapaukset ovat yhtä todennäköiset.

b) Kahta noppaa heitetään; otosavaruus on $\Omega = \{$(ykkönen, ykkönen), …, (kuutonen, kuutonen)$\}$ ja alkeistapaukset ovat yhtä todennäköiset.

c) Kahta rahaa heitetään; otosavaruus on $\Omega = \{(R, R), (R, L), (L, R)$ ja $(L, L)\}$ ja alkeistapaukset ovat yhtä todennäköiset.

Määritelmä 2.1 *Satunnaismuuttuja* X otosavaruudessa Ω on funktio $X\colon \Omega \to \mathbb{R}$. ■

Satunnaismuuttuja siis liittää kuhunkin alkeistapaukseen jonkin reaaliluvun. Mikä tämä reaaliluku on, riippuu siitä, mitä satunnaismuuttujan halutaan kuvaavan.

Määritelmä 2.2 Satunnaismuuttuja X on *diskreetti*, jos se voi saada vain äärellisen tai numeroituvasti äärettömän määrän arvoja. ■

Merkitään, että diskreetin satunnaismuuttujan saamien arvojen joukko on K. Diskreeteillä satunnaismuuttujilla muotoa $P(X = k)$ olevat todennäköisyydet ovat tärkeimmät. Näille todennäköisyyksille määritellään funktio.

Määritelmä 2.3 Diskreetin satunnaismuuttujan X *todennäköisyysfunktio* on $p(k) = P(X = k)$, $k \in K$. ■

Käytetään myös nimityksiä pistetodennäköisyysfunktio ja tiheysfunktio (englannissa käytetään usein termiä probability density function ja lyhennettä pdf).

Todennäköisyysfunkiolla $p(k)$ on seuraavat ominaisuudet:

- $0 \le p(k) \le 1$, $k \in K$,
- $\sum_{k\in K} p(k) = 1$.

Todennäköisyyslaskennassa halutaan usein laskea myös todennäköisyyksiä, jotka ovat muotoa $P(X \le a)$, $P(X > a)$ ja $P(a < X \le b)$. Nämä voidaan kaikki ilmaista muotoa $P(X \le a)$ olevien todennäköisyyksien avulla. Ensinnäkin $P(X > a) = 1 - P(X \le a)$. Toiseksi jos $a < b$, niin $P(X \le b) = P(x \le a) + P(a < X \le b)$, joten $P(a < X \le b) = P(X \le b) - P(X \le a)$. Koska siis muotoa $P(X \le a)$ olevat todennäköisyydet ovat keskeisiä, on tälle todennäköisyydelle määritelty oma funktio:

Määritelmä 2.4 Satunnaismuuttujan X *jakaumafunktio* on $F(x) = P(X \le x)$, $x \in \mathbb{R}$. ■

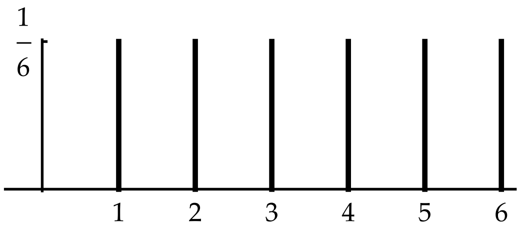
Käytetään myös nimityksiä kumulatiivinen jakaumafunktio ja kertymäfunktio (englannissa käytetään usein termiä cumulative distribution function ja lyhennettä cdf).

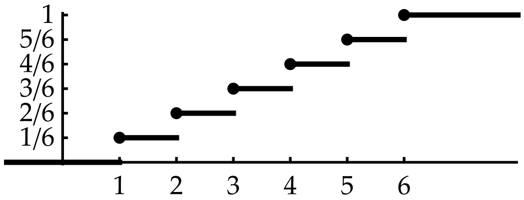
Huomaa, että todennäköisyysfunktio määritellään vain pisteissä $k \in K$ mutta jakaumafunktio kaikilla $x \in \mathbb{R}$.

Todennäköisyysfunktiosta saadaan jakaumafunktio ja jakaumafunktiosta todennäköisyysfunktio seuraavasti:

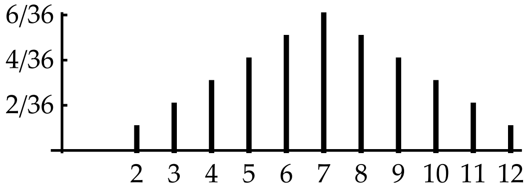
- $F(x) = \sum_{k \leq x} p(k)$,
- $p(k) = F(k) - F(k-1)$.

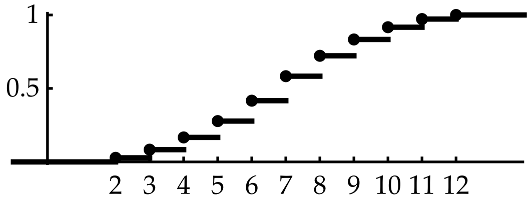
Esimerkki 2.2 *a)* Noppaa heitetään; olkoon X nopanheiton tulosta vastaava kokonaisluku.





b) Kahta noppaa heitetään; olkoon X silmälukujen summa.





Diskreetin satunnaismuuttujan jakaumafunktiolla $F(x)$ on seuraavat ominaisuudet:

- $0 \leq F(x) \leq 1$, $x \in \mathbb{R}$;
- $F(x)$ on oikealta jatkuva porrasfunktio;
- $F(x)$ on ei-vähenevä;
- $\lim_{x \to \infty} F(x) = 1$;
- $\lim_{x \to -\infty} F(x) = 0$.

■ 2.1.2 Odotusarvo

Esimerkki 2.3 Seuraavassa on *Mathematica*-ohjelman avulla simuloitu 100 nopanheittoa:

```
nn = RandomInteger[DiscreteUniformDistribution[6], 100]
```

```
{2, 2, 2, 5, 1, 3, 4, 1, 6, 3, 3, 6, 4, 5, 2, 4, 1, 5, 3, 5, 1, 4, 2, 2, 5,
 4, 3, 1, 6, 4, 4, 2, 2, 3, 4, 2, 3, 5, 4, 6, 3, 1, 3, 3, 6, 6, 3, 6, 1, 4,
 4, 3, 3, 2, 2, 6, 1, 5, 4, 4, 2, 2, 6, 4, 1, 3, 4, 1, 5, 1, 4, 1, 2, 4, 3,
 4, 5, 2, 2, 5, 1, 6, 1, 2, 3, 1, 4, 5, 2, 1, 4, 2, 5, 6, 1, 5, 3, 5, 1, 6}
```

Tulosten artimeettinen keskiarvo on

```
N[Mean[nn]]
```

```
3.29
```

Jos lasketaan sellainen silmälukujen painotettu keskiarvo, jossa kunkin silmäluvun painona on vastaavan tapauksen todennäköisyys, niin saadaan $\sum_{k=1}^{6} k\,\frac{1}{6} = \frac{7}{2} = 3.5$; tätä arvoa sanotaan silmäluvun *odotusarvoksi*. Simuloidun aineiston keskiarvo on lähellä tätä odotusarvoa. Näin on yleisesti: jos kokeita tehdään paljon, niin koetulosten keskiarvo on lähellä odotusarvoa. ■

Määritelmä 2.5 Diskreetin satunnaismuuttujan X *odotusarvo* on

$$E(X) = \sum_{k\in K} k\,p(k),$$

jos $\sum_{k\in K} |k|\, p(k) < \infty$. ■

Esimerkki 2.4

Jos diskreetin satunnaismuutujan X arvojoukko K on äärellinen, niin odotusarvokin on äärellinen. Jos kuitenkin arvojoukko on numeroituvasti ääretön, niin odotusarvon määritelmä voi antaa äärettömän tuloksen. Voi myös olla niin, että summalle $\sum_{k\in K} k\,p(k)$ saadaan erilaisia arvoja siitä riippuen, missä järjestyksessä termit lasketaan yhteen. Tietty analyysin tulos sanoo, että kaikki laskentajärjestykset johtavat samaan tulokseen silloin ja vain silloin, kun summa suppenee itseisesti eli kun $\sum_{k\in K} |k|\, p(k) < \infty$. Jotta siis odotusarvo olisi äärellinen ja hyvin määritelty, niin vaaditaan määritelmän 2.5 tapaan, että summa suppenee itseisesti. Jos näin ei ole, niin sanotaan, että odotusarvo ei ole olemassa.

■ 2.1.3 Muunnoksen odotusarvo

Esimerkki 2.5 Oletetaan, että $P(X = -1) = 0.2$, $P(X = 0) = 0.4$ ja $P(X = 1) = 0.4$.

Satunnaismuuttujan X funktion $Y = g(X)$ (g on mitallinen funktio) odotusarvo saadaan laskemalla ensin Y:n todennäköisyysfunktio $q(l)$, $l \in L$, ja laskemalla sitten $E(Y) = \sum_{l\in L} l\,q(l)$. Seuraavan lauseen mukaan Y:n odotusarvo saadaan myös suoraan X:n todennäköisyysfunktion avulla.

Lause 2.1

a) $E[g(X)] = \sum_{k\in K} g(k)\,p(k)$, jos $\sum_{k\in K} |g(k)|\,p(k) < \infty$.

b) $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$.

c) $E(a + b\,X) = a + b\,E(X)$.

Esimerkki 2.6

■ 2.1.4 Varianssi

Usein merkitään $E(X) = \mu$.

Määritelmä 2.6 Satunnaismuuttujan X *varianssi* on

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right].$$ ■

Lause 2.2 Varianssi voidaan laskea myös seuraavista kaavoista:

a) $\mathrm{Var}(X) = E\left(X^2\right) - \mu^2$.

b) $\mathrm{Var}(X) = E[X(X-1)] + \mu - \mu^2$.

Diskreetin satunnaismuuttujan varianssi voidaan siis määritelmän 2.6 ja lauseen 2.2 mukaan laskea kaavoista

$$\mathrm{Var}(X) = \begin{cases} \sum\limits_{k \in K} (k-\mu)^2\, p(k), \\ \sum\limits_{k \in K} k^2\, p(k) - \mu^2, \\ \sum\limits_{k \in K} k(k-1)\, p(k) + \mu - \mu^2. \end{cases}$$

Kaksi jälkimmäistä kaavaa ovat usein mukavimmat.

Esimerkki 2.7

Satunnaismuuttujan varianssi mittaa sitä, kuinka laajalle satunnaismuuttujan arvot ovat levinneet. Esimerkiksi kahden nopan silmälukujen summan mahdollisten arvojen joukko {2, 3, …, 12} on laajempi kuin yhden nopan tulosjoukko {1, 2, …, 6}. Niinpä silmälukujen summan varianssi onkin suurempi kuin yhden nopan tuloksen varianssi.

Lause 2.3 *a)* $\mathrm{Var}(a + b\,X) = b^2\,\mathrm{Var}(X)$.

b) $\mathrm{Var}(X) \geq 0$; $\mathrm{Var}(X) = 0$ silloin ja vain silloin, kun on olemassa sellainen vakio c, että $P(X = c) = 1$ (tällainen satunnaismuuttuja on ns. *degeneroitunut satunnaismuuttuja*).

c) $E\left[(X-c)^2\right]$ on pienin, kun $c = \mu$; pienin arvo on $\mathrm{Var}(X)$.

Määritelmä 2.7 Satunnaismuuttujan X *hajonta* on varianssin neliöjuuri. ■

Usein merkitään $\mathrm{Var}(X) = \sigma^2$. Hajontaa merkitään usein σ:lla: $\sigma = \sqrt{\mathrm{Var}(X)}$. Hajonnalle käytetään englannissa termiä standard deviation.

2.2 Joitakin diskreettejä jakaumia

■ 2.2.1 Diskreetti tasainen jakauma

Lause 2.4 Jos satunnaismuuttujalla X on n yhtä todennäköistä arvoa $1, 2, \ldots, n$, niin

$$P(X = k) = \tfrac{1}{n}, \quad k = 1, 2, \ldots, n,$$

$$E(X) = \tfrac{n+1}{2}, \quad \mathrm{Var}(X) = \tfrac{n^2-1}{12}.$$

Sanotaan, että satunnaismuuttujalla X on *diskreetti tasainen jakauma* parametrillä n; merkitään $X \sim \mathrm{DU}(n)$.

Esimerkki 2.8

■ 2.2.2 Poisson-jakauma

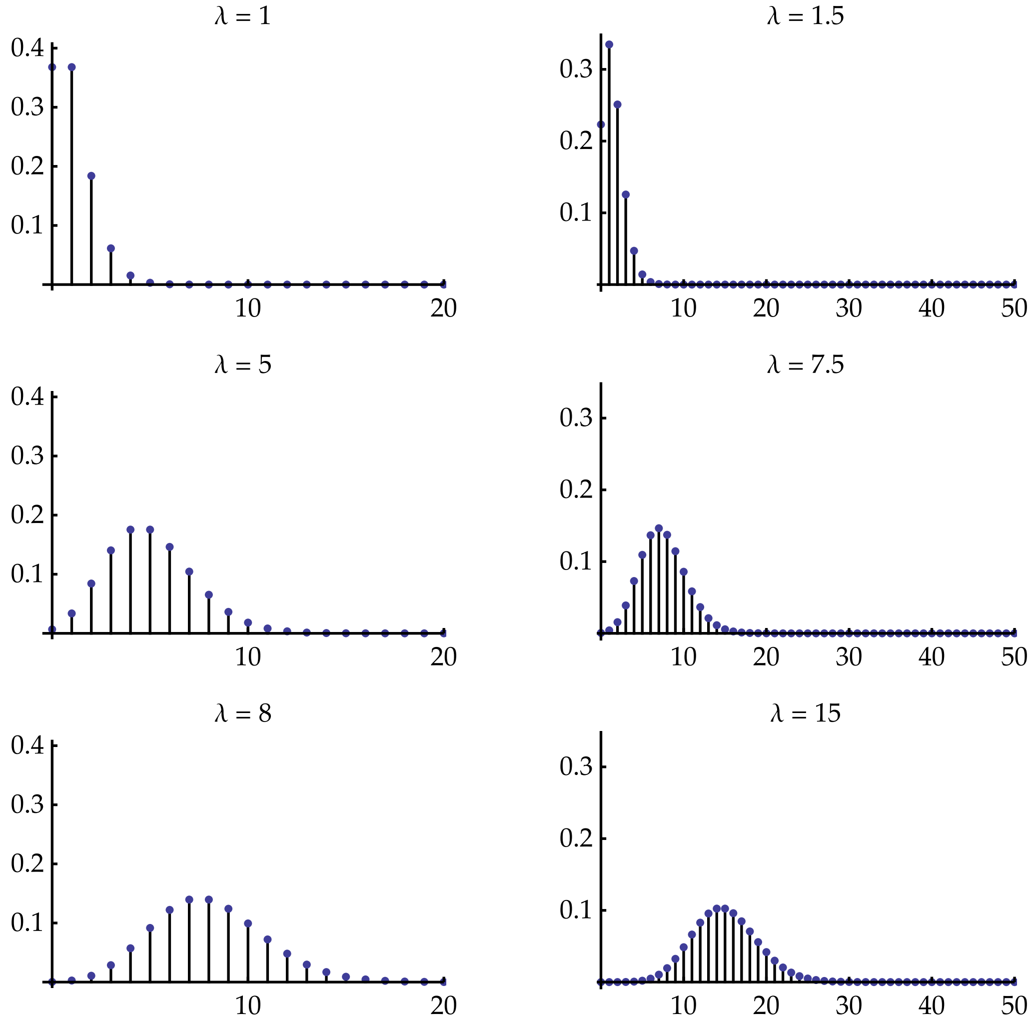
Lause 2.5 Jos

$$P(X = k) = e^{-\lambda}\,\tfrac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots,$$

niin sanotaan, että satunnaismuuttujalla X on *Poisson-jakauma* parametrillä λ; merkitään $X \sim \mathrm{Po}(\lambda)$. Tällöin

$$E(X) = \lambda, \quad \mathrm{Var}(X) = \lambda.$$

Seuraavassa kuvassa on joidenkin Poisson-jakaumien todennäköisyysfunktiot.



Esimerkki 2.9 Vuonna 1910 Rutherford ja Geiger tutkivat poloniumin lähettämää α-säteilyä. He laskivat vastaanotettujen α-partikkeleiden määrän $1/8$ minuutin pituisina aikaväleinä. Aikavälejä oli kaikkiaan 2608 kappaletta. Tulos oli oheisen taulukon kahden ensimmäisen sarakkeen mukainen.

Osoittautui, että jos X on $1/8$ minuutin aikana vastaanotettujen α-partikkeleiden lukumäärä, niin X:llä on likimain Poisson-jakauma. Arvioidaan jakauman parametri λ seuraavalla tavalla. Koska $\lambda = E(X)$ ja odotusarvoa $E(X)$ voidaan approksimoida otoskeskiarvolla m, niin myös λ:aa voidaan approksimoida otoskeskiarvolla m. Nyt m eli keskimäärin $1/8$ minuutin aikavälillä vastaanotettujen α-partikkelien lukumäärä on noin $(0 \cdot 57 + 1 \cdot 203 + 2 \cdot 383 + \ldots + 11 \cdot 6)/2608 = 3.87$. Oletetaan siis, että $X \sim \mathrm{Po}(3.87)$. Lasketaan tämän jakauman todennäköisyydet $P(X = k)$, $k = 0, 1, 2, \ldots, 10$, ja $P(X \geq 11)$ ja kerrotaan todennäköisyydet välien lukumäärällä 2608. Näin saadaan Poisson-jakauman mukaiset teoreettiset frekvenssit. Esimerkiksi $P(X = 0) = e^{-3.87}\, 3.87^0 \big/ 0! = 0.02086$, jolloin teoreettinen frekvenssi on $2608 \cdot 0.02086 = 54.4$. Näin saadaan taulukon kolmas sarake. Neljänteen sarakkeeseen on laskettu havaittujen ja teoreettisten frekvenssien erotukset. Koska erotukset ovat varsin pienet, niin tämä tukee sitä johtopäätöstä, että vastaanotettujen α-partikkelien lukumäärällä $1/8$ minuutin aikana on likimain Po(3.87)-jakauma. ■

Partikkelien lukumäärä	Havaittu frekvenssi	Teoreettinen frekvenssi	Frekvenssien erotus
0	57	54.4	2.6
1	203	210.5	−7.5
2	383	407.4	−24.4
3	525	525.5	−0.5
4	532	508.4	23.6
5	408	393.5	14.5
6	273	253.8	19.2
7	139	140.3	−1.3
8	45	67.9	−22.9
9	27	29.2	−2.2
10	10	11.3	−1.3
11+	6	5.8	0.2

■ 2.2.3 Binomijakauma

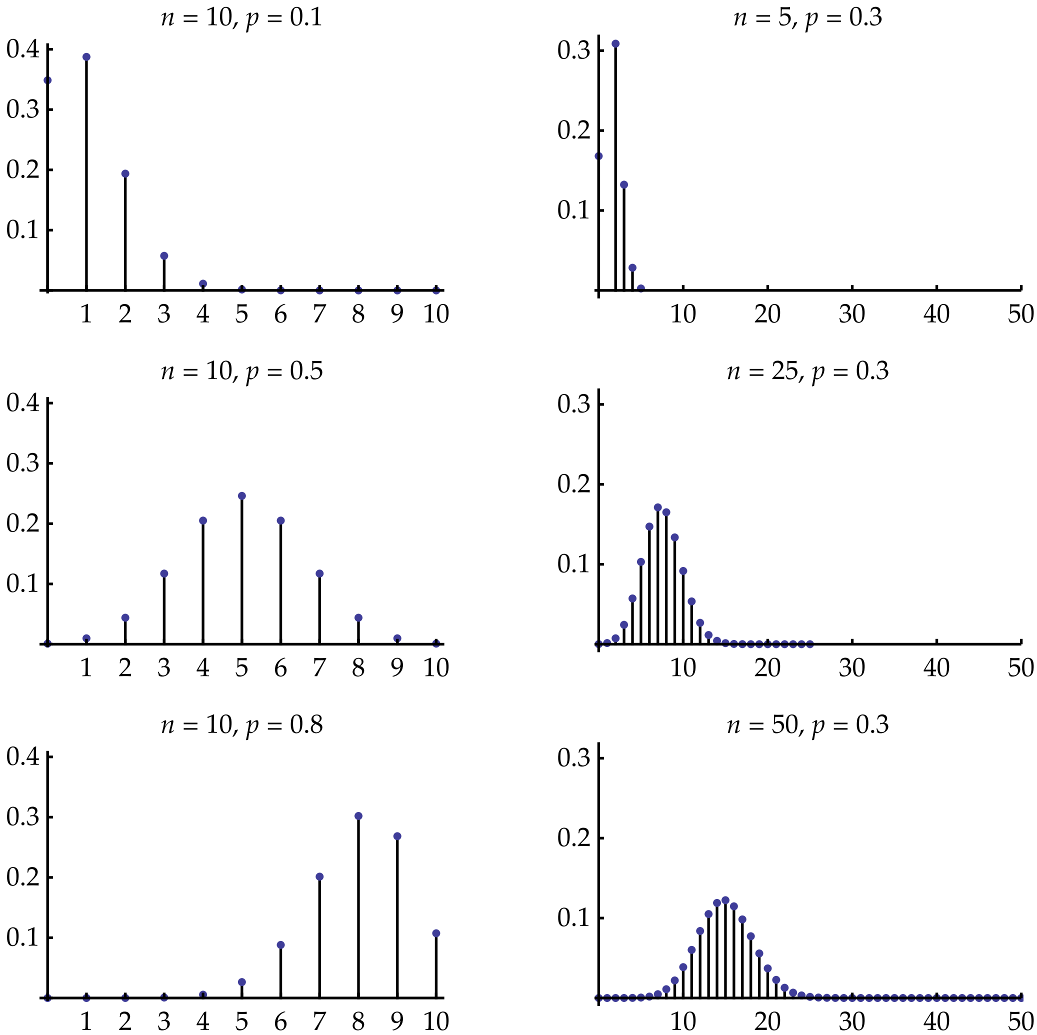
Lause 2.6 Koe toistetaan riippumattomasti n kertaa. Kukin koe onnistuu todennäköisyydellä p; merkitään $q = 1 - p$. Olkoon X onnistuneiden kokeiden lukumäärä. Tällöin

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n,$$

$$E(X) = n\,p, \quad \mathrm{Var}(X) = n\,p\,q.$$

Sanotaan, että satunnaismuuttujalla X on *binomijakauma* parametreillä n ja p; merkitään $X \sim \mathrm{Bin}(n, p)$.

Seuraavassa kuvassa on joidenkin binomijakaumien todennäköisyysfunktiot.

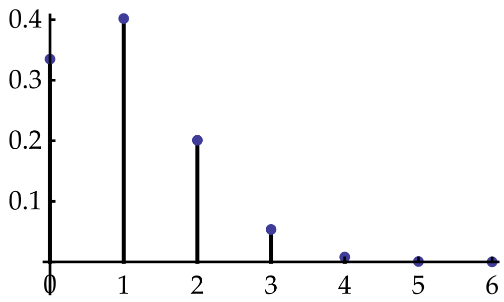


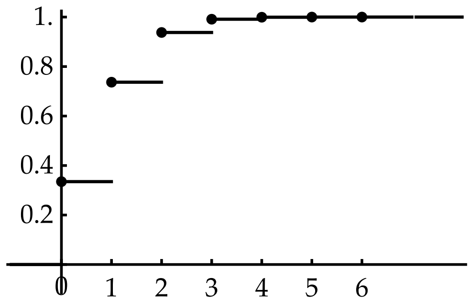
Kuvan vasemmassa sarakkeessa on $n = 10$ ja oikeassa sarakkeessa on $p = 0.3$. Huomataan, että pienehköillä tai suurehkoilla p:n arvoilla ja suurehkoilla n:n arvoilla monet todennäköisyydet ovat hyvin pieniä ja satunnaismuuttuja saa suurella todennäköisyydellä arvoja melko suppealta väliltä.

Kuvassa n:n ja p:n arvot ovat sellaiset, että odotusarvo $n\,p$ on sama kuin Poisson-jakauman odotusarvo λ aikaisemmin Poisson-jakaumalle esitetyissä kuvissa. Kun yllä olevaa kuvaa verrataan Poisson-jakauman kuvaan, niin huomataan, että vasemmassa sarakkeessa, jossa n ei ole suuri ($n = 10$), binomi- ja Poisson-todennäköisyydet eroavat selvästi, mutta oikeassa sarakkeessa, jossa p on melko pieni ($p = 0.3$), todennäköisyydet ovat lähempänä toisiaan, varsinkin sarakkeen kahdessa alimmassa kuvassa, joissa myös n on melko suuri ($n = 25$ ja $n = 50$). Hetken kuluttua osoitetaankin, että binomijakaumaa voidaan tietyin edellytyksin approksimoida Poisson-jakaumalla.

Esimerkki 2.10 Noppaa heitetään 6 kertaa. Olkoon X kuutosten lukumäärä.

k	$p(k)$	$F(k)$
0	0.334898	0.334898
1	0.401878	0.736776
2	0.200939	0.937714
3	0.0535837	0.991298
4	0.00803755	0.999336
5	0.000643004	0.999979
6	0.0000214335	1





Esimerkki 2.11 Oletetaan, että tytön syntymän todennäköisyys on 0.49. Olkoon X kolmilapsisessa perheessä olevien tyttöjen lukumäärä.

k	$p(k)$
0	0.132651
1	0.382347
2	0.367353
3	0.117649

Lause 2.7 Jos $X \sim \mathrm{Bin}(n, p)$ ja n on suuri ja p pieni, niin X:llä on likimain $\mathrm{Po}(n\,p)$-jakauma; merkitään $X \approx \mathrm{Po}(n\,p)$.

Lauseen 2.7 mukainen Poisson-approksimaatio on usein riittävän tarkka, jos $p \leq 0.05$ ja $n \geq 20$.

Esimerkki 2.12 Oletetaan, että kirjan yhdellä sivulla on keskimäärin 2 painovirhettä. Olkoon X yhdellä sivulla olevien painovirheiden lukumäärä. Mikä jakauma voisi X:llä olla?

k	Binomitod.	Poisson−tod.
0	0.13520	0.13534
1	0.27067	0.27067
2	0.27081	0.27067
3	0.18054	0.18045
4	0.09022	0.09022
5	0.03605	0.03609

Lause 2.8 Olkoon $X = 1$, jos koe onnistuu, ja $X = 0$, jos koe epäonnistuu. Oletetaan, että koe onnistuu todennäköisyydellä p; merkitään $q = 1 - p$. Tällöin

$$P(X = 1) = p, \quad P(X = 0) = q,$$

$$E(X) = p, \quad \mathrm{Var}(X) = p\,q.$$

Sanotaan, että satunnaismuuttujalla X on *Bernoulli-jakauma* parametrillä p; merkitään $X \sim \mathrm{Ber}(p)$.

Esimerkki 2.13

2.2.4 Hypergeometrinen jakauma

Lause 2.9 Uurnassa on N palloa, joista M on mustia ja $N - M$ valkoisia. Uurnasta otetaan n palloa palauttamatta. Olkoon X saatujen mustien pallojen lukumäärä. Tällöin

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad \max\{0, n - (N - M)\} \le k \le \min\{M, n\},$$

$$E(X) = n\,p, \quad \mathrm{Var}(X) = n\,p\,q\,\tfrac{N-n}{N-1}, \quad \text{missä } p = \tfrac{M}{N}, \ q = 1 - p.$$

Sanotaan, että satunnaismuuttujalla X on *hypergeometrinen jakauma* parametreillä N, M ja n; merkitään $X \sim \mathrm{Hyp}(N, M, n)$.

Esimerkki 2.14 Laatikossa on 10 arpaa, joista 3 voittaa. Laatikosta otetaan 4 arpaa palauttamatta. Olkoon X saatavien voittoarpojen lukumäärä.

Esimerkki 2.15 Lottoarvonnassa on 39 palloa, jotka on numeroitu 1, 2, …, 39. Voidaan ajatella niin, että palloista 7 on niitä (mustia), jotka ovat omassa ruudukossamme, ja 32 on muita palloja (valkoisia). Lotossa saa siis k oikein, jos lottokone valitsee 7:stä omassa ruudukossamme olevasta numerosta k kappaletta ja 32:sta muusta numerosta $7 - k$ numeroa. Olkoon X oikeiden numeroiden lukumäärä.

k	$p(k)$
0	0.218833
1	0.412416
2	0.274944
3	0.0818286
4	0.0112867
5	0.000677202
6	0.0000145635
7	0.00000006501

Lause 2.10 Jos $X \sim \mathrm{Hyp}(N, M, n)$ ja N on suuri ja $\frac{n}{N}$ pieni, niin X:llä on likimain $\mathrm{Bin}\left(n, \frac{M}{N}\right)$-jakauma; merkitään $X \approx \mathrm{Bin}\left(n, \frac{M}{N}\right)$.

Binomiapproksimaatio on usein riittävän hyvä, jos $N > 50$ ja $\frac{n}{N} < 0.1$.

■ 2.2.5 Geometrinen jakauma

Lause 2.11 Koetta toistetaan riippumattomasti, kunnes koe onnistuu ensimmäisen kerran. Kukin koe onnistuu todennäköisyydellä p; merkitään $q = 1 - p$. Olkoon X tarvittavien kokeiden lukumäärä. Tällöin

$$P(X = k) = q^{k-1}\,p, \quad P(X \le k) = 1 - q^k, \quad k = 1, 2, \ldots,$$

$$E(X) = \tfrac{1}{p}, \quad \mathrm{Var}(X) = \tfrac{q}{p^2}.$$

Sanotaan, että satunnaismuuttujalla X on *geometrinen jakauma* parametrillä p; merkitään $X \sim \mathrm{Geom}(p)$.

Lauseen todistuksessa tarvitaan seuraavia kaavoja (joissa oletetaan, että $|r| < 1$):

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}, \quad \sum_{k=1}^{\infty} k\,r^{k-1} = \frac{1}{(1-r)^2}, \quad \sum_{k=2}^{\infty} k(k-1)\,r^{k-2} = \frac{2}{(1-r)^3}.$$

Ensimmäinen kaava on geometrisen sarjan summakaava, toinen kaava saadaan derivoimalla ensimmäinen kaava ja kolmas kaava saadaan derivoimalla toinen kaava.

Esimerkki 2.16 Noppaa heitetään, kunnes saadaan kuutonen. Olkoon X tarvittavien heittojen lukumäärä.

k	$p(k)$
1	0.166667
2	0.138889
3	0.115741
4	0.0964506
5	0.0803755
6	0.0669796
7	0.0558163
8	0.0465136
9	0.0387613
10	0.0323011

Esimerkki 2.17 Oletetaan, että kun riippumattomia kokeita suoritettiin toistuvasti, niin n ensimmäistä koetta epäonnistuivat. Mikä on tällöin ehdollinen todennäköisyys, että tarvitaan vielä k koetta, ennen kuin saadaan onnistunut koe? Helposti tulee ajatelleeksi, että seuraavalla tai muutamalla seuraavalla kerralla kokeen onnistumisen todennäköisyys olisi tavallista suurempi.

Olkoon X tarvittavien kokeiden lukumäärä, ennen kuin saadaan ensimmäinen onnistunut koe. Kysytty todennäköisyys on

$$P(X = n + k \mid X > n) = \frac{P(X = n + k)}{P(X > n)} = \frac{q^{(n+k)-1}\,p}{q^n} = q^{k-1}\,p = P(X = k).$$

Jos siis n ensimmäistä koetta on epäonnistunut, niin todennäköisyys, että tarvitaan vielä k koetta, ennen kuin saadaan onnistunut koe, on sama kuin todennäköisyys, että alun perin olisi tarvittu k koetta. Tämä on ns. *muistittomuusominaisuus*: X:n jakauma ei "muista" n:ää aiemmin epäonnistunutta koetta. Geometrinen jakauma on ainoa positiivinen, kokonaislukuarvoinen jakauma, jolla on muistittomuusominaisuus.

Huomataan, että erityisesti $P(X = n+1 \mid X > n) = p$. Jos esimerkiksi rahaa on heitetty kauan eikä vieläkään ole tullut yhtään kruunaa, niin helposti ajattelee, että todennäköisyys, että seuraavalla kerralla tulee kruuna, on tavallista suurempi; tämä on ns. pelurin harhaluulo. Todennäköisyys on kuitenkin edelleenkin vain $1/2$.

Totta sen sijaan on, että pitkällä aikavälillä kruunien suhteellinen osuus kaikista tuloksista lähestyy arvoa $1/2$ (tämä osoitetaan kurssilla Todennäköisyyslaskenta II). Suppeneminen ei välttämättä ole kovin nopeaa, kuten pykälän 1.4.3 simuloinnista huomattiin.

Jos rahaa heitetään, kunnes saadaan kruuna, ja X on tarvittavien kokeiden lukumäärä, niin $X \sim \mathrm{Geom}\left(\frac{1}{2}\right)$, joten $E(X) = \frac{1}{1/2} = 2$ ja esimerkiksi $P(X \le 10) = 1 - \left(\frac{1}{2}\right)^{10} = 0.999$. Yleensä ei siis tarvita kovin monta heittoa kruunan saamiseksi. ■

■ 2.2.6 Negatiivibinomijakauma

Lause 2.12 Koetta toistetaan riippumattomasti, kunnes koe onnistuu n:nnen kerran. Kukin koe onnistuu todennäköisyydellä p; merkitään $q = 1 - p$. Olkoon X tarvittavien kokeiden lukumäärä. Tällöin

$$P(X = k) = \binom{k-1}{n-1} p^n q^{k-n}, \quad k = n, n+1, \ldots,$$

$$E(X) = \frac{n}{p}, \quad \mathrm{Var}(X) = \frac{nq}{p^2}.$$

Sanotaan, että satunnaismuuttujalla X on *negatiivibinomijakauma* parametreillä n ja p; merkitään $X \sim \mathrm{Negbin}(n, p)$.

Huomataan, että $\mathrm{Negbin}(1, p) = \mathrm{Geom}(p)$.

Esimerkki 2.18 Noppaa heitetään, kunnes saadaan kuutonen toisen kerran. Olkoon X tarvittavien heittojen lukumäärä.

2.3 Jatkuva satunnaismuuttuja

2.3.1 Jakauma- ja tiheysfunktio

Pykälässä 2.1.1 määriteltiin jakaumafunktio. Esitetään sama määritelmä uudestaan:

Määritelmä 2.8 Satunnaismuuttujan X *jakaumafunktio* on $F(x) = P(X \leq x)$, $x \in \mathbb{R}$. ■

Diskreetti satunnaismuuttuja saa vain äärellisen tai numeroituvasti äärettömän määrän arvoja. Pykälässä 2.1.1 huomattiin, että diskreetin satunnaismuuttujan jakaumafunktio on porrasfunktion tyyppiä, siis epäjatkuva.

Sanotaan, että satunnaismuuttuja on *jatkuva*, jos se voi saada ylinumeroituvan määrän arvoja; tyypillisesti kyse on jostain reaaliakselin välistä. Jatkuvan satunnaismuuttujan jakaumafunktio on jatkuva. Jatkuva satunnaismuuttuja määritellään kuitenkin ns. tiheysfunktion avulla:

Määritelmä 2.9 Satunnaismuuttuja X on *jatkuva*, jos on olemassa sellainen ei-negatiivinen funktio $f(x)$, että kaikille reaalilukujoukoille A on $P(X \in A) = \int_A f(x)\, dx$. Funktio $f(x)$ on satunnaismuuttujan X *tiheysfunktio*. ■

Lause 2.13 Jatkuvan satunnaismuuttujan jakauma- ja tiheysfunktiolla on seuraavat ominaisuudet:

a) $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

b) $F(x) = \int_{-\infty}^{x} f(t)\, dt$.

c) $f(x) = F'(x)$ pisteissä x, joissa $f(x)$ on jatkuva.

Jos siis tiheysfunktio on annettu, niin jakaumafunktio saadaan lauseen b-kohdan avulla. Jos taas jakaumafunktio on annettu, niin tiheysfunktio saadaan lauseen c-kohdan avulla; pisteissä, joissa $F'(x)$ ei ole olemassa, voidaan määritellä $f(x) = 0$.

Jokainen tiheysfunktio toteuttaa ehdot

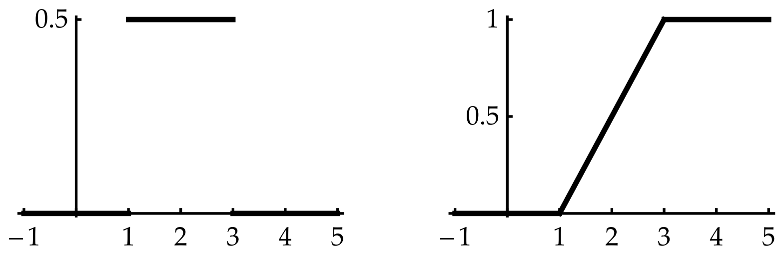
- $f(x) \geq 0$, $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

Voitaisiin myös osoittaa, että jos jokin funktio $f(x)$ toteuttaa nämä ehdot, niin on olemassa satunnaismuuttuja, jonka tiheysfunktio on $f(x)$. Mainitut kaksi ehtoa ovat siis välttämättömät ja riittävät ehdot sille, että funktio $f(x)$ on tiheysfunktio. Huomaa, että tiheysfunktio voi hyvin saada ykköstä suurempia arvoja.

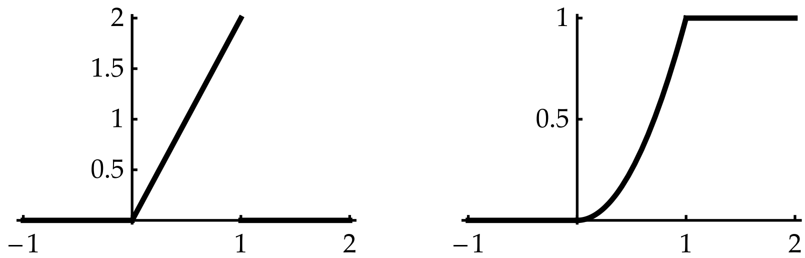
Lauseen 2.13 b-kohdasta nähdään, että jatkuvan satunnaismuuttujan jakaumafunktiolla on seuraavat ominaisuudet:

- $0 \le F(x) \le 1$, $x \in \mathbb{R}$;
- $F(x)$ on jatkuva;
- $F(x)$ on ei-vähenevä;
- $\lim_{x\to\infty} F(x) = 1$;
- $\lim_{x\to-\infty} F(x) = 0$.

Esimerkki 2.19 Reaaliluku X valitaan satunnaisesti väliltä (a, b).

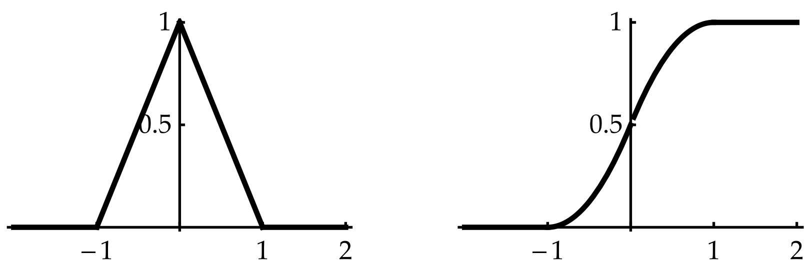


Esimerkki 2.20 Piste valitaan satunnaisesti yksikköympyrästä. Olkoon X pisteen etäisyys ympyrän keskipisteestä.



Esimerkki 2.21 Oletetaan, että satunnaismuuttujan X tiheysfunktio on muotoa

$$f(x) = \begin{cases} 0, & x \le -1, \\ c(1+x), & -1 < x \le 0, \\ c(1-x), & 0 < x \le 1, \\ 0, & x > 1. \end{cases}$$

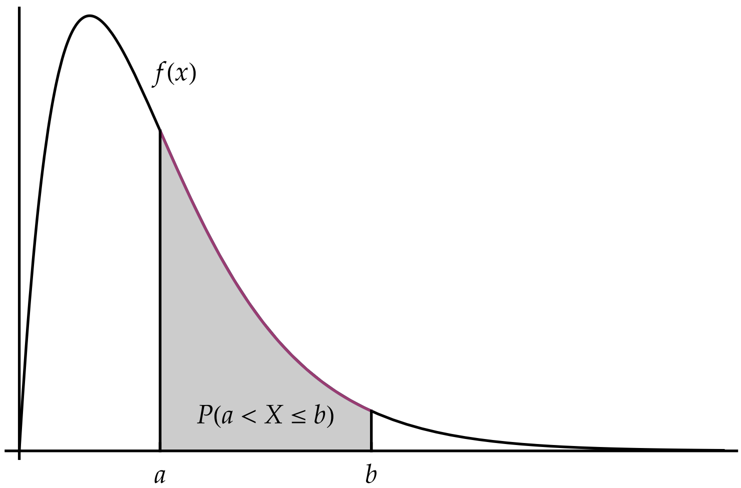


2.3.2 Todennäköisyyksien laskeminen

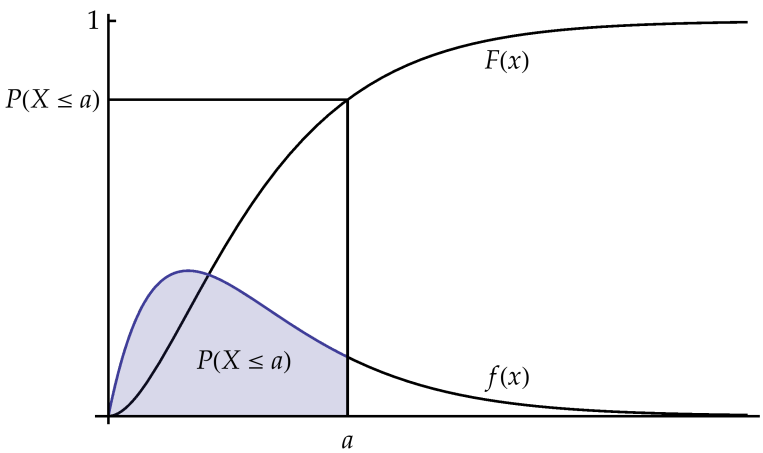
Jatkuvaan satunnaismuuttujaan liittyviä todennäköisyyksiä voidaan laskea seuraavasti ($a \le b$):

Todennäköisyys	$F(x)$:n avulla	$f(x)$:n avulla
$P(X \le a)$	$F(a)$	$\int_{-\infty}^{a} f(t)\,dt$
$P(X > a)$	$1 - F(a)$	$\int_{a}^{\infty} f(t)\,dt$
$P(a \le X \le b)$	$F(b) - F(a)$	$\int_{a}^{b} f(t)\,dt$

Todennäköisyyksiä voidaan siis laskea sekä jakauma- että tiheysfunktion avulla. Jos jakaumafunktio on käytettävissä, niin sitä kannattaa käyttää, koska silloin vältytään intergoinnilta. Jos todennäköisyys lasketaan tiheysfunktion avulla, niin silloin tulee itse asiassa lasketuksi tiheysfunktion ja tarkasteltavan x-akselin osan välisen alueen pinta-ala. Tämä on havainnollistettu seuraavassa kuvassa.



Seuraavassa kuvassa on saman satunnaismuuttujan jakaumafunktio ja tiheysfunktio. Kuva ilmoittaa, millä tavalla todennäköisyys $P(X \le a)$ liittyy kumpaankin funktioon.



Koska $f(x) = F'(x)$, niin $f(x)$ ilmoittaa, kuinka voimakkaasti $F(x)$ kasvaa pisteessä x. Mitä voimakkaampi kasvu, sitä nopeammin todennäköisyyttä kertyy funktioon $F(x) = P(X \leq x)$, kun x kasvaa.

Voidaan myös kirjoittaa

$$P(a < X < a + \epsilon) = \int_a^{a+\epsilon} f(x)\, dx \simeq \epsilon\, f(a),$$

joten $f(a)$:n suuruus määrää, kuinka todennäköistä on, että X saa arvon läheltä pistettä a. Näin tiheysfunktion arvo $f(a)$ ilmoittaa, kuinka "tiheässä" todennäköisyyttä on pisteen a lähellä tai mikä on sen todennäköisyyden "intensiteetti", että X saa arvon a. Tietyn pituisen välin todennäköisyys on suurempi siellä, missä tiheysfunktio saa suurempia arvoja, kuin siellä, missä tiheysfunktio saa pienempiä arvoja.

Jatkuvalle satunnaismuuttujalle pätee myös

$$P(X = a) = 0,$$

sillä $P(a \leq X \leq a) = \int_a^a f(t)\, dt = 0$. Näin ollen todennäköisyys, että jatkuva satunnaismuuttuja saa tietyn arvon, on nolla. Huomaa siis erityisesti, että $f(a)$ ei ole sama kuin $P(X = a)$. Vain tiheysfunktion määrätyillä integraaleilla on todennäköisyystulkinnat.

Kaavasta $P(X = a) = 0$ seuraa, että kun lasketaan välien todennäköisyyksiä jatkuvalle satunnaismuuttujalle, niin ei ole merkitystä, ovatko välin päätepisteet mukana vai eivät. Esimerkiksi

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Esimerkki 2.22

2.3.3 Odotusarvo ja varianssi

Määritelmä 2.10 Jatkuvan satunnaismuuttujan X *odotusarvo* on

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx,$$

jos $\int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty$. ■

Esimerkki 2.23

Diskreettien satunnaismuuttujien kohdalla esitetty varianssin määritelmä pätee myös jatkuville satunnaismuuttujille:

Määritelmä 2.11 Satunnaismuuttujan X *varianssi* on

$$\mathrm{Var}(X) = E\left[(X - \mu)^2\right].$$ ■

Lauseiden 2.2, 2.3 ja 2.4 tapaan voidaan osoittaa seuraavat tulokset:

Lause 2.14

a) $E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$, jos $\int_{-\infty}^{\infty} |g(x)|\, f(x)\, dx < \infty$.

b) $\mathrm{Var}(X) = E\left(X^2\right) - \mu^2$.

c) $E(a + b\, X) = a + b\, E(X)$.

d) $\mathrm{Var}(a + b\, X) = b^2\, \mathrm{Var}(X)$.

Varianssi voidaan siis laskea joko määritelmän 2.11 tai lauseen 2.14 b-kohdan avulla:

$$\mathrm{Var}(X) = \begin{cases} \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx, \\ \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2. \end{cases}$$

Jälkimmäinen tapa on yleensä mukavampi.

Esimerkki 2.24

Lause 2.15 Jos X saa ei-negatiivisia arvoja, niin

$E(X) = \int_0^{\infty} [1 - F(x)]\, dx,$

$E(X^2) = 2 \int_0^{\infty} x[1 - F(x)]\, dx.$

2.3.4 Moodi, mediaani ja kvantiili

Odotusarvo $\mu = E(X)$ on käytetyin satunnaismuuttujan sijaintia ilmaiseva tunnusluku. Sijaintia voidaan mitata muillakin tunnusluvuilla.

Määritelmä 2.12 Diskreetin satunnaismuuttujan *moodi* on se k:n arvo (tai ne k:n arvot), jossa (tai joissa) todennäköisyysfunktio $p(k)$ saa maksimiarvon.

Jatkuvan satunnaismuuttujan *moodi* on se x:n arvo (tai ne x:n arvot), jossa (tai joissa) tiheysfunktio $f(x)$ saa maksimiarvon. ■

Moodia voidaan sanoa satunnaismuuttujan *todennäköisimmäksi arvoksi*. Jatkuvan satunnaismuuttujan kunkin arvon todennäköisyys on kylläkin nolla, mutta moodin ympärillä on "todennäköisyysmassaa" eniten.

Esimerkki 2.25

Määritelmä 2.13 Diskreetin satunnaismuuttujan *mediaani* on se k:n arvo (tai ne k:n arvot), jossa (tai joissa) jakaumafunktio $F(x)$ toteuttaa $F(k-1) \le \frac{1}{2} \le F(k)$.

Jatkuvan satunnaismuuttujan *mediaani* on se x:n arvo (tai ne x:n arvot), jossa (tai joissa) jakaumafunktio toteuttaa $F(x) = \frac{1}{2}$. ■

Mediaani jakaa populaation kahteen yhtä suureen osaan: mediaanin ala- ja yläpuolella on puolet populaatiosta.

Esimerkki 2.26

Määritelmä 2.14 Diskreetin satunnaismuuttujan *p-kvantiili* on se k:n arvo (tai ne k:n arvot), jossa (tai joissa) jakaumafunktio $F(x)$ toteuttaa $F(k-1) \le p \le F(k)$.

Jatkuvan satunnaismuuttujan *p-kvantiili* on se x:n arvo (tai ne x:n arvot), jossa (tai joissa) jakaumafunktio toteuttaa $F(x) = p$. ■

Huomataan, että $\frac{1}{2}$-kvantiili on mediaani.

2.4 Normaalijakauma

2.4.1 Standardoitu normaalijakauma

Lause 2.16 Jos satunnaismuuttujan X tiheysfunktio on

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty,$$

niin sanotaan, että X:llä on *standardoitu normaalijakauma*; merkitään $X \sim N(0, 1)$. Tällöin

$$E(X) = 0, \quad \mathrm{Var}(X) = 1.$$

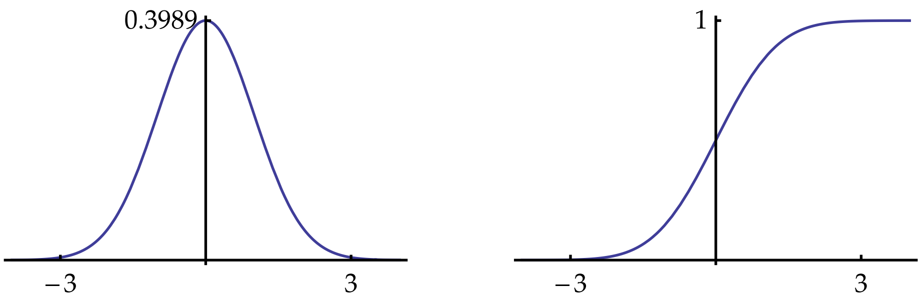
Jakaumafunktiota merkitään

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt, \quad -\infty < x < \infty.$$

On voimassa

$$\Phi(-x) = 1 - \Phi(x).$$

Seuraavassa kuvassa on standardoidun normaalijakauman tiheys-ja jakaumafunktio.



Huomataan, että tiheysfunktio $\phi(x)$ on symmetrinen origon suhteen ja että välin $(-3, 3)$ ulkopuolella on hyvin vähän todennäköisyyttä.

Standardoidun normaalijakauman jakaumafunktiota ei voida ilmaista alkeisfunktioiden avulla. Tämän johdosta jakaumafunktion arvoja on taulukoitu; yksi taulukko on tämän monisteen lopussa. Taulukoista saadaan yleensä todennäköisyyksiä $\Phi(x) = P(X \leq x)$, kun x on positiivinen. Negatiivisilla x:n arvoilla voidaan käyttää lauseen 2.16 kaavaa $\Phi(-x) = 1 - \Phi(x)$.

Esimerkki 2.27

2.4.2 Yleinen normaalijakauma

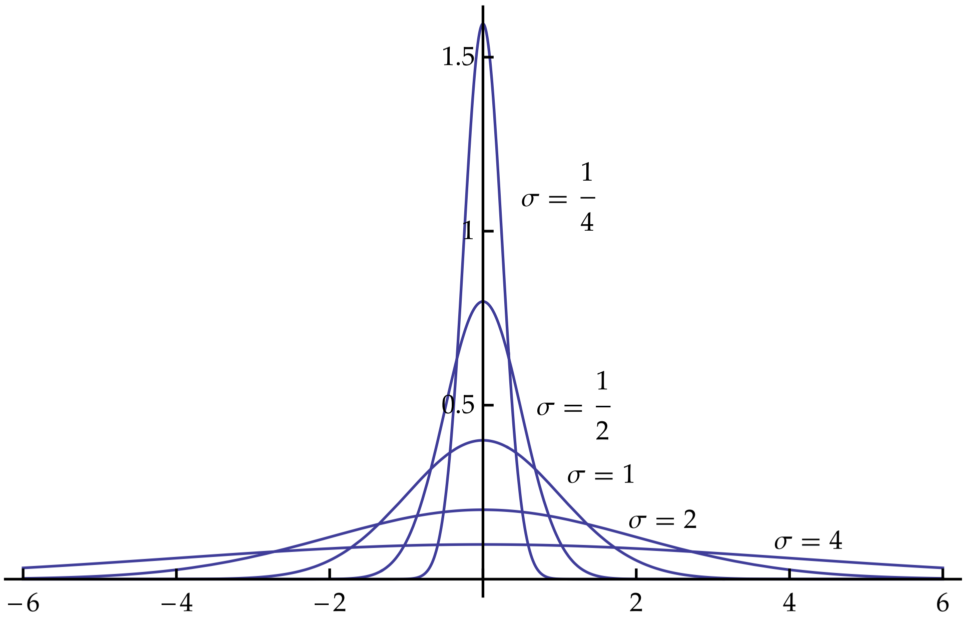
Lause 2.17 Jos $Y \sim N(0, 1)$, niin sanotaan, että satunnaismuuttujalla $X = \mu + \sigma Y$, missä $\sigma > 0$, on *normaalijakauma* (eli *Gaussin jakauma*) $N(\mu, \sigma^2)$. Tällöin

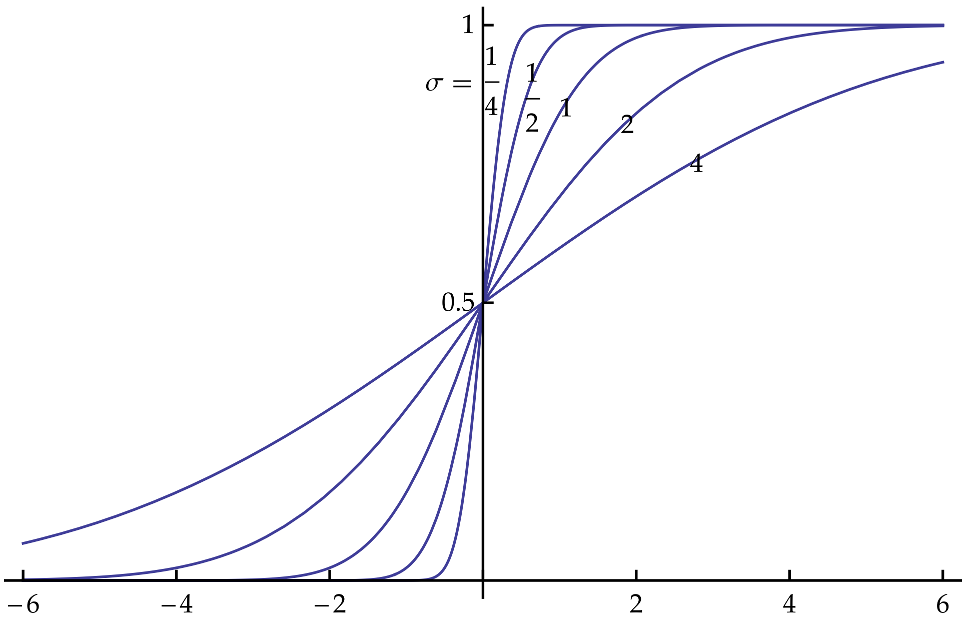
$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty,$$

$$E(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2.$$

Huomaa, että tällä kurssilla normaalijakauman $N(\mu, \sigma^2)$ toinen parametri on varianssi σ^2 (siis ei hajonta σ, kuten on esim. MAOL-taulukossa).

Seuraavissa kuvissa on normaalijakauman $N(0,\sigma^2)$ tiheys- ja jakaumafunktio, kun $\sigma = \frac{1}{4}, \frac{1}{2}, 1, 2,$ ja 4.





Jos $X \sim N(\mu,\sigma^2)$, niin lauseen 2.17 mukaan todennäköisyyksiä voidaan laskea standardoidun normaalijakauman jakaumafunktion $\Phi(x)$ avulla kaavasta

$$\boxed{P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right).}$$

Tähän kaavaan voidaan päätyä myös toteamalla, että yhtälön $X = \mu + \sigma Y$ mukaan $Y = \frac{X-\mu}{\sigma}$. Näin ollen $\frac{X-\mu}{\sigma} \sim N(0,1)$, joten

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Koska $E\left(\frac{X-\mu}{\sigma}\right) = 0$ ja $\mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = 1$, niin satunnaismuuttujaa $\frac{X-\mu}{\sigma}$ sanotaan *standardoiduksi satunnaismuuttujaksi*.

Esimerkki 2.28

Lause 2.18 Jos $X \sim N(\mu, \sigma^2)$, niin

$$P(\mu - 1.96\,\sigma < X < \mu + 1.96\,\sigma) = 0.95,$$

$$P(\mu - 2.58\,\sigma < X < \mu + 2.58\,\sigma) = 0.99,$$

$$P(\mu - 3.29\,\sigma < X < \mu + 3.29\,\sigma) = 0.999.$$

Lauseen 2.18 mukaisia välejä $\mu - a\,\sigma < X < \mu + a\,\sigma$ sanotaan satunnaismuuttujan X *luottamusväleiksi*. Lauseessa on annettu 95:n, 99:n ja 99.9:n prosentin luottamusvälit (ne on mainittu myös todennäköisyyslaskennan kaavakokoelmassa). Huomataan esimerkiksi, että jos $X \sim N(\mu, \sigma^2)$, niin noin 95 % tapauksista on välillä, joka ulottuu kahden hajonnan verran odotusarvosta vasemmalle ja kahden hajonnan verran odotusarvosta oikealle. Jos odotusarvosta mennään noin kolmen hajonnan verran vasemmalle ja oikealle, niin tällä välillä on jo lähes kaikki tapaukset. Koska odotusarvo μ on luottamusvälien keskipisteessä, niin välit ovat ns. *odotusarvokeskisiä* välejä.

Esimerkki 2.29 Raskauden kestolla X on likimain jakauma $N(266, 16^2)$.

Normaalijakaumaa käytti ensin ranskalainen matemaatikko Abraham DeMoivre vuonna 1733, kun hän laski rahanheittoon liittyviä todennäköisyyksiä; hän kutsui tiheysfunktiota eksponentiaaliseksi *kellomuotoiseksi* funktioksi.

Vuonna 1809 saksalainen matemaatikko Karl Fridrich Gauss käytti normaalijakaumaa, kun hän ennusti havaintojen perusteella astronomisten kohteiden paikkoja; pian tämän jälkeen jakaumaa ruvettiin kutsumaan *Gaussin jakaumaksi*.

Koska huomattiin, että monet havaintoaineistot noudattivat Gaussin jakaumaa, alettiin ajatella, että tämä jakauma on ikään kuin ”normaali” havaintojen jakauma. Tämän johdosta, varsinkin brittiläisen tilastotieteilijän Karl Pearsonin vaikutuksesta, jakaumaa alettiin kutsua *normaalijakaumaksi*.

■ 2.4.3 Summan odotusarvo ja varianssi

Tämän pykälän ja myös kolmen seuraavan pykälän asia kuuluu kurssiin Todennäköisyyslaskenta II, jossa perehdytään satunnaisvektoreihin. Näitä asioita tarkastellaan kuitenkin myös tällä kurssilla, koska ne ovat hyvin tärkeitä ja myös melko helppo omaksua ilman perehtymistä satunnaisvektoreiden teoriaan.

Seuraavassa tullaan tarvitsemaan satunnaismuuttujien *riippumattomuuden* käsitettä. Riippumattomuus määritellään matemaattisesti kurssilla Todennäköisyyslaskenta II, mutta tällä kurssilla riittänee todeta, satunnaismuuttujat ovat riippumattomat, jos niillä ei ole vaikutusta toistensa saamiin arvoihin. Jos esimerkiksi X on ensimmäisen nopan tulos ja Y toisen nopan tulos, niin X ja Y ovat riippumattomat.

Satunnaismuuttujien summa ja aritmeettinen keskiarvo

$$S_n = X_1 + \ldots + X_n,$$
$$\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$$

ovat tärkeitä satunnaismuuttujia. Summan odotusarvo ja varianssi voidaan laskea seuraavan lauseen avulla.

Lause 2.19

a) $E(S_n) = \sum_{i=1}^{n} E(X_i)$.

b) $\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$, jos satunnaismuuttujat $X_1, \ldots, X_n$ ovat riippumattomat.

Huomaa, että odotusarvoa koskeva kaava pätee, vaikka satunnaismuuttujat olisivat riippuvia.

Seuraus 2.1 Jos satunnaismuuttujat $X_1, \ldots, X_n$ ovat riippumattomat ja niillä on sama odotusarvo μ ja sama varianssi σ^2, niin

a) $E(S_n) = n\,\mu$, $\mathrm{Var}(S_n) = n\,\sigma^2$.

b) $E(\overline{X}_n) = \mu$, $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$.

On yllättävää, että odotusarvon ja varianssin lisäksi tiedetään (tietyin edellytyksin) summan likimääräinen jakaumakin, vaikka summattavien jakauma olisi tuntematon. Voidaan nimittäin osoittaa, että riippumattomien ja samoin jakautuneiden satunnaismuuttujien summalla on likimäärin *normaalijakauma*, jos summattavia on paljon. Tämä on ns. keskeinen raja-arvolause.

■ 2.4.4 Keskeinen raja-arvolause

Lause 2.20 (*Keskeinen raja-arvolause*) Oletetaan, että satunnaismuuttujat X_i, $i = 1, 2, \ldots$, ovat riippumattomat ja samoin jakautuneet ja että $E(X_i) = \mu$ ja $\mathrm{Var}(X_i) = \sigma^2$. Silloin suurilla n:n arvoilla

a) S_n:llä on likimain $N(n\,\mu,\, n\,\sigma^2)$-jakauma;

b) $\overline{X}_n$:llä likimain $N\left(\mu,\, \frac{\sigma^2}{n}\right)$-jakauma.

Sanotaan myös, että S_n ja $\overline{X}_n$ ovat *asymptoottisesti normaalisti jakautuneet*, ja merkitään

$$S_n \sim \mathrm{AsN}(n\,\mu,\, n\,\sigma^2), \quad \overline{X}_n \sim \mathrm{AsN}\left(\mu,\, \frac{\sigma^2}{n}\right).$$

Seurauksen 2.1 perusteella S_n:n ja $\overline{X}_n$:n likimääräisten normaalijakaumien parametrit ovat selvät, mutta lauseen uutuus on siis se, että kyseessä on likimääräinen normaalijakauma. Likimääräisiä todennäköisyyksiä voidaan siis laskea seuraavasti:

$$P(S_n \le x) \simeq \Phi\left(\frac{x - n\,\mu}{\sqrt{n}\,\sigma}\right), \quad P(\overline{X}_n \le x) \simeq \Phi\left(\frac{x - \mu}{\sigma / \sqrt{n}}\right).$$

Esimerkki 2.30 Noppaa heitetään 100 kertaa. Olkoon X_i i:nnen heiton tulos ja $S_{100} = X_1 + \ldots + X_{100}$ silmälukujen summa.

Voidaan osoittaa, että jos satunnaismuuttujilla X_i on $N(\mu, \sigma^2)$-jakauma, $i = 1, \ldots, n$, niin satunnaismuuttujilla S_n ja $\overline{X}_n$ on *tarkat* normaalijakaumat $N(n\,\mu, n\,\sigma^2)$ ja $N\left(\mu, \frac{\sigma^2}{n}\right)$.

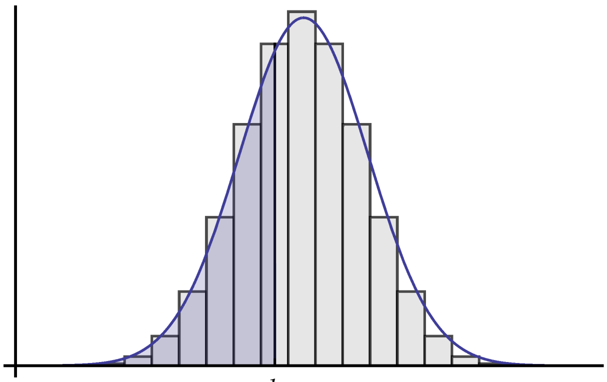
2.4.5 Jatkuvuuskorjaus

Keskeinen raja-arvolause antaa summan todennäköisyydelle likiarvon

$$P(S_n \le x) \simeq \Phi\left(\frac{x - n\,\mu}{\sqrt{n}\;\sigma}\right),$$

kun n on suuri. Jos kuitenkin satunnaismuuttujat X_i ovat diskreettejä, niin tätä todennäköisyyden approksimaatiota voidaan hiukan parantaa ns. jatkuvuuskorjauksella.

Seuraavassa kuvassa on esitetty erään diskreetin satunnaismuuttujan X todennäköisyysfunktio pylväskuviona. Kunkin pylvään korkeus on todennäköisyys, että X saa vastaavan kokonaislukuarvon. Koska pylväiden leveys on yksi, niin myös pylvään pinta-ala on todennäköisyys, että X saa vastaavan kokonaislukuarvon. Todennäköisyys $P(X \le k)$ saadaan siis laskemalla vastaavien pylväiden pinta-alojen summa.



Kuvaan on piirretty myös sellaisen normaalijakauman tiheysfunktio, joka approksimoi X:n diskreettiä jakaumaa. Jos lasketaan $P(X \le k)$ tämän normaalijakauman avulla, niin tulee lasketuksi tiheysfunktion ja x-akselin välisen alueen pinta-ala pisteestä k vasemmalle. Huomataan, että tällöin tapauksen $X = k$ todennäköisyydestä eli tähän tapaukseen liittyvän pylvään pinta-alasta tulee otetuksi mukaan vain noin puolet. Tästä syystä kannattakin laskea tapauksen $X \le k + 0.5$ todennäköisyys. Tätä menettelyä kutsutaan *jatkuvuuskorjaukseksi*. Siinä korjataan se pieni virhe, joka syntyy, kun diskreettiä jakaumaa approksimoidaan jatkuvalla normaalijakaumalla.

Jos satunnaismuuttujat X_i ovat diskreettejä, niin myös S_n on diskreetti, joten kun lasketaan S_n:ään liittyviä todennäköisyyksiä, kannattaa käyttää jatkuvuuskorjausta:

$$P(S_n \le k) \simeq \Phi\left(\frac{k + 0.5 - n\,\mu}{\sqrt{n}\;\sigma}\right).$$

Samaan tapaan

$$P(k \le S_n \le l) \simeq \Phi\left(\frac{l + 0.5 - n\,\mu}{\sqrt{n}\;\sigma}\right) - \Phi\left(\frac{k - 0.5 - n\,\mu}{\sqrt{n}\;\sigma}\right).$$

Esimerkki 2.31 Tarkastellaan uudestaan esimerkkiä 2.30.

Esimerkki 2.32 Uudelle asuinalueelle suunnitellaan asuntoja tuhannelle perheelle. Kussakin perheessä on 0, 1, 2 tai 3 lasta todennäköisyyksin 0.35, 0.45, 0.15 ja 0.05. Tarkastellaan lasten yhteismäärää. Olkoon X_i i:nnen perheen lasten lukumäärä.

2.4.6 Jakaumien approksimointi

Keskeisen raja-arvolauseen avulla voidaan useille jakaumille johtaa normaalijakauma-approksimaatiot. Seuraavassa lauseessa on tulokset binomi- ja Poisson-jakaumalle.

Lause 2.21 *a)* Jos $X \sim \text{Bin}(n, p)$, niin $X \sim \text{AsN}(n\,p,\ n\,p\,q)$, jos n on suuri.
b) Jos $X \sim \text{Po}(\lambda)$, niin $X \sim \text{AsN}(\lambda,\ \lambda)$, jos λ on suuri.

Normaalijakauma-approksimaatio on yleensä riittävän tarkka binomijakaumalle, jos $n\,p\,q \geq 10$, ja Poisson-jakaumalle, jos $\lambda > 15$.

Esimerkki 2.33 Noppaa heitetään 100 kertaa. Olkoon X kuutosten lukumäärä.

Esimerkki 2.34 Olkoon X tietyn tyyppisten onnettomuuksien lukumäärä tietyllä alueella tietyllä aikavälillä.

2.5 Muita jatkuvia jakaumia

2.5.1 Tasainen jakauma

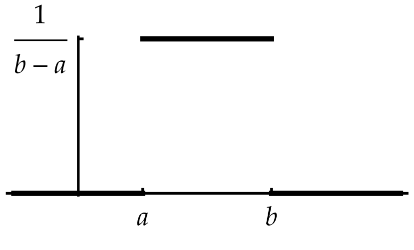
Lause 2.22 Jos satunnaismuuttujan X tiheysfunktio on

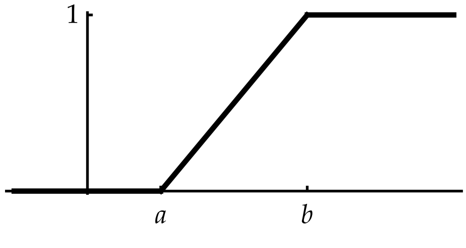
$$f(x) = \begin{cases} 0 & x \leq a, \\ \frac{1}{b-a} & a < x < b, \\ 0 & x \geq b, \end{cases}$$

missä $a < b$, niin sanotaan, että X:llä on *tasainen jakauma* välillä (a, b); merkitään $X \sim U(a, b)$. Tällöin

$$F(x) = \begin{cases} 0 & x \leq a, \\ \frac{x-a}{b-a} & a < x < b, \\ 1 & x \geq b, \end{cases} \quad E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Seuraavassa kuvassa on tiheys- ja jakaumafunktio:





Tasaisen jakauman nimitys tulee siitä, että tiheysfunktio on vakio eli tasainen välillä (a, b). Kuten esimerkissä 2.28 todettiin, voidaan välien todennäköisyyksiä laskea janojen pituuksien suhteina: $P(c \le X \le d) = \frac{d-c}{b-a}$, jos $a < c < d < b$. Tästä johtuu, että tasaisella jakaumalla on se ainutlaatuinen ominaisuus, että X saa arvon tietyltä väliltä yhtä suurella todennäköisyydellä kuin miltä tahansa saman pituiselta väliltä. Sanotaan myös, että X saa arvoja *satunnaisesti* väliltä (a, b).

Esimerkki 2.35 Bussi kulkee tietyn pysäkin kautta 10 minuutin välein. Oletetaan, että opiskelija saapuu pysäkille satunnaisena hetkenä. Olkoon X odotusaika, ennen kuin bussi saaapuu.

■ 2.5.2 Eksponenttijakauma

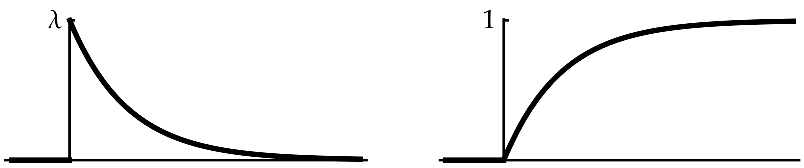
Lause 2.23 Jos satunnaismuuttujan X tiheysfunktio on

$$f(x) = \begin{cases} 0 & x \le 0, \\ \lambda e^{-\lambda x} & x > 0, \end{cases}$$

missä $\lambda > 0$, niin sanotaan, että X:llä on *eksponenttijakauma* parametrillä λ; merkitään $X \sim \mathrm{Exp}(\lambda)$. Tällöin

$$F(x) = \begin{cases} 0 & x \le 0, \\ 1 - e^{-\lambda x} & x > 0, \end{cases} \quad E(X) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Seuraavassa kuvassa on tyypillinen eksponenttijakauman tiheys- ja jakaumafunktio:



Esimerkki 2.36 Olkoon X tietyn komponentin elinikä; oletetaan, että $X \sim \mathrm{Exp}(\lambda)$. Tiedetään, että komponentin keskimääräinen elinikä on 1000.

Esimerkki 2.37 Radioaktiivisen isotoopin (esim. strontium 90) atomit pysyvät vakaina satunnaisen ajan, jonka jälkeen ne hajoavat (muuttuvat joksikin toisenlaiseksi atomiksi) ja lähettävät samalla säteilyä tai partikkeleita. Tällaisen atomin elinaikaa T (eli aikaa ennen hajoamista) voidaan kuvata $\mathrm{Exp}(\lambda)$-jakaumalla; parametriä λ sanotaan isotoopin *hajoamisnopeudeksi*. Radioaktiivisen isotoopin *puoliintumisaika* h on se aika, jonka kuluessa puolet isotoopeista on hajonnut, eli se aika, jonka atomi korkeintaan elää tn:llä $1/2$.

Lause 2.24 Eksponenttijakaumalla on ns. *muistittomuusominaisuus*:

$$P(X > t + h \mid X > t) = P(X > h).$$

Muistittomuusominaisuus on havainnollista tulkita laitteen eliniän avulla. Oletetaan siis, että laitteen eliniällä X on eksponenttijakauma. Jos laite on kestänyt hetkeen t saakka ($X > t$), niin todennäköisyys, että laite kestää vielä ainakin h aikayksikköä ($X > t + h$), on sama kuin todennäköisyys, että uusi laite kestää ainakin h aikayksikköä ($X > h$). Tieto nykyisestä elinajasta ei siis anna mitään informaatiota jäljellä olevasta elinajasta. Tämä merkitsee, että laite ei kulu tai vanhene käytön mukana: käytössä ollut laite on yhtä hyvä kuin uusi laite. Laite ei tavallaan ”muista” sitä, että se on jo ollut käytössä. Tällöin laitteen jäljellä olevaan elinaikaan eivät vaikuta laitteen *sisäiset* seikat (koska laite ei vanhene) eikä se, kuinka kauan laite on jo ollut toiminnassa, vaan *ulkoiset* seikat: laite rikkoontuu, kun sen käyttöympäristössä tapahtuu riittävän suuri (laitteen kestokyvyn ylittävä) häiriö (esim. jänitepiikki).

Eksponenttijakauma on ainoa ei-negatiivinen, jatkuva jakauma, jolla on muistittomuusominaisuus.

Eksponenttijakaumaa käytetään usein approksimoimaan mm. seuraavien satunnaismuuttujien jakaumaa: laitteen elinikä, rikkoontuneen laitteen korjausaika, asiakkaiden saapumisten välinen aika jonosysteemissä, asiakkaan palveluaika.

3 Tilastotiedettä

3.1 Jakauman parametrien estimointi

■ 3.1.1 Momenttimenetelmä

Toisinaan tiedetään, että jollain satunnaismuuttujalla on (ainakin likimain) tietty jakauma mutta tämän jakauman parametrit ovat tuntemattomia. Parametrit voidaan arvioida eli estimoida, jos on käytettävissä satunnaismuuttujan arvoja eli havaintoja. Yleisimmin käytetyt estimointimenetelmät ovat momenttimenetelmä ja suurimman uskottavuuden menetelmä.

Tarkastellaan esimerkkinä tiettyä ihmispopulaatiota. Tässä populaatiossa ihmisten pituus on satunnaismuuttuja X, jolla on tietty $N(\mu, \sigma^2)$-jakauma. Ollaan kiinnostuneita estimoimaan jakaumassa esiintyvät parametrit μ ja σ^2. Sitä varten valitaan populaatiosta n satunnaista ihmistä ja mitataan heidän pituutensa. Olkoon X_i i:nnen ihmisen pituus; X_i on satunnaismuuttuja, jolla on sama jakauma kuin muuttujalla X. Kun mittaus on tehty, on satunnaismuuttujalle X_i saatu arvo x_i. Tällaista riipumattomien, samoin jakautuneiden satunnaismuuttujien joukkoa $\{X_1, \ldots, X_n\}$ sanotaan satunnaisotokseksi.

> **Määritelmä 3.1** Satunnaismuuttujien joukko $\{X_1, \ldots, X_n\}$ on *satunnaisotos* jakaumafunktiosta $F(x)$, jos satunnaismuuttujat ovat riippumattomat ja kunkin jakaumafunktio on $F(x)$. ■

Lyhyemmin sanotaan, että $\{X_1, \ldots, X_n\}$ on satunnaisotos, jos satunnaismuuttujat ovat riippumattomat ja samoin jakautuneet.

Kuten pykälässä 2.3.4 todettiin, sanotaa odotusarvoa $E(X^k)$ satunnaismuuttujan X k:nneksi *momentiksi*. Myös satunnaisotoksesta voidaan laskea momentteja; k:s *otosmomentti* on

$$M'_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Ensimmäista otosmomenttia on tapana merkitä $\overline{X}$:lla: $M'_1 = \overline{X}$. Otosmomentit M'_k ovat satunnaismuuttujia. Otosmomentin laskettua arvoa merkitään $m'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$; laskettu arvo on luku.

Ns. *momenttimenetelmä* perustuu siihen, että momenttien $E(X^k)$ lasketut lausekkeet sisältävät jakauman parametreja ja momentteja voidaan estimoida otosmomenteilla M'_k. Menetelmässä merkitään momentteja ja otosmomentteja yhtä suuriksi niin paljon, että saadaan riittävä määrä yhtälöitä, joista jakauman parametrit voidaan ratkaista. Ratkaisua pidetään parametrien *estimaattoreina*. Estimaattori on satunnaismuuttuja, koska se sisältää otosmomentteja, jotka ovat satunnaismuuttujia. Kun estimaattorin arvo lasketaan tietyille havainnoille, saadaan tietty luku, jota sanotaan *estimaatiksi*.

Jos jakaumalla on vain yksi parametri, niin momenttimenetelmässä tarvitaan vain yksi yhtälö: $E(X) = \overline{X}$; tästä yhtälöstä ratkaistaan parametri. Jos jakaumalla on kaksi parametria, niin ne ratkaistaan yhtälöistä $E(X) = \overline{X}$ ja $E(X^2) = M'_2$.

Esimerkki 3.1

3.1.2 Suurimman uskottavuuden menetelmä 1

Tarkastellaan diskreettiä satunnaismuuttujaa X, jonka todennäköisyysfunktio on $p(k) = P(X = k)$. Satunnaismuuttujasta on käytettävissä n riippumatonta havaintoa $X_1, \ldots, X_n$, joiden arvot ovat $k_1, \ldots, k_n$. Nyt $P(X_i = k_i) = p(k_i)$ ja havaintojen riippumattomuuden vuoksi $P(X_1 = k_1, \ldots, X_n = k_n) = p(k_1) \cdots p(k_n)$. Tämä lauseke on satunnaismuuttujan X jakauman parametrien $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_r\}$ tietty lauseke. Sitä sanotaan *uskottavuusfunktioksi* (likelihood function) ja merkitään $L(\boldsymbol{\theta})$:

$$L(\boldsymbol{\theta}) = p(k_1) \cdots p(k_n).$$

Suurimman uskottavuuden menetelmässä etsitään sellaiset parametrien arvot, että uskottavuusfunktio saa niillä arvoilla suurimman arvonsa. Menetelmässä siis etsitään sellaiset parametrien arvot, että todennäköisyys $P(X_1 = k_1, \ldots, X_n = k_n)$ saada juuri ne havainnot $k_1, \ldots, k_n$, jotka on saatu, on mahdollisimman suuri, ts. menetelmässä etsitään se (annettua muotoa oleva) jakauma, joka todennäköisimmin antaa ne havainnot, jotka on saatu.

Uskottavuusfunktio $L(\boldsymbol{\theta})$ maksimoituu samoilla parametrien arvoilla kuin ns. *logaritminen uskottavuusfunktio* $l(\boldsymbol{\theta}) = \ln L(\theta)$:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(k_i).$$

Koska funktion $l(\boldsymbol{\theta})$ maksimointi on usein helpompaa kuin funktion $L(\boldsymbol{\theta})$ maksimointi, niin usein käytetäänkin logaritmista uskottavuusfunktiota.

Esimerkki 3.2

Esimerkki 3.3

3.1.3 Suurimman uskottavuuden menetelmä 2

Tarkastellaan sitten jatkuvaa satunnaismuuttujaa X, jonka tiheysfunktio on $f(x)$. Satunnaismuuttujasta on käytettävissä n riippumatonta havaintoa $X_1, \ldots, X_n$, joiden arvot ovat $x_1, \ldots, x_n$. Tiheysfunktion arvojen tulo $f(x_1) \cdots f(x_n)$ on satunnaismuuttujan X jakauman parametrien $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_r\}$ tietty lauseke. Sitä sanotaan uskottavuusfunktioksi ja merkitään $L(\boldsymbol{\theta})$:

$$L(\boldsymbol{\theta}) = f(x_1) \cdots f(x_n).$$

Suurimman uskottavuuden menetelmässä etsitään sellaiset parametrien arvot, että uskottavuusfunktio saa niillä arvoilla suurimman arvonsa. Usein maksimoidaan logaritminen uskottavuusfunktio:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i).$$

Esimerkki 3.4

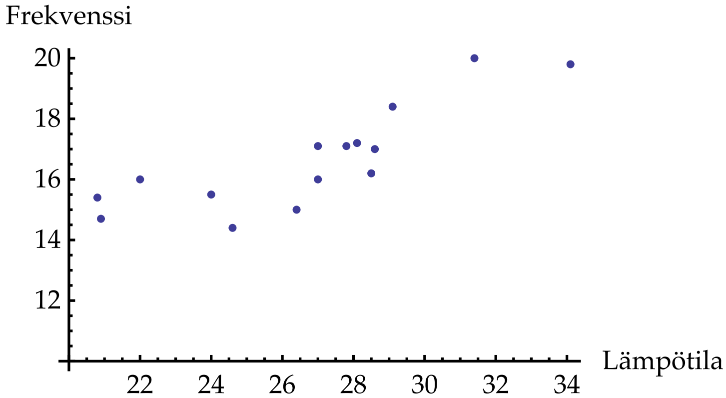
Esimerkki 3.5

3.1.4 Yksinkertainen lineaarinen regressio

Esimerkki 3.6 Kun mitattiin tietyn heinäsirkkalajin siritystä 15:ssä eri lämpötilassa, saatiin seuraavat tulokset:

Lämpötila	20.8	20.9	22.	24.	24.6	26.4	27.	27.	27.8	28.1	28.5	28.6	29.1	31.4	34.1
Frekvenssi	15.4	14.7	16.	15.5	14.4	15.	16.	17.1	17.1	17.2	16.2	17.	18.4	20.	19.8

Frekvenssi tarkoittaa värähdystä sekunnisssa. Piirretään havainnot:



Mitä korkeampi lämpötila, sen suurempi frekvenssi. Riippuvuus näyttäisi olevan likimain lineaarista. Olkoon x lämpötila ja Y frekvenssi. Etsitään muotoa $Y = a + b\,x$ oleva funktio, joka kuvaa frekvenssin riippuvuutta lämpötilasta. ∎

Kun tietty muuttuja x sai arvot $x_1, \ldots, x_n$, niin tietty toinen muuttuja Y sai arvot $y_1, \ldots, y_n$. Sanotaan, että x on ns. riippumaton muuttuja tai selittävä muuttuja ja Y ns. riippuva muuttuja tai selitettävä muuttuja. Muuttujien välille halutaan etsiä muotoa $Y = a + b\,x + \epsilon$ oleva malli. Tätä mallia sanotaan *yksikertaiseksi lineaariseksi regressiomalliksi*. Tässä ϵ kuvaa riippuvan muuttujan satunnaista vaihtelua; oletetaan, että $\epsilon \sim N(0, \sigma^2)$. Parametrit a, b ja σ^2 ovat tuntemattomia. Tavoitteena on etsiä parametreille sellaiset estimaatit, että näin saatu estimoitu malli kuvaa havaintoja mahdollisimman hyvin.

Merkitään

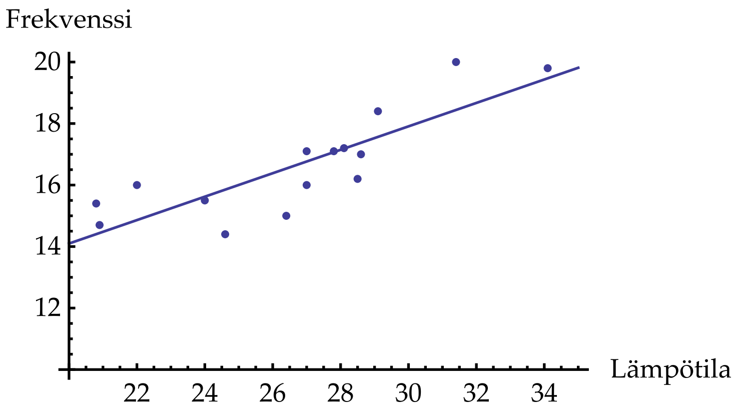
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\,y_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \hat{y}_i = \hat{a} + \hat{b}\,x_i.$$

Lause 3.1 Mallin $Y = a + b\,x + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, parametrien a, b ja σ^2 suurimman uskottavuuden estimaatit ovat

$$\hat{b} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \quad \hat{a} = \overline{y} - \hat{b}\,\overline{x}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Lauseen mukaisia suurimman uskottavuuden estimaatteja sanotaan myös *pienimmän neliösumman estimaateiksi*, koska niihin voidaan päätyä myös minimoimalla neliösumma $\sum_{i=1}^{n} (y_i - a - b\,x_i)^2$.

Esimerkki 3.7 Lasketaan regressiomallin $Y = a + b\,x + \epsilon$ suurimman uskottavuuden estimaatit edellisen esimerkin havaintoaineistolle.



3.2 Odotusarvon ja varianssin estimointi

■ 3.2.1 Otoskeskiarvo ja otosvarianssi

Olkoon $\{X_1, \ldots, X_n\}$ satunnaisotos jakaumasta, jonka odotusarvo on μ ja varianssi σ^2. Osoittautuu, että μ voidaan estimoida ns. otoskeskiarvon (eli ensimmäisen momentin) avulla ja σ^2 ns. otosvarianssin avulla.

Määritelmä 3.2 Satunnaisotoksen $\{X_1, \ldots, X_n\}$ *otoskeskiarvo* on

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ja *otosvarianssi* on

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2. \blacksquare$$

Sanotaan, että otoskeskiarvo $\overline{X}$ on odotusarvon μ estimaattori. Kun otoskeskiarvo lasketaan tietyille havainnolle x_i, saadaan $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$; tätä sanotaan odotusarvon estimaatiksi.

Otosvarianssi S^2 on varianssin σ^2 estimaattori. Kun otosvarianssi lasketaan tietyille havainnolle x_i, saadaan $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$; tätä sanotaan varianssin estimaatiksi.

Pykälässä 3.2.2 osoitetaan seuraavan lauseen avulla, että otoskeskiarvo ja otosvarianssi ovat populaation odotusarvon ja varianssin *hyviä* estimaattoreita. Määritellään k:s *keskusmomentti* $\mu_k = E\left[(X - \mu)^k\right]$. Toinen keskusmomentti on sama kuin varianssi: $\mu_2 = \sigma^2$.

Lause 3.2 Jos $\{X_1, \ldots, X_n\}$ on satunnaisotos jakaumasta, jonka odotusarvo on μ, varianssi σ^2 ja neljäs keskusmomentti μ_4, niin

a) $E(\overline{X}) = \mu$, $\quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

b) $E\left(S^2\right) = \sigma^2$, $\quad \mathrm{Var}\left(S^2\right) = \frac{\mu_4}{n} + \frac{3-n}{n(n-1)} \left(\sigma^2\right)^2$.

3.2.2 Harhattomuus ja tarkentuvuus

Parametrin estimaattori (niin kuin otoskeskiarvo tai otosvarianssi) on yleisesti jokin havaintojen X_i lauseke ja siis satunnaismuuttuja. Jos siis satunnaisotoksia otetaan useita, niin joka kerta saadaan satunnaisuuden vuoksi hiukan erilaisia tuloksia. Estimaattorilla on satunnaismuuttujana myös jokin jakauma. Voidaan sanoa, että parametrin estimaattori on sitä parempi, mitä paremmin estimaattorin jakauma keskittyy estimoitavan parametrin ympärille; tällöin nimittäin estimaattori saa suurella todennäköisyydellä arvon, joka on lähellä parametria. Estimaattorin hyvyyttä voidaan tarkastella estimaattorin odotusarvon ja varianssin avulla:

- Jos estimaattorin odotusarvo olisi parametrin arvo, niin silloin estimaattorin jakauma olisi juuri oikealla kohdalla, joten estimaattori saisi arvoja *parametrin ympäriltä*.
- Jos vielä estimaattorin varianssi olisi pieni, niin silloin estimaattorin jakauma olisi hyvin keskittynyt odotusarvon ympärille, joten estimaattori saisi suurella todennäköisyydellä arvon *läheltä parametria*.

Seuraava määritelmä antaa estimaattorille kaksi hyvää ominaisuutta.

Määritelmä 3.3 Parametrin θ estimaattori $\hat{\theta}$ on *harhaton*, jos $E(\hat{\theta}) = \theta$, ja *tarkentuva*, jos $\mathrm{Var}(\hat{\theta}) \to 0$, kun $n \to \infty$. ■

Harhattomuus takaa, että estimaattoriin ei liity systemaattista virhettä eli harhaa. Tarkentuvuus takaa, että mitä enemmän meillä on havaintoja, sen suurempi on todennäköisyys, että estimaattori saa arvon läheltä parametria; estimaattori tällä tavoin tarkentuu kohti parametria, kun otoskoko kasvaa. (Tarkentuvuus määritellään oikeastaan yleisemmin niin, että estimaattori suppenee todennäköisyysmielessä kohti parametria, mutta tällä kurssilla riittänee yllä oleva yksinkertaisempi määritelmä.)

Seuraava tulos osoittaa, että otoskeskiarvo ja otosvarianssi todellakin ovat populaation odotusarvon ja varianssin hyviä estimaattoreita.

Seuraus 3.1 *a)* Otoskeskiarvo $\overline{X}$ on populaation odotusarvon μ harhaton ja tarkentuva estimaattori.
b) Otosvarianssi S^2 on populaation varianssin σ^2 harhaton ja tarkentuva estimaattori.

Esimerkki 3.8 Tehdään n riippumatonta koetta. Kukin koe onnistuu todennäköisyydellä p. Oletetaan, että saatiin k kappaletta onnistuneita kokeita.

Seuraava lause antaa mahdollisuuden määrittää sellainen otoskoko n, että otoskeskiarvo toteuttaa annetun tarkkuusvaatimuksen $\left|\overline{X}_n - \mu\right| < a$ annetulla todennäköisyydellä p.

Lause 3.3 Jos $\{X_1, \ldots, X_n\}$ on satunnaisotos jakaumasta, jonka odotusarvo on μ ja varianssi σ^2, niin $P\left(\left|\overline{X}_n - \mu\right| < a\right) \simeq p$, jos n toteuttaa ehdon $\Phi\left(\frac{a\sqrt{n}}{\sigma}\right) = \frac{1+p}{2}$.

Esimerkki 3.9 Halutaan estimoida ihmisten pituuden odotusarvoa μ sellaisella otoskeskiarvolla, joka poikkeaa μ:stä korkeintaan a:n verran todennäköisyydellä 0.95.

3.3 Luottamusvälit

3.3.1 Odotusarvon luottamusväli 1

Olkoon $\{X_1, \ldots, X_n\}$ satunnaisotos jakaumasta, jonka odotusarvo μ ja varianssi σ^2 ovat tuntemattomia. Estimoidaan μ otoskeskiarvolla $\overline{X}$ ja varianssi σ^2 otosvarianssilla S^2. Pykälässä 3.2.2 todettiin, että $\overline{X}$ on μ:n hyvä estimaattori. Satunnaismuuttujaa $\overline{X}$ sanotaan myös μ:n *piste-estimaattoriksi*, koska tämä estimaattori on yksi luku eli piste.

Koska estimaattori $\overline{X}$ on satunnaismuuttuja, sen arvo vaihtelee otoksesta toiseen. Jos otettaisiin useita otoksia ja kustakin laskettaisiin otoskeskiarvo $\overline{x}$, saataisiin käsitys siitä, missä rajoissa $\overline{X}$ vaihtelee. Se auttaisi päättelemään, mikä voisi olla μ:n todellinen arvo. Tavallisesti on kuitenkin käytettävissä vain yksi otos. Osoittautuu, että senkin avulla saadaan piste-estimaatin lisäksi myös informaatiota siitä, millä välillä todellinen μ sijaitsee suurella todennäköisyydellä.

Seuraavassa onkin tavoitteena etsiä sellainen väli (c, d), että μ on tällä välillä todennäköisyydellä α. Tällaista väliä sanotaan μ:n 100 α prosentin *luottamusväliksi* (confidence interval). Luottamusvälin laskemista sanotaan myös *väliestimoinniksi* erotukseksi piste-estimoinnista. Yleisimmin lasketaan 95:n, 99:n ja 99.9:n prosentin luottamusvälejä. Seuraava lause antaa likimääräisen 95 %:n luottamusvälin, mutta samaan tapaan saadaan muita välejä. Jos esimerkiksi halutaan laskea 99:n tai 99.9:n prosentin luottamusväli, niin lauseen kaavassa oleva luku 1.96 vain korvataan luvulla 2.58 tai 3.29.

Lause 3.4 Odotusarvon μ likimääräinen 95 prosentin luottamusväli suurilla n:n arvoilla on

$$P\left(\overline{X} - 1.96\,\frac{S}{\sqrt{n}} < \mu < \overline{X} + 1.96\,\frac{S}{\sqrt{n}}\right) \simeq 0.95.$$

Esimerkki 3.10 Oletetaan, että kun saatiin 100 havaintoa, niin $\overline{x} = 54.07$ ja $s^2 = 83.40$.

3.3.2 Odotusarvon luottamusväli 2

Lauseessa 3.4 jouduttiin σ korvaamaan estimaatillaan s, ja tällöin kaavan tarkkuus heikkenee. Tarkempi luottamusväli saadaan ns. t-jakauman avulla seuraavasti.

Lause 3.5 Jos otos on normaalijakaumasta, niin μ:n 100 $(1-\alpha)$ prosentin luottamusväli on

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

Tässä $t_{n-1;\,\alpha/2}$ on $t(n-1)$-jakauman 100 $\frac{\alpha}{2}$ prosentin *kriittinen arvo* eli sellainen arvo, että $t(n-1)$-jakautunut satunnaismuuttuja saa sitä suuremman arvon todennäköisyydellä $\frac{\alpha}{2}$. Tällaisia kriittisiä arvoja on taulukoitu eri n:n ja α:n arvoille; yksi taulukko on todennäköisyyslaskennan kaavakokoelmassa. Taulukon mukaan on esimerkiksi $t_{20,\,0.10} = 1.325$, mikä siis tarkoittaa, että jos $T \sim t(20)$, niin $P(T > 1.325) = 0.10$. Jos $T \sim t(n)$, niin sanotaan myös, että T:llä on t-jakauma n:llä *vapausasteella* (degrees of freedom, lyh. df). t-jakauman tiheysfunktio on symmetrinen origon suhteen.

Tässä lauseessa oletettiin, että havainnot noudattavat normaalijakaumaa. Jos jakauma ei ole normaali, niin luottamusvälin kaava pätee likimäärin ja sitä paremmin, mitä suurempi on n.

Voidaan myös osoittaa, että $t(n)$-jakauma lähestyy standardoitua normaalijakaumaa, kun n kasvaa, joten jos n on suuri, ei ole juuri merkitystä, käytetäänkö lauseen 3.5 kaavaa vai lauseen 3.4 kaavaa, joka perustuu normaalijakaumaan.

Esimerkki 3.11 Oletetaan esimerkin 3.10 tapaan, että $\overline{x} = 54.07$ ja $s^2 = 83.40$, mutta oletetaan nyt, että havaintoja on vain 20.

3.4 Hypoteesien testaus

3.4.1 Odotusarvon testaus

Esimerkki 3.12 Ohjelman suoritusajalla on ollut normaalijakauma odotusarvolla 50. Ohjelmaa yritettiin parantaa, ja sen jälkeen havaittiin, että sadan testiajon keskiarvo oli $\overline{x} = 44$ ja varianssi $s^2 = 900$. Voidaanko katsoa, että ohjelman suoritusajan odotusarvo on edelleen 50, vai onko odotusarvo todella pienentynyt?

Testiajojen keskiarvo 44 oli pienempi kuin 50, mutta otoskeskiarvo on satunnaismuuttuja, joka vaihtelee otoksesta toiseen. Tehtyihin sataan ajoon onkin voinut sattumalta tulla mukaan paljon poikkeuksellisen pieniä aikoja. Vanha odotusarvo 50 voi edelleen olla lähellä todellista arvoa. Halutaan tutkia, onko uusi otoskeskiarvo 44 *tilastollisesti merkitsevästi* pienempi kuin vanha odotusarvo 50.

Merkitään $\mu_0 = 50$ ja tehdään seuraavat hypoteesit:

H_0: $\mu = 50$

H_1: $\mu < 50$

H_0 on ns. *nollahypoteesi*, jonka mukaan suoritusajan odotusarvossa ei ole tapahtunut muutosta. H_1 on ns. *vastahypoteesi*, jonka mukaan suoritusajan odotusarvo ei enää olekaan μ_0 vaan jokin sitä pienempi arvo. Oletetaan, että nollahypoteesi on voimassa, ja tutkitaan, mitä hypoteesista seuraa. Jos havainnot näyttävät olevan ristiriidassa näiden seurausten kanssa, niin tämä antaisi aiheen olettaa, että nollahypoteesi ei olekaan voimassa.

Jos nollahypoteesi on voimassa, niin statistikalla

$$T = \frac{\overline{X} - \mu_0}{S \big/ \sqrt{n}}$$

on $t(n-1)$-jakauma. Tiedetään siis, että T on esimerkiksi vain todennäköisyydellä $\alpha = 0.05$ pienempi kuin $-t_{n-1;\,\alpha} = -t_{99;\,0.05} \simeq -t_{100;\,0.05} = -1.660$. Lasketaan T:n arvo:

$$t = \frac{\overline{x} - \mu_0}{s \big/ \sqrt{n}} = \frac{44 - 50}{30 \big/ \sqrt{100}} = -2.0.$$

Tämä on siis pienempi kuin -1.660. Näin pienen (-2.0) T:n arvon saaminen on harvinaista (todennäköisyys on 0.05), jos nollahypoteesi on voimassa. Jos siis tehtäisiin useita sadan ajon testejä ja jokaisesta testistä laskettaisiin T:n arvo, niin vain noin 5 % lasketuista T:n arvoista olisi pienempiä kuin -1.660, jos nollahypoteesi on voimassa. Kun tehtiin yksi sadan ajon testi, niin *voidaan luottaa siihen (varmuus on 95 %), että jos nollahypoteesi on voimassa, niin testiajoista laskettu T:n arvo ei ole pienempi kuin* -1.660. Laskettu T:n arvo oli kuitenkin pienempi kuin -1.660. Näin ollen voidaan suurella varmuudella sanoa, että nollahypoteesi ei ole voimassa. Nollahypoteesi voidaan hylätä *merkitsevyystasolla* 0.05. Vastahypoteesi on todennäköisempi. Näin ollen ohjelman suoritusaika näyttää pienentyneen tilastollisesti merkitsevästi. Uusi suoritusajan odotusarvo on lähellä arvoa $\bar{x} = 44$. ■

Esimerkki 3.13 Ohjelman suoritusajalla on ollut normaalijakauma odotusarvolla 50. Ohjelmaan tehtiin muutoksia, ja sen jälkeen havaittiin, että sadan testiajon keskiarvo oli $\bar{x} = 54$ ja varianssi $s^2 = 900$. Voidaanko katsoa, että ohjelman suoritusajan odotusarvo on edelleen 50, vai onko odotusarvo todella suurentunut?

Tehdään hypoteesit

H_0: $\mu = 50$

H_1: $\mu > 50$

Jos nollahypoteesi on voimassa, niin T-statistika on esimerkiksi vain todennäköisyydellä $\alpha = 0.05$ suurempi kuin $t_{n-1;\,\alpha} = t_{99;\,0.05} \simeq t_{100;\,0.05} = 1.660$. Lasketaan T:n arvo:

$$t = \frac{\bar{x} - \mu_0}{s \big/ \sqrt{n}} = \frac{54 - 50}{30 \big/ \sqrt{100}} = 1.33.$$

Tämä on siis pienempi kuin 1.660. Näin pienen (1.33) T:n arvon saaminen on hyvinkin mahdollista (todennäköisyys on 0.95), jos nollahypoteesi on voimassa. Jos siis tehtäisiin useita sadan ajon testejä ja jokaisesta testistä laskettaisiin T:n arvo, niin noin 95 % lasketuista T:n arvoista olisi pienempiä kuin 1.660, jos nollahypoteesi on voimassa. Näin ollen havainnot eivät näytä olevan ristiriidassa nollahypoteesin kanssa, joten nollahypoteesia ei voida hylätä merkitsevyystasolla 0.05: ohjelman suoritusajan odotusarvo on edelleen 50. ■

Esimerkki 3.14 Ohjelman suoritusajalla on ollut normaalijakauma odotusarvolla 50. Ohjelmaan tehtiin muutoksia, ja sen jälkeen havaittiin, että sadan testiajon keskiarvo oli $\bar{x} = 54$ ja varianssi $s^2 = 900$. Voidaanko katsoa, että ohjelman suoritusajan odotusarvo on edelleen 50, vai onko odotusarvo muuttunut?

Tehdään hypoteesit

H_0: $\mu = 50$

H_1: $\mu \neq 50$

Jos nollahypoteesi on voimassa, niin T-statistika on esimerkiksi vain todennäköisyydellä $\alpha = 0.05$ välin $(-t_{n-1;\,\alpha/2},\ t_{n-1;\,\alpha/2}) = (-t_{99;\,0.025},\ t_{99;\,0.025}) \simeq (-1.984,\ 1.984)$ ulkopuolella. Lasketaan T:n arvo:

$$t = \frac{\bar{x} - \mu_0}{s \big/ \sqrt{n}} = \frac{54 - 50}{30 \big/ \sqrt{100}} = 1.33.$$

Tämä arvo kuuluu väliin $(-1.984, 1.984)$. T:n arvon 1.33 saaminen on hyvinkin mahdollista (todennäköisyys on 0.95), jos nollahypoteesi on voimassa. Jos siis tehtäisiin useita sadan ajon testejä ja jokaisesta testistä laskettaisiin T:n arvo, niin noin 95 % lasketuista T:n arvoista olisi välillä $(-1.984, 1.984)$, jos nollahypoteesi on voimassa. Näin ollen havainnot eivät näytä olevan ristiriidassa nollahypoteesin kanssa, joten nollahypoteesia ei voida hylätä merkitsevyystasolla 0.05: ohjelman suoritusajan odotusarvo on edelleen 50. ■

Edelliset esimerkit perustelevat seuraavan lauseen.

Lause 3.6 Olkoon $\{X_1, \ldots, X_n\}$ otos $N(\mu, \sigma^2)$-jakaumasta, H_0: $\mu = \mu_0$ ja $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.

a) Jos H_1: $\mu < \mu_0$, niin H_0 voidaan hylätä merkitsevyystasolla α, jos $t < -t_{n-1;\alpha}$.

b) Jos H_1: $\mu > \mu_0$, niin H_0 voidaan hylätä merkitsevyystasolla α, jos $t > t_{n-1;\alpha}$.

c) Jos H_1: $\mu \neq \mu_0$, niin H_0 voidaan hylätä merkitsevyystasolla α, jos $t < -t_{n-1;\alpha/2}$ tai $t > t_{n-1;\alpha/2}$.

Jos H_1: $\mu < \mu_0$ tai H_1: $\mu > \mu_0$, niin puhutaan *yksipuolisesta* testistä. Jos H_1: $\mu \neq \mu_0$, niin puhutaan *kaksipuolisesta* testistä.

■ 3.4.2 Testauksen virheistä

Tilastolliseen testaukseen liittyy aina epävarmuutta. Ehdottoman varmaa johtopäätöstä ei voida tehdä. Joskus siis tulee väistämättä tehdyksi virheellinen päätös. Seuraavassa taulukossa on lyhyt yhteenveto:

		Hypoteesi on	
		oikea	väärä
Hypoteesi	hyväksytään	*Oikea päätös*	*Väärä päätös* **Tyypin II virhe**
	hylätään	*Väärä päätös* **Tyypin I virhe**	*Oikea päätös*

Oikea päätös tehdään seuraavissa tapauksissa:

- Nollahypoteesi on oikea eikä nollahypoteesia hylätä.
- Nollahypoteesi on väärä ja nollahypoteesi hylätään.

Väärä päätös tehdään seuraavissa tapauksissa:

- Nollahypoteesi on oikea mutta nollahypoteesi hylätään. Tämä on *tyypin I virhe*.
- Nollahypoteesi on väärä ja mutta nollahypoteesia ei hylätä. Tämä on *tyypin II virhe*.

Jos testauksessa käytetään merkitsevyystasoa α, niin tyypin I virhe tehdään todennäköisyydellä α. Jos esimerkiksi H_0: $\mu = \mu_0$ ja H_1: $\mu < \mu_0$, niin H_0 hylätään, jos $t < -t_{n-1;\alpha}$. Jos H_0 on tosi, niin kuitenkin todennäköisyydellä α on t pienempi kuin $-t_{n-1;\alpha}$, joten todennäköisyydellä α tulee oikea nollahypoteesi hylätyksi.

Tyypin II virheen todennäköisyyttä merkitään usein β:lla, ja se voidaan laskea kullekin testille erikseen. Lukua $1-\beta$ sanotaan testin voimakkuudeksi. Jos esimerkiksi H_0: $\mu = \mu_0$ ja H_1: $\mu < \mu_0$, niin H_0 hylätään, jos $t < -t_{n-1;\alpha}$. Jos H_0 on epätosi, niin t voi silti olla suurempi kuin $-t_{n-1;\alpha}$, joten tällaisessa tapauksessa väärä nollahypoteesi ei tule hylätyksi.

Kummankin virheen todennäköisyys olisi hyvä saada niin pieneksi kuin mahdollista. On kuitenkin niin, että jos tyypin I virheen todennäköisyyttä pienennetään valitsemalla pienempi α, niin tyypin II virheen todennäköisyys kasvaa. Yleisesti käytetäänkin α:n arvoa 0.05. Se on riittävän pieni, jotta tyypin I virheen todennäköisyys on hyväksyttävän pieni, mutta ei kuitenkaan liian pieni, niin että tyypin II virheen todennäköisyyskin pysyy riittävän pienenä. Voidaan käyttää myös α:n arvoja 0.01 ja 0.001, jos halutaan olla erityisen varmoja siitä, että oikeaa nollahypoteesia ei hylätä.

Selvästi on niin, että kun havaintojen määrä n kasvaa, niin kummankin virhetyypin todennäköisyys pienenee. Saadaanhan suuremmalla havaintomäärällä tarkempaa ja luotettavampaa informaatiota populaatiosta.

Todennäköisyyslaskennan kaavoja

Unionin todennäköisyys

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P\left(A_i \cap A_j\right) + - \ldots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right)$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} P\left(\bigcap_{j=1}^{i} A_j\right),$$ jos leikkaukset keskenään yhtä todennäköiset

$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$, jos A_i : t toisensa poissulkevat

$P(\bigcup_{i=1}^{n} A_i) = 1 - \prod_{i=1}^{n} P(A_i{}^c)$, jos A_i : t riippumattomat

Leikkauksen todennäköisyys

$P(A \cap B) = P(A)\, P(B \mid A) = P(A \mid B)\, P(B)$

$P(\bigcap_{i=1}^{n} A_i) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \ldots \cap A_{n-1})$

$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$, jos A_i : t riippumattomat

Ehdollinen todennäköisyys

$P(A \mid B) = P(A \cap B) \,/\, P(B)$

$P(B) = \sum_{i=1}^{n} P(B \mid A_i)\, P(A_i)$, jos A_i : t muodostavat partition

$P(A_k \mid B) = P(B \mid A_k)\, P(A_k) \,/\, P(B)$

Diskreetit satunnaismuuttujat

$F(x) = P(X \le x),\ p(k) = P(X = k),\ \sum_{k \in K} p(k) = 1$

$E(X) = \sum_{k \in K} k\, p(k),\ E[g(X)] = \sum_{k \in K} g(k)\, p(k)$

$X \ge 0 : E(X) = \sum_{k=0}^{\infty} [1 - F(k)],\ E[X(X-1)] = 2 \sum_{k=1}^{\infty} k[1 - F(k)]$

$\mathrm{Var}(X) = E\left[(X - \mu)^2\right] = E\left(X^2\right) - \mu^2 = E[X(X-1)] + \mu - \mu^2$

$G(z) = E\left(z^X\right),\ p(k) = \frac{1}{k!} G^{(k)}(0),\ E(X) = G'(1),\ E[X(X-1)] = G''(1)$

Jatkuvat satunnaismuuttujat

$F(x) = P(X \le x),\ P\,(X \in A) = \int_A f(x)\, dx$

$F(x) = \int_{-\infty}^{x} f(t)\, dt,\ f(x) = F'(x),\ \int_{-\infty}^{\infty} f(x)\, dx = 1$

$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx,\ E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$

$X \ge 0 : E(X) = \int_0^{\infty} [1 - F(x)]\, dx,\ E\left(X^2\right) = 2 \int_0^{\infty} x[1 - F(x)]\, dx$

$M(t) = E\left(e^{tX}\right),\ E\left(X^j\right) = M^{(j)}(0)$

Yhteisjakaumat

$F_{X,Y}(x, y) = P(X \le x,\ Y \le y),\ P\,[(X, Y) \in A] = \iint_{(x,y) \in A} f_{X,Y}\,(x, y)\, dx dy$

$F_{X,Y}(x, y) = \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{X,Y}(s, t)\, dt\right] ds,\ f_{X,Y}(x, y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x, y)$

$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy,\ f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$

Yleiset muunnokset

$Y = g(X) : f_Y(y) = f_X\left[g^{-1}(y)\right] \left| \frac{d}{dy} g^{-1}(y) \right|$

$$\begin{cases} Z = g\,(X, Y) \\ V = h\,(X, Y) \end{cases}, \begin{cases} X = r\,(Z, V) \\ Y = s\,(Z, V) \end{cases} : \ f_{Z,V}(z, v) = f_{X,Y}[r(z, v), s(z, v)]\, \left| J(z, v) \right|$$

Erityiset muunnokset

$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(z-v, v)\, dv$, $f_{X-Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(z+v, v)\, dv$

$f_{XY}(z) = \int_{-\infty}^{\infty} f_{X,Y}\left(\frac{z}{v}, v\right) \frac{1}{|v|}\, dv$, $f_{\frac{X}{Y}}(z) = \int_{-\infty}^{\infty} f_{X,Y}(z\,v, v)\,|v|\, dv$

$Y = \min\{X_1, \ldots, X_n\} : F_Y(y) = 1 - \prod_{i=1}^{n} [1 - F_{X_i}(y)]$

$Z = \max\{X_1, \ldots, X_n\} : F_Z(z) = \prod_{i=1}^{n} F_{X_i}(z)$

Summat

$\mathrm{Cov}(X, Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(X\,Y) - \mu_X\,\mu_Y$

$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$, $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j)$

$X = \sum_{i=1}^{n} I_{A_i} : E(X) = n\,P(A_i)$, $\mathrm{Var}(X) = n\,P(A_i)\,P(A_i{}^c) + n(n-1)\left[P(A_i \cap A_j) - P(A_i)\,P(A_j)\right]$

$P(|X-\mu| \geq t) \leq \sigma^2 \big/ t^2$

Ehdolliset jakaumat

$p_{X|Y=l}(k) = p_{X,Y}(k, l) / p_Y(l)$, $E(X \mid Y=l) = \sum_{k \in K} k\, p_{X|Y=l}(k)$

$p_X(k) = \sum_{l \in L} p_{X|Y=l}(k)\, p_Y(l)$, $p_{Y|X=k}(l) = p_{X|Y=l}(k)\, p_Y(l) / p_X(k)$

$f_{X|Y=y}(x) = f_{X,Y}(x, y) / f_Y(y)$, $E(X \mid Y=y) = \int_{-\infty}^{\infty} x\, f_{X|Y=y}(x)\, dx$

$P(a \leq X \leq b \mid Y=y) = \int_a^b f_{X|Y=y}(x)\, dx$

$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y=y}(x)\, f_Y(y)\, dy$, $f_{Y|X=x}(y) = f_{X|Y=y}(x)\, f_Y(y) / f_X(x)$

$p_X(k) = \int_{-\infty}^{\infty} p_{X|Y=y}(k)\, f_Y(y)\, dy$, $f_{Y|X=k}(y) = p_{X|Y=y}(k)\, f_Y(y) / p_X(k)$

$f_X(x) = \sum_{l \in L} f_{X|Y=l}(x)\, p_Y(l)$, $p_{Y|X=x}(l) = f_{X|Y=l}(x)\, p_Y(l) / f_X(x)$

Kokonaisodotusarvokaava

$$E(X) = E[E(X \mid Y)] = \begin{cases} \sum_{l \in L} E(X \mid Y=l)\, p_Y(l), & Y \text{ diskreetti} \\ \int_{-\infty}^{\infty} E(X \mid Y=y)\, f_Y(y)\, dy, & Y \text{ jatkuva} \end{cases}$$

$\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[E(X \mid Y)]$

$$P(A) = E[P(A \mid X)] = \begin{cases} \sum_{k \in K} P(A \mid X=k)\, p_X(k), & X \text{ diskreetti} \\ \int_{-\infty}^{\infty} P(A \mid X=x)\, f_X(x)\, dx, & X \text{ jatkuva} \end{cases}$$

$f_{X|A}(x) = P(A \mid X=x)\, f_X(x) / P(A)$

$$E(X \mid A) = \begin{cases} \sum_k k\, P(X=k \mid A), & X \text{ diskreetti} \\ \int_{\infty}^{\infty} x\, f_{X|A}(x)\, dx, & X \text{ jatkuva} \end{cases}$$

$E(X) = \sum_{i=1}^{n} E(X \mid A_i)\, P(A_i)$

Satunnaisen pituinen summa

$G_{S_N}(z) = G_N[G_X(z)]$, $M_{S_N}(t) = G_N[M_X(t)]$

$E(S_N) = E(N)\, E(X)$, $\mathrm{Var}(S_N) = E(N)\, \mathrm{Var}(X) + \mathrm{Var}(N)\, [E(X)]^2$

Korrelaatio

$\mathrm{Cor}(X, Y) = \mathrm{Cov}(X, Y) \big/ \sqrt{\mathrm{Var}(X)\, \mathrm{Var}(Y)}$

$l(Y) = \mu_X + \rho\, \frac{\sigma_X}{\sigma_Y}\, (Y - \mu_Y)$

Syntymis–kuolemis– prosessi

$$\rho_0 = 1,\ \rho_i = \frac{\lambda_0 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} : \pi_0 = \left(\sum_{i=0}^{\infty} \rho_i\right)^{-1},\ \pi_i = \rho_i\, \pi_0$$

Diskreettejä jakaumia, äärellinen arvojoukko

DU(n): $\frac{1}{n}$, $k = 1, \ldots, n$, $E(X) = \frac{n+1}{2}$, $V(X) = \frac{n^2-1}{12}$, $G(z) = \frac{1}{n} \sum_{k=1}^{n} z^k = \frac{z(1-z^n)}{n(1-z)}$

Ber(p): $p^k q^{1-k}$, $k = 0, 1$, $E(X) = p$, $V(X) = p\,q$, $G(z) = p\,z + q$

Bin(n, p): $\binom{n}{k} p^k q^{n-k}$, $k = 0, \ldots, n$, $\begin{cases} E(X) = n\,p \\ V(X) = n\,p\,q \end{cases}$, $G(z) = (p\,z + q)^n$, $X \approx \mathrm{Po}(n\,p)$, $X \sim \mathrm{AsN}(n\,p, n\,p\,q)$

Hyp(N, M, n): $\binom{M}{k}\binom{N-M}{n-k} \Big/ \binom{N}{n}$, $\begin{matrix} 0 \le k \le M \\ 0 \le n-k \le N-M \end{matrix}$, $\begin{cases} E(X) = n\,p, \; p = M/n \\ V(X) = n\,p\,q\,\frac{N-n}{N-1} \end{cases}$, $X \approx \mathrm{Bin}\left(n, \frac{M}{N}\right)$

Multinomi(n, p_1, …, p_r): $P(X_1 = k_1, \ldots, X_r = k_r) = \frac{n!}{k_1! \cdots k_r!} {p_1}^{k_1} \cdots {p_r}^{k_r}$, $\sum_{i=1}^{r} p_i = 1$, $\sum_{i=1}^{r} k_i = n$

Multihyp(N, M_1, …, M_r, n): $P(X_1 = k_1, \ldots, X_r = k_r) = \binom{M_1}{k_1} \cdots \binom{M_r}{k_r} \Big/ \binom{N}{n}$, $\sum_{i=1}^{r} M_i = N$, $\sum_{i=1}^{r} k_i = n$

Diskreettejä jakaumia, ääretön arvojoukko

Po(λ): $e^{-\lambda} \frac{\lambda^k}{k!}$, $k \ge 0$, $E(X) = \lambda$, $V(X) = \lambda$, $G(z) = e^{\lambda(z-1)}$, $X \sim \mathrm{AsN}(\lambda, \lambda)$

Geom(p): $q^{k-1} p$, $k \ge 1$, $F(k) = 1 - q^k$, $E(X) = \frac{1}{p}$, $V(X) = \frac{q}{p^2}$, $G(z) = \frac{p\,z}{1-q\,z}$

ModGeom(p): $q^k p$, $k \ge 0$, $F(k) = 1 - q^{k+1}$, $E(X) = \frac{q}{p}$, $V(X) = \frac{q}{p^2}$, $G(z) = \frac{p}{1-q\,z}$

Negbin(n, p): $\binom{k-1}{n-1} p^n q^{k-n}$, $k \ge n$, $E(X) = \frac{n}{p}$, $V(X) = \frac{n\,q}{p^2}$, $G(z) = \left(\frac{p\,z}{1-q\,z}\right)^n$

ModNegbin(n, p): $\binom{k+n-1}{n-1} p^n q^k$, $k \ge 0$, $E(X) = \frac{n\,q}{p}$, $V(X) = \frac{n\,q}{p^2}$, $G(z) = \left(\frac{p}{1-q\,z}\right)^n$

Normaalijakauma

$N(0, 1)$: $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt$, $E(X) = 0$, $V(X) = 1$, $M(t) = e^{\frac{1}{2}t^2}$

$N(\mu, \sigma^2)$: $f(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right)$, $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, $E(X) = \mu$, $V(X) = \sigma^2$, $M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$P(\mu - [1.96, 2.58, 3.29]\,\sigma < X < \mu + [1.96, 2.58, 3.29]\,\sigma) = [0.95, 0.99, 0.999]$

$S_n \sim \mathrm{AsN}(n\,\mu, n\,\sigma^2)$, $\overline{X}_n \sim \mathrm{AsN}\left(\mu, \frac{\sigma^2}{n}\right)$

Muita yleisimpiä jatkuvia jakaumia

$U(a, b)$: $\frac{1}{b-a}$, $a < x < b$, $F(x) = \frac{x-a}{b-a}$, $E(X) = \frac{a+b}{2}$, $V(X) = \frac{(b-a)^2}{12}$, $M(t) = \frac{e^{b\,t} - e^{a\,t}}{(b-a)\,t}$

Exp(λ): $\lambda\, e^{-\lambda x}$, $x > 0$, $F(x) = 1 - e^{-\lambda x}$, $E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$, $M(t) = \frac{\lambda}{\lambda - t}$

Gamma(α, λ): $\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x > 0$, $E(X) = \frac{\alpha}{\lambda}$, $V(X) = \frac{\alpha}{\lambda^2}$, $M(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha$

Erlang(n, λ): $\frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}$, $x > 0$, $F(x) = 1 - e^{-\lambda x} \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!}$, $E(X) = \frac{n}{\lambda}$, $V(X) = \frac{n}{\lambda^2}$, $M(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$

Beta(α, β): $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$, $0 < x < 1$, $E(X) = \frac{\alpha}{\alpha+\beta}$, $V(X) = \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}$

Muita jatkuvia jakaumia, $x > 0$ ellei toisin mainita

LogN$\left(\mu, \sigma^2\right)$: $\frac{1}{\sigma x}\,\phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$, $F(x) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$, $E(X) = e^{\mu+\frac{1}{2}\sigma^2}$, $E\left(X^2\right) = e^{2\mu+2\sigma^2}$

Weibull(α, λ): $\alpha\,\lambda\,x^{\alpha-1}\,e^{-\lambda x^\alpha}$, $F(x) = 1 - e^{-\lambda x^\alpha}$, $E(X) = \left(\frac{1}{\lambda}\right)^{\frac{1}{\alpha}}\,\Gamma\left(1+\frac{1}{\alpha}\right)$, $E\left(X^2\right) = \left(\frac{1}{\lambda}\right)^{\frac{2}{\alpha}}\,\Gamma\left(1+\frac{2}{\alpha}\right)$

Hypoexp$(\vec{\lambda})$: $\sum_{i=1}^{n}\alpha_i\,\lambda_i\,e^{-\lambda_i x}$, $\lambda_i \neq \lambda_j$, $\alpha_i = \prod_{j=i,j\neq i}^{n}\frac{\lambda_j}{\lambda_j-\lambda_i}$, $F(x) = \sum_{i=1}^{n}\alpha_i\left(1-e^{-\lambda_i x}\right)$, $E(X) = \sum_{i=1}^{n}\frac{1}{\lambda_i}$, $V(X) = \sum_{i=1}^{n}\frac{1}{\lambda_i^2}$

Hyperexp$\left(\vec{\lambda}, \vec{\alpha}\right)$: $\sum_{i=1}^{n}\alpha_i\,\lambda_i\,e^{-\lambda_i x}$, $\sum_{i=1}^{n}\alpha_i = 1$, $F(x) = \sum_{i=1}^{n}\alpha_i\left(1-e^{-\lambda_i x}\right)$, $E(X) = \sum_{i=1}^{n}\frac{\alpha_i}{\lambda_i}$, $E\left(X^2\right) = 2\sum_{i=1}^{n}\frac{\alpha_i}{\lambda_i^2}$

Pareto(α, λ): $\frac{\lambda}{\alpha}\left(\frac{\alpha}{x}\right)^{\lambda+1}$, $x > \alpha$, $F(x) = 1 - \left(\frac{\alpha}{x}\right)^{\lambda}$, $E(X) = \frac{\alpha\,\lambda}{\lambda-1}$ $(\lambda > 1)$, $V(X) = \frac{\lambda\,\alpha^2}{(\lambda-1)^2\,(\lambda-2)}$ $(\lambda > 2)$

Muita jatkuvia jakaumia, $x \in \mathbb{R}$

Laplace(μ, λ): $\frac{1}{2}\,\lambda\,e^{-\lambda|x-\mu|}$, $F(x) = \begin{cases}\frac{1}{2}\,e^{-\lambda(\mu-x)} & x < \mu\\ 1-\frac{1}{2}\,e^{-\lambda(x-\mu)} & x \geq \mu\end{cases}$, $E(X) = \mu$, $V(X) = \frac{2}{\lambda^2}$, $M(t) = \frac{\lambda^2\,e^{\mu t}}{\lambda^2-t^2}$

Logistic$\left(\mu, \sigma^2\right)$: $\frac{a\,e^{-a y}}{\sigma(1+e^{-a y})^2}$, $y = \frac{x-\mu}{\sigma}$, $a = \frac{\pi}{\sqrt{3}}$, $F(x) = \frac{1}{1+e^{-a y}}$, $E(X) = \mu$, $V(X) = \sigma^2$

Cauchy(a, b): $\frac{1}{\pi b}\left[1+\left(\frac{x-a}{b}\right)^2\right]^{-1}$, $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-a}{b}\right)$

$N\left(\mu_1, \mu_2, \sigma_1{}^2, \sigma_2{}^2, \rho\right)$: $E(X \mid Y = y) = \mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}\,(y-\mu_2)$, $V(X \mid Y = y) = \sigma_1{}^2\left(1-\rho^2\right)$

Tilastollisia jakaumia

$t(n)$: $\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$, $x \in \mathbb{R}$, $E(X) = 0$ $(n \geq 2)$, $V(X) = \frac{n}{n-2}$ $(n \geq 3)$

$\chi^2(n)$: $\left[2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)\right]^{-1}\,x^{\frac{n}{2}-1}\,e^{-\frac{x}{2}}$, $x > 0$, $E(X) = n$, $V(X) = 2\,n$, $M(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$

$F(m, n)$: $\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\sqrt{\frac{m^m\,n^n\,x^{m-2}}{(n+m\,x)^{m+n}}}$, $x > 0$, $E(X) = \frac{n}{n-2}$ $(n \geq 3)$, $V(X) = \frac{2\,n^2(m+n-2)}{m(n-2)^2\,(n-4)}$ $(n \geq 5)$

Luottamusvälit

$P\left(\overline{X} - 1.96\,\frac{S}{\sqrt{n}} < \mu < \overline{X} + 1.96\,\frac{S}{\sqrt{n}}\right) \simeq 0.95$

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

$P\left(\frac{(n-1)\,S^2}{\chi^2_{n-1;\,\alpha/2}} < \sigma^2 < \frac{(n-1)\,S^2}{\chi^2_{n-1;\,1-\alpha/2}}\right) = 1-\alpha$

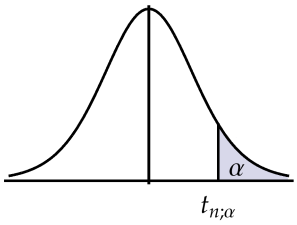
$P\left(\hat{p} - 1.96\sqrt{\hat{p}\,\hat{q}\,/\,n} < p < \hat{p} + 1.96\sqrt{\hat{p}\,\hat{q}\,/\,n}\right) \simeq 0.95$

Hypoteesien testaus

$H_0 : \mu = \mu_0 : T = \left(\overline{X} - \mu_0\right)\big/\left(S\,/\sqrt{n}\right) \sim t(n-1)$

$H_0 : p = p_0 : Z = \left(\hat{p} - p_0\right)\big/\sqrt{p_0\,q_0\,/\,n} \sim N(0, 1)$

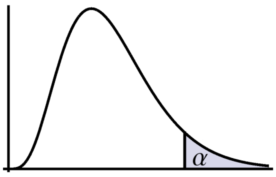
$H_0 : P(X = x_i) = p_i : Q = \sum_{i=1}^{k}\frac{(N_i - n\,p_i)^2}{n\,p_i} \sim \chi^2(k-m-1)$



$t(n)$ **–jakauman kriittisiä arvoja** $t_{n;\alpha}$ **:** $P\left(X > t_{n;\alpha}\right) = \alpha$

Esim. $t_{10;\,0.05} = 1.812$: jos $X \sim t(10)$, niin $P(X > 1.812) = 0.05$

	α				
n	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	3.078	6.314	12.71	31.82	63.66
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
150	1.287	1.655	1.976	2.351	2.609
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576



$\chi^2_{n;\alpha}$

$\chi^2(n)$ **–jakauman kriittisiä arvoja** $\chi^2_{n;\alpha}$ **:** $P\left(X > \chi^2_{n;\alpha}\right) = \alpha$

Esim. $\chi^2_{10;\,0.05} = 3.940$: jos $X \sim \chi^2(10)$, niin $P(X > 3.940) = 0.05$.

	α							
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.95**	**0.975**	**0.99**	**0.995**
1	0.00003927	0.0001571	0.0009821	0.003932	3.841	5.024	6.635	7.879
2	0.01003	0.02010	0.05064	0.1026	5.991	7.378	9.210	10.60
3	0.07172	0.1148	0.2158	0.3518	7.815	9.348	11.34	12.84
4	0.2070	0.2971	0.4844	0.7107	9.488	11.14	13.28	14.86
5	0.4117	0.5543	0.8312	1.145	11.07	12.83	15.09	16.75
6	0.6757	0.8721	1.237	1.635	12.59	14.45	16.81	18.55
7	0.9893	1.239	1.690	2.167	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	31.41	34.17	37.57	40.00
21	8.034	8.897	10.28	11.59	32.67	35.48	38.93	41.40
22	8.643	9.542	10.98	12.34	33.92	36.78	40.29	42.80
23	9.260	10.20	11.69	13.09	35.17	38.08	41.64	44.18
24	9.886	10.86	12.40	13.85	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	43.77	46.98	50.89	53.67
31	14.46	15.66	17.54	19.28	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	47.40	50.73	54.78	57.65
34	16.50	17.79	19.81	21.66	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	49.80	53.20	57.34	60.27
36	17.89	19.23	21.34	23.27	51.00	54.44	58.62	61.58
37	18.59	19.96	22.11	24.07	52.19	55.67	59.89	62.88
38	19.29	20.69	22.88	24.88	53.38	56.90	61.16	64.18
39	20.00	21.43	23.65	25.70	54.57	58.12	62.43	65.48
40	20.71	22.16	24.43	26.51	55.76	59.34	63.69	66.77

$N(0, 1)$-jakauman todennäköisyyksiä $P(X \leq x) = \alpha$

Esim. $P(X \leq 1.23) = 0.8907$. $\Phi(-x) = 1 - \Phi(x)$.

x	0	1	2	3	4	5	6	7	8	9	
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	**0.0**
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	**0.1**
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	**0.2**
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	**0.3**
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	**0.4**
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	**0.5**
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	**0.6**
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	**0.7**
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	**0.8**
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	**0.9**
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	**1.0**
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	**1.1**
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	**1.2**
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177	**1.3**
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319	**1.4**
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441	**1.5**
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545	**1.6**
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633	**1.7**
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706	**1.8**
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767	**1.9**
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817	**2.0**
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857	**2.1**
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890	**2.2**
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916	**2.3**
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936	**2.4**
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952	**2.5**
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	**2.6**
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974	**2.7**
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981	**2.8**
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986	**2.9**
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990	**3.0**
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993	**3.1**
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	**3.2**
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	**3.3**
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	**3.4**
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	**3.5**
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	**3.6**
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	**3.7**
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	**3.8**
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	**3.9**
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	