MAB09
Tilastolliset ja todennäköisyysjakaumat

Iitin lukio
2023-2024

Aleksi Alenius

Sisältö

1. Tilastollinen jakauma
- Tilastoaineisto
- Frekvenssijakauma
- Tunnuslukuja
- Jakaumien kuvailu ja vertailu

2. Diskreetti todennäköisyysjakauma
- Kertausta
- Toistokoe
- Binomijakauma

3. Jatkuva todennäköisyysjakauma
- Normaalijakauma
- Normitettu normaalijakauma
- Sovelluksia

4. Tilastollinen päättely
- Keskiarvon keskivirhe ja luottamusväli
- Suhteellisen osuuden virhemarginaali ja luottamusväli

Tilastollinen jakauma

Tilastoaineisto

Havaintoaineisto

Määritelmä

- Perusjoukko on tutkimuksen kohteena oleva ryhmä.
- Havainto (tai havaintoarvo) on yksittäinen vastaus.

Esimerkki

Tutkitaan luokan oppilaiden pituuksia ja vaatekokoja. Tällöin perusjoukkona on luokka. Havaintoina ovat mittaus- ja kyselytulokset.

Tilastomuuttuja

Määritelmä

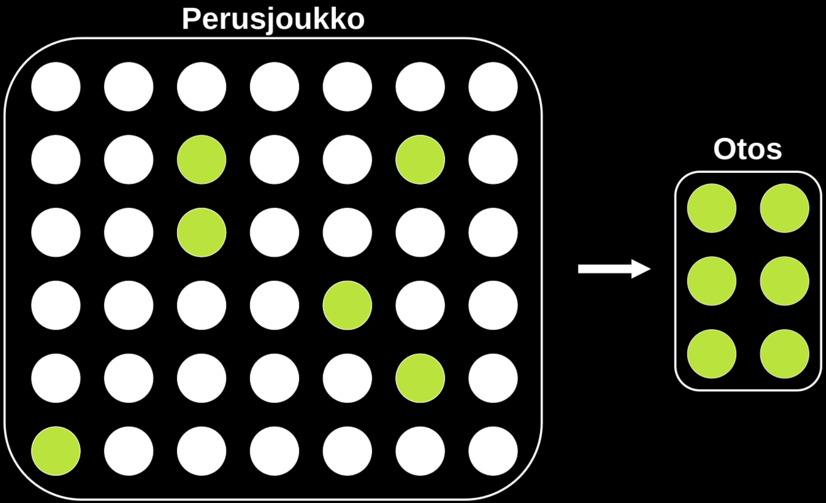
Tilastomuuttuja on tutkittava ominaisuus, joka voi olla:

- Kvantitatiivinen (luku)
  - erilliset arvot $\Longrightarrow$ diskreetti
  - jatkuva
- Kvalitatiivinen (laatu)

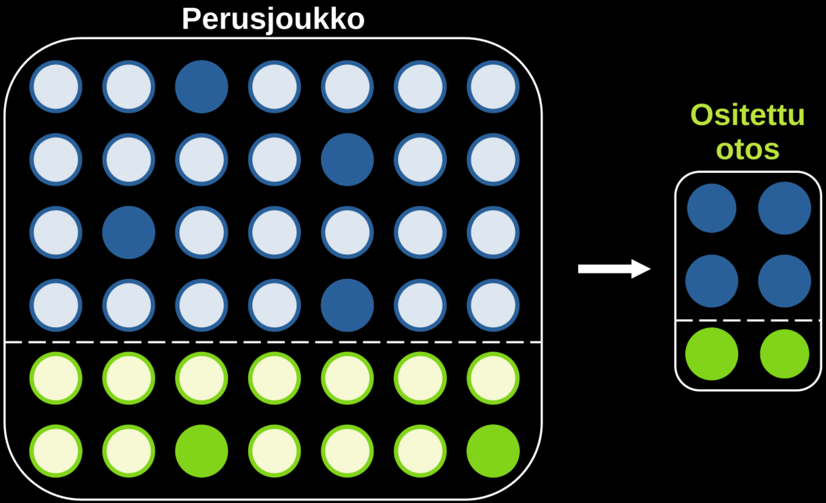
Esimerkki

Pituus on jatkuva kvantitatiivinen muuttuja, kun taas vaatekoko on diskreetti. Esimerkiksi vaatteen merkki on kvalitatiivinen muuttuja.

Perusotos



Ositettu otos



Frekvenssijakauma

Frekvenssi

Määritelmä (Frekvenssi)

Frekvenssi eli esiintymistiheys on havainnon esiintymiskertojen lukumäärä.

Esimerkki

20 opiskelijan luokalta kysytään heidän sisarustensa lukumäärää. Tulokset kootaan alla olevaan taulukkoon.

Sisarusta	f
0	4
1	7
2	8
3	1

Suhteellinen rekvenssi

Määritelmä (Suhteellinen frekvenssi)

Suhteellinen frekvenssi $f\%$ kertoo havainnon esiintymiskertojen %-osuuden kaikista tapahtumista.

Esimerkki

20 opiskelijan luokalta kysytään heidän sisarustensa lukumäärää. Tulokset kootaan alla olevaan taulukkoon.

Sisarusta	f	$f\%$
0	4	20%
1	7	35%
2	8	40%
3	1	5%

Summafrekvenssi

Määritelmä (Summafrekvenssi)

Summafrekvenssi eli kumulatiivinen esiintymistiheys on yksittäisten havaintojen yhteenlaskettu arvo.

Esimerkki

20 opiskelijan luokalta kysytään heidän sisarustensa lukumäärää. Tulokset kootaan alla olevaan taulukkoon.

Sisarusta	f	$f\%$	sf
0	4	20%	4
1	7	35%	11
2	8	40%	19
3	1	5%	20

Suhteellinen summafrekvenssi

Määritelmä (Suhteellinen summafrekvenssi)

Suhteellinen summafrekvenssi on vastaava summafrekvenssi prosentteina.

Esimerkki

20 opiskelijan luokalta kysytään heidän sisarustensa lukumäärää. Tulokset kootaan alla olevaan taulukkoon.

Sisarusta	f	$f\%$	sf	$sf\%$
0	4	20%	4	20%
1	7	35%	11	55%
2	8	40%	19	95%
3	1	5%	20	100%

Moodi ja mediaani

Määritelmä (Moodi)

Moodi Mo on aineistossa eniten esiintynyt havainto.

Määritelmä (Mediaani)

Mediaani Md on suuruusjärjestykseen lajitellun jakauman keskimmäinen luku, joka jakaa aineiston kahteen yhtä suureen osaan.

Esimerkki

Esimerkin taulukon perusteella sisarustutkimuksen moodi ja mediaani ovat

$$Mo = 2$$
$$Md = 1.$$

Sisarusta	f	$f\%$	sf	$sf\%$
0	4	20%	4	20%
1	7	35%	11	55%
2	8	40%	19	95%
3	1	5%	20	100%

Jatkuvan tilastomuuttujan jakauma

Kaikkia tilastomuuttujia ei ole mahdollista järjestää helposti luettavaan taulukkoon.
$\Longrightarrow$ luokittelu:

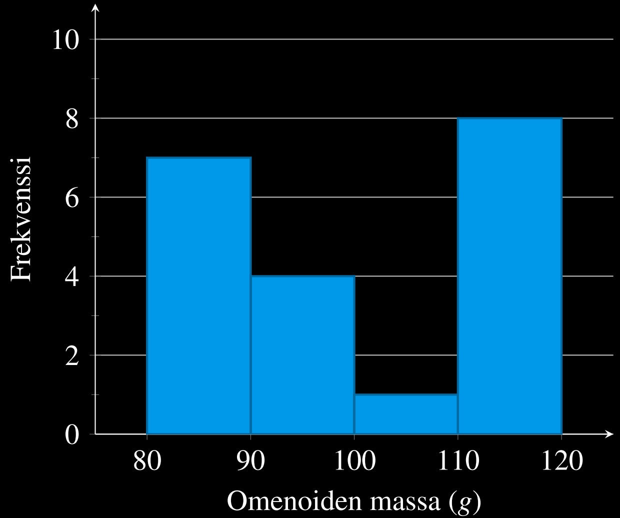
1. Vaihteluväli
2. Luokkavälit
3. Luokat

Esimerkki (Massan tutkiminen)

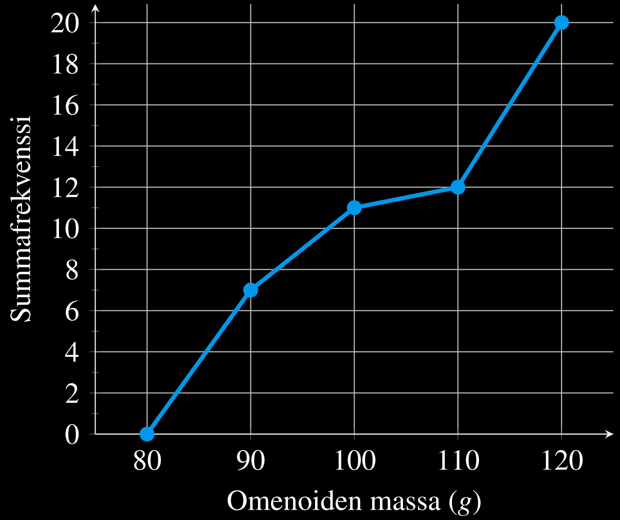
Tutkitaan omenien massoja, jotka vaihtelevat välillä 80-119 g.

Massa (g)	Luokkakeskus (g)	f	$f\%$	sf	$sf\%$
80-89	84,5	7	35	7	35
90-99	94,5	4	20	11	55
100-109	104,5	1	5	12	60
109-119	114,5	8	40	20	100

Histogrammi



Summakäyrä


Summafrekvenssi
0
2
4
6
8
10
12
14
16
18
20
80
90
100
110
120
Omenoiden massa (g)

Tunnuslukuja

Keskiluvut

Keskiluvut kuvaavat muuttujien keskimääräisiä suuruuksia. Niihin lukeutuu moodin ja mediaanin lisäksi myös keskiarvo.

Määritelmä (keskiarvo)

Keskiarvo $\bar{x}$ on havaintoarvojen summan $\sum x$ ja lukumäärän n osamäärä eli:

$$\bar{x} = \frac{\sum x}{n}.$$

Esimerkki

Lukujen $1, 1, 5, 6$ ja 7 keskiarvo on:

$$\frac{1+1+5+6+7}{5} = \frac{20}{5} = 4.$$

Keskihajonta

Hajontaluvut kuvaavat muuttujien vaihtelua.

Määritelmä (Keskihajonta)

Keskihajonta s_n (myös σ) kertoo, kuinka kaukana havainnot keskimäärin ovat keskiarvosta.

$$s_n = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

Esimerkki

Lukujen $1, 1, 5, 6$ ja 7 keskiarvo $\bar{x} = 4$. Niiden keskihajonta on siis:

$$\begin{aligned} s_5 &= \sqrt{\frac{(1-4)^2+(1-4)^2+(5-4)^2+(6-4)^2+(7-4)^2}{5}} \\ &= \sqrt{\frac{9+9+1+4+9}{5}} \\ &= \sqrt{\frac{32}{5}} \approx 2,52\ldots \end{aligned}$$

Otoskeskihajonta

Otoskeskihajonta eroaa vain vähän keskihajonnasta.

Määritelmä (Otoskeskihajonta)

Otoskeskihajonta s_{n-1} määritellään seuraavasti:

$$s_{n-1} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Huomio

Keskihajonnan ja otoskeskihajonnan lausekkeiden ainoana erona ovat niiden nimittäjät.

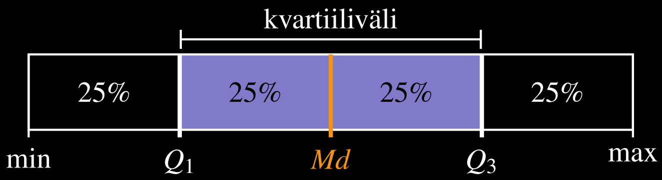
Esimerkki

Lukujen $1, 1, 5, 6$ ja 7 otoskeskihajonta on siis:

$$s_{5-1} = \sqrt{\frac{32}{5-1}} \approx 2,82...$$

Muita tunnuslukuja

Mediaani ($Md = Q_2$) jakaa havaintoaineiston kahteen yhtä suureen osaan. Aineisto voidaan jakaa myös neljänneksiin alakvarteelin Q_1 ja yläkvarteelin Q_3 avulla.



Luvut min, Q_1, Md, Q_3 ja max muodostavat viiden luvun yhteenvedon. Vertaa suhteelliseen summafrekvenssiin $sf\%$.

Jakaumien kuvailu ja vertailu

Hajontaluvut

Hajontaluvut kuvaavat muuttujan arvoissa esiintyvää vaihtelua. Niistä käytetyimmät ovat vaihteluväli ja keskihajonta s.

Muistetaan:

- vaihteluvälin määrittävät muuttujan x mahdolliset arvot
- keskihajonta:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}},$$

  missä $\bar{x}$ on muuttujien x keskiarvo.

Jos arvo poikkeaa yli kahden keskihajonnan verran keskiarvosta, poikkeamaa kutsuttaan merkittäväksi.

Normitettu arvo

Muuttujan poikkeamaa keskiarvosta voidaan tutkia normitetun arvon avulla.

Määritelmä (Normitettu arvo)

Normitetun arvon Z määritelmä on:

$$Z = \frac{x - \bar{x}}{s},$$

missä x on muuttujan arvo, $\bar{x}$ keskiarvo ja s keskihajonta.

Esimerkki

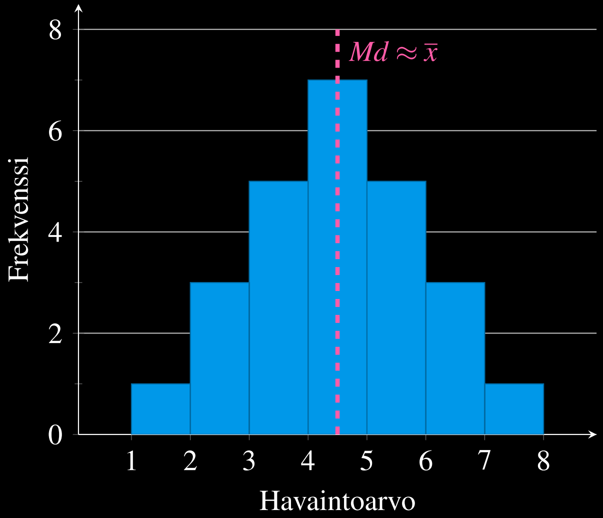
Olkoot $s = 9$ cm ja $\bar{x} = 170$ cm. Normitetaan pituudet 161 cm ja 192 cm:

$$Z_{161} = \frac{161 - 170}{9} = \frac{-9}{9} = -1$$
$$Z_{192} = \frac{192 - 170}{9} = \frac{22}{9} = 2,44\ldots \approx 2,4$$

Tämän perusteella voidaan todeta, että arvon 191 cm poikkeama on merkittävä, mutta arvon 161 cm ei.

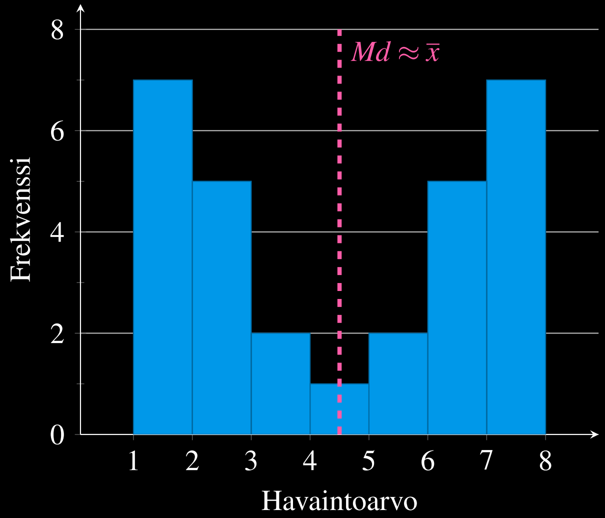
Symmetrinen jakauma

Jos jakauma on symmetrinen, niin keskihajonta ja mediaani ovat likimäärin samat.



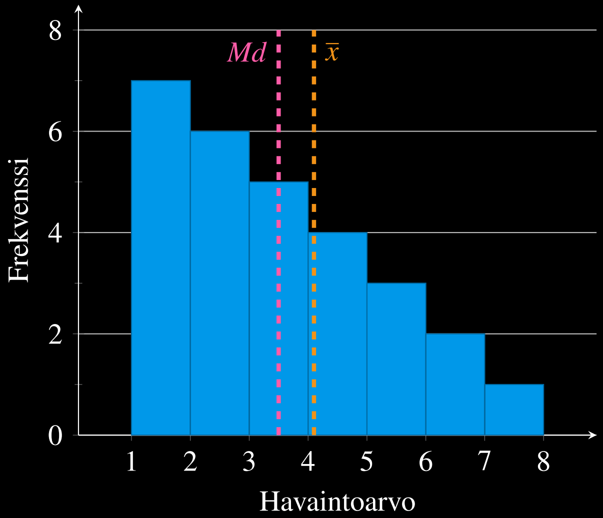
Symmetrinen jakauma

Symmetrisellä jakaumalla voi olla useampi huippu.



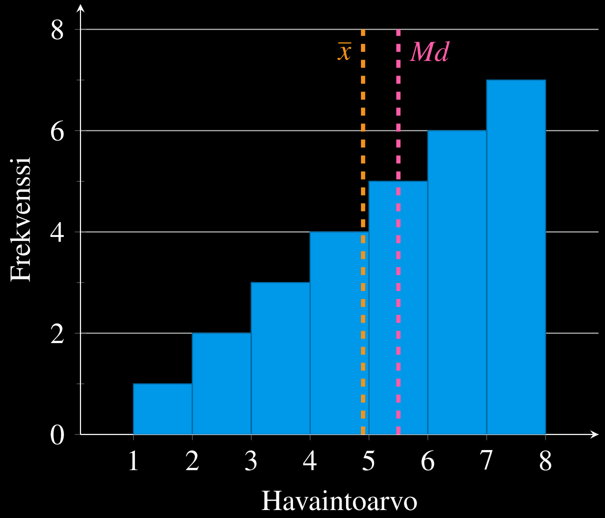
Vino jakauma

Jos keskiarvo ja mediaani poikkeavat toisistaan, tuloksena on vino jakauma. Kun $Md < \bar{x}$, niin arvot keskittyvät vasemmalle.



Vino jakauma

Vastaavasti, kun $Md > \bar{x}$, niin arvot keskittyvät oikealle.



Poikkeavat havainnot

Joskus havaintoaineistossa ilmenee selvästi muista arvoista poikkeavia havaintoja. Erityisesti keskiarvo on herkkä tällaisille poikkeuksille.

Esimerkki

Luokan arvosanat on koottu oheiseen taulukkoon. Voidaan laskea, että

$$\bar{x} = 7{,}73\ldots$$
$$Md = 8$$

Jos taulukkoon lisätään yksi arvosana 4, niin moodi ei muutu, mutta keskiarvoksi saadaan $\bar{x} = 7{,}38\ldots$

arvosana	f
4	0
5	0
6	2
7	6
8	7
9	3
10	1

Diskreetti todennäköisyysjakauma

Kertausta

Todennäköisyys

Määritelmä

Tapahtuman A todennäköisyydelle $P(A)$ pätee:

- $0 \leq P(A) \leq 1$
- $$P(A) = \frac{\text{suotuisat tapaukset}}{\text{kaikki tapaukset}}$$

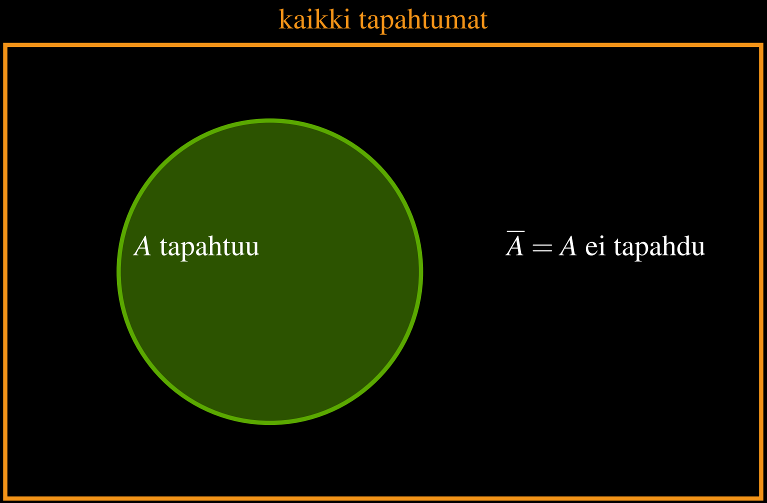
Esimerkki

Heitetään noppaa. Todennäköisyys saada tulokseksi tasan 2 on

$$P(2) = \frac{1}{6},$$

koska vain yhdessä nopan kuudesta sivusta on kaksi silmää.

Vastatapahtuma



Vastatapahtuma

Määritelmä

Tapahtuman A vastatapahtuma eli komplementtitapahtuma $\overline{A}$ on tilanne, jossa tapahtuma A ei toteudu.

Seuraus

Määritelmän perusteella

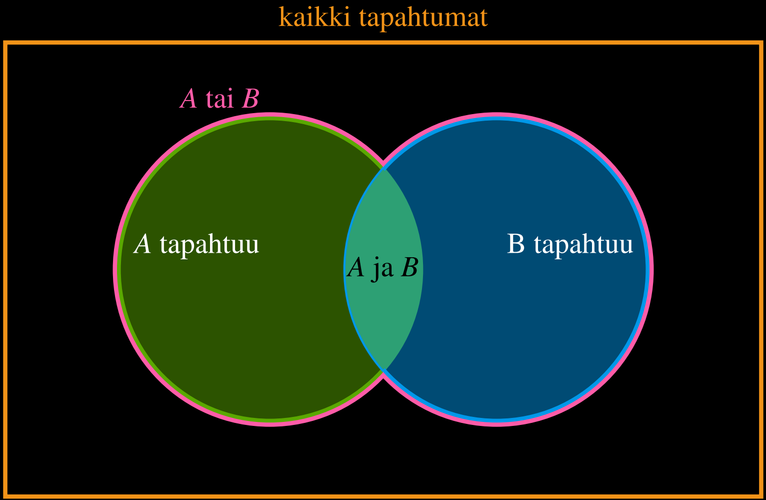
$$P(A) = 1 - P(\overline{A}).$$

Esimerkki

Koska todennäköisyys saada silmäluku 2 noppaa heitettäessä on $P(2) = \frac{1}{6}$, niin

$$P(\text{ei } 2) = 1 - P(2) = \frac{5}{6}.$$

Joukko-oppia



"Tai"

Määritelmä

Olkoot A ja B tapahtumia. Kun sanotaan "A tai B tapahtuu", niin jokin seuraavista tilanteista on tosi:

- A tapahtuu
- B tapahtuu
- A ja B tapahtuvat

Seuraus

Todennäköisyys, että A tai B tapahtuu on

$$P(A \text{ tai } B) = P(A) + P(B) - P(A \text{ ja } B)$$

Erilliset tapahtumat

Määritelmä

Tapahtumat A ja B ovat erillisiä, jos ne eivät voi tapahtua yhtä aikaa.

Seuraus

Tällöin

$$P(A \text{ tai } B) = P(A) + P(B),$$

koska $P(A \text{ ja } B) = 0.$

Esimerkki

Noppaa heitettäessä ei voi saada yhtä aikaa tulokseksi sekä kahta että yhtä. Tällöin

$$P(1 \text{ tai } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Riippumattomuus

Määritelmä

Tapahtumia A ja B kutsutaan riippumattomiksi, jos yhden tapahtuminen ei vaikuta toisen tapahtumisen todennäköisyyteen. Tällöin ne toteuttavat yhtälön

$$P(A \text{ ja } B) = P(A) \cdot P(B)$$

Esimerkki

Heitetään kahta noppaa yhtä aikaa. Selvästi eri noppien silmälukemat ovat toisistaan riippumattomat. Lisäksi $P(6) = \frac{1}{6}$, joten todennäköisyys heittää kaksi kuutosta on

$$P(6 \text{ ja } 6) = P(6) \cdot P(6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

Toistokoe

Toistokoe

Määritelmä

Toistokoe on tilanne, jossa aiemmista tuloksistaan riippumaton tapahtuma toistetaan useita kertoja.

Esimerkki

- kolikon heitto
- nopan heitto

Nopan heitto

Suoritetaan toistokoe heittämällä noppaa. Selvitetään, millä todennäköisyydella saadaan tasan 2 kertaa silmäluvuksi 6, kun noppaa heitetään 3 kertaa.
Koska todennäköisyys $P(6) = \frac{1}{6}$, niin merkitään, että todennäköisyys onnistua $P(\text{onnistuu}) = P(6) = \frac{1}{6}$. Nyt

$$P(\text{ei onnistu}) = 1 - P(\text{onnistuu}) = 1 - \frac{1}{6} = \frac{5}{6}.$$

Koska nopan heitot ovat toisistaan riippumattomia, niin

$$\begin{aligned} P(\text{onnistuu, onnistuu, ei onnistu}) &= P(\text{onnistuu}) \cdot P(\text{onnistuu}) \cdot P(\text{ei onnistu}) \\ &= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{216} \end{aligned}$$

Tapahtumat

Nopan heitto voi onnistua (O) tai epäonnistua (E) korkeintaan 3 kertaa. Tarkastellaan kaikkia mahdollisia tilanteita:

$$\begin{array}{cccc} O,O,O & O,O,E & O,E,O & O,E,E \\ E,O,O & E,O,E & E,E,O & E,E,E \end{array}$$

Tapahtumia, joissa silmäluvuksi saadaan 6 tasan 2 kertaa on yhteensä 3. Jokaisen tällaisen tapauksen todennäköisyys on $\frac{5}{216}$ ja ne ovat myös erillisiä, joten:

$$P(\text{onnistuu, onnistuu, ei onnistu}) = 3 \cdot \frac{5}{216} = \frac{15}{216} = \frac{5}{72}$$

Binomikerroin

Tapahtumia ei kannata luetella, jos niitä on liian monta. Mieluisten tapahtumien lukumäärä voidaan laskea binomikertoimella.

Määritelmä

Olkoot n ja k luonnollisia lukuja siten, että $n \geq k$. Tällöin niiden binomikerroin on:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

Binomikerroin kertoo, kuinka monta k alkion joukkoa voidaan muodostaa n alkiosta.

Huom.

Binomikertoimen kaavaa ei tällä kurssilla tarvitse käyttää tai muistaa ulkoa. Laskimen komento "nCr"toimii seuraavasti:

$$nCr = \binom{n}{r} \text{ eli esim. } 3C2 = \binom{3}{2} = 3$$

Binomitodennäköisyys

Määritelmä

Toistokokeen tuloksen todennäköisyys on

$$P(n \text{ toistoa ja , } k \text{ onnistumista}) = \binom{n}{k} \cdot p^k \cdot q^{n-k},$$

missä

- n on toistojen lukumäärä
- k on onnistumisten lukumäärä
- $n - k$ on epäonnistumisten lukumäärä
- p on onnistumisen todennäköisyys
- $q = 1 - p$ on epäonnistumisen todennäköisyys

Binomijakauma

Pistetodennäköisyys

Heitetään kolikkoa 5 kertaa. Tavoitteena on saada silmäluvuksi 5 tai 6. Onnistuneiden heittojen lukumäärää voidaan kuvata satunnaismuuttujalla X seuraavasti:

X	**Pistetodennäköisyys**
0	$P(X=0) = \binom{5}{0} \cdot \left(\frac{1}{3}\right)^0 \cdot \left(\frac{2}{3}\right)^5 \approx 0,132$
1	$P(X=1) \approx 0,329$
2	$P(X=2) \approx 0,329$
3	$P(X=3) \approx 0,165$
4	$P(X=4) \approx 0,041$
5	$P(X=5) \approx 0,004$

Pistetodennäköisyystaulukosta voidaan lukea, että 5 tai 6 onnistutaan heittämään tasan 3 kertaa noin $16,5\%$ todennäköisyydellä.

Kertymätodennäköisyys

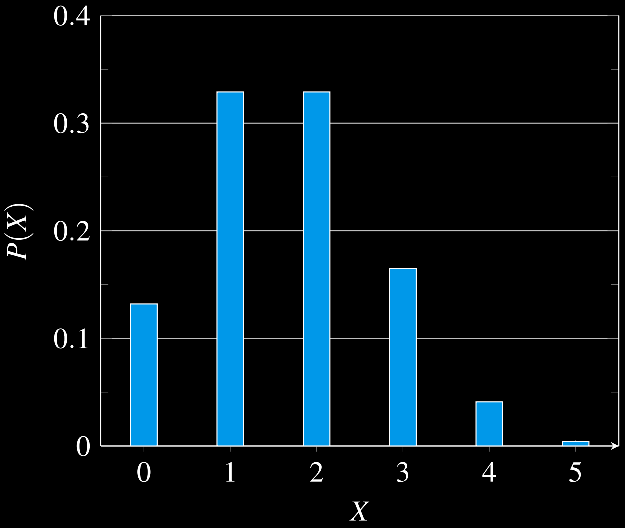
Joskus on kiinnostavaa selvittää, millä todennäköisyydellä satunnaismuuttujan X arvo on suurempi tai pienempi kuin jokin raja. Tätä voidaan tutkia kertymätodennäköisyyden eli kertymän avulla.

X	**Kertymätodennäköisyys**
0	$P(X \leq 0) = P(X = 0) \approx 0,132$
1	$P(X \leq 1) = P(X = 0) + P(X = 1) \approx 0,461$
2	$P(X \leq 2) \approx 0,79$
3	$P(X \leq 3) \approx 0,955$
4	$P(X \leq 4) \approx 0,996$
5	$P(X \leq 5) \approx 1$

Kertymätaulukon mukaan todennäköisyys saada silmäluvuksi 5 tai 6 <u>korkeintaan</u> kolme kertaa on noin $95,5\%$.

Todennäköisyyden kuvaaminen

Pistetodennäköisyydet muistuttavat ominaisuuksiltaan suhteellista frekvenssiä, joten niistä voidaan muodostaa pylväsdiagrammi.



Binomijakauma

Määritelmä

Toistokokeen, jonka tulokset on jaettu kahtia, satunnaismuuttujaa X kutsutaan binomijakautuneeksi. Sen merkintä on

$$X \sim \mathrm{Bin}(n, p),$$

missä n on toistojen lukumäärä ja p on onnistumisen todennäköisyys.

Esimerkki

Nopanheittotilanteessamme $X \sim \mathrm{Bin}(5, \frac{1}{3})$.

Määritelmä (Odotusarvo)

Satunnaismuuttujan X odotusarvo on

$$E(X) = p_1 \cdot x_1 + p_2 \cdot x_2 + \cdots + p_n \cdot x_n = n \cdot p$$

jossa muuttujat x_i ovat satunnaismuuttujan X mahdollisia arvoja ja p_i niitä vastaavat todennäköisyydet.

Odotusarvo

X	P
0	$0,132$
1	$0,329$
2	$0,329$
3	$0,165$
4	$0,041$
5	$0,004$

Satunnaismuuttujan X odotusarvo $E(X)$ on rinnastettavissa keskiarvoon. Tässä tilanteessa

$$\begin{aligned} E(X) &= 0,132 \cdot 0 + 0,329 \cdot 1 + 0,329 \cdot 2 + 0,165 \cdot 3 + 0,041 \cdot 4 + 0,004 \cdot 5 \\ &= 5 \cdot \frac{1}{3} \\ &= 1,666 \\ &\approx 1,7 \end{aligned}$$

Voidaan siis odottaa, että 5 nopan heiton aikana saadaan 5 tai 6 noin $1,7$ kertaa.

Keskihajonta

Määritelmä

Binomijakautuneen satunnaismuuttujan X keskihajonta $D(X)$ (tai σ) on

$$D(X) = \sqrt{n \cdot p \cdot (1-p)}$$

Esimerkki

Nopanheittotilanteessamme

$$D(X) = \sqrt{5 \cdot \frac{1}{3} \cdot \frac{2}{3}} = \sqrt{\frac{10}{9}} = 1,054...$$

Jatkuva todennäköisyysjakauma

Normaalijakauma

Jatkuva satunnaismuuttuja

Määritelmä

Jatkuva satunnaismuuttuja voi saada minkä tahansa reaalilukuarvon tietyltä väliltä.

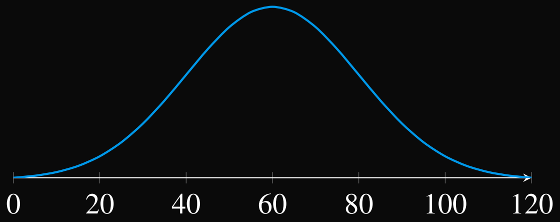
Esimerkki

- pituus
- paino
- nopeus
- tilavuus
- lämpötila jne.

Tiheysfunktio

Määritelmä

Jatkuvan satunnaismuuttujan todennäköisyysjakauma voidaan esittää tiheysfunktion avulla, jossa vaaka-akselilla ovat satunnaismuuttujan mahdolliset arvot ja pystyakselilla niiden todennäköisyyden tiheys. Pystyakselia ei useimmiten näytetä.



Normaalijakauma

Määritelmä

Normaalijakauma on todennäköisyysjakauma, jonka määrittelee funktio:

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

missä X on satunnaismuuttuja, μ sen odotusarvo ja σ sen keskihajonta.
Jos satunnaismuuttuja X noudattaa normaalijakaumaa, niin sen sanotaan olevan normaalijakautunut ja merkitään

$$X \sim N(\mu, \sigma).$$

Huom.

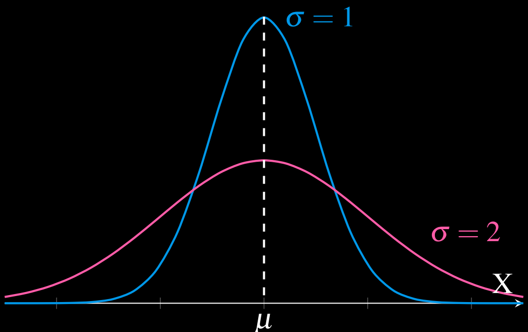
Kurssilla ei tarvitse muistaa normaalijakauman lauseketta, mutta sen perusominaisuudet on tunnettava.

Normaalijakauman ominaisuuksia

Seuraus

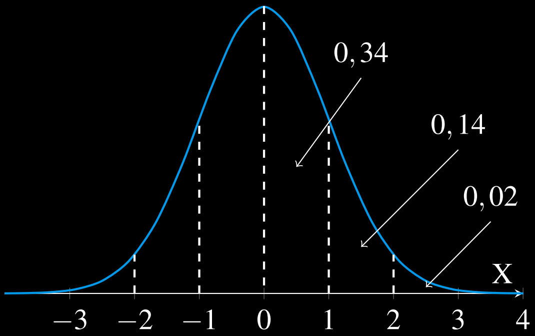
Normaalijakauman kuvaajalla on seuraavat ominaisuudet:

- *kuvaajan huippu on odotusarvon* μ *kohdalla*
- *kuvaaja on symmetrinen*
- *kuvaajan korkeus ja leveys riippuvat keskihajonnasta* σ
- *kuvaajan käyrän rajaaman alueen pinta-ala on* 100%



Kuvaajan soveltaminen

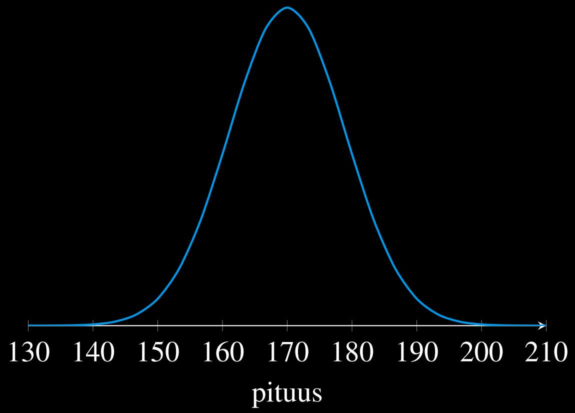
Tutkitaan normaalijakauman kuvaajaa, jossa odotusarvona on 0 ja keskihajontana 1.



- $P(X \geq 0) = 50\%$
- $P(X \geq 1) = 0,14 + 0,02 = 0,16 = 16\%$
- $P(X \leq -2) = 0.02 = 2\%$

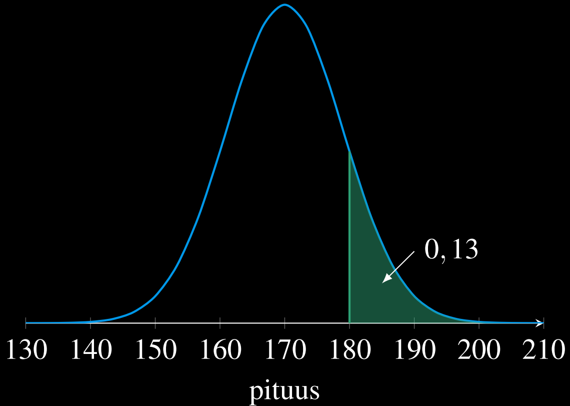
Esimerkki

Havaintoaineistoon on koottu tietoja otosryhmän pituuksista. Pituuksien odotusarvo on 170 cm ja niiden keskihajonta 9 cm. Oletetaan, että pituudet ovat normaalijakautuneet eli pituus $\sim N(170, 9)$, jolloin voimme piirtää alla olevan kuvaajan



Esimerkki

Millä todennäköisyydella sattumanvarainen otosryhmän jäsen on 180 cm pitkä? Selvitetään $P(X \geq 180)$ käyrän rajaaman pinta-alan avulla.



Koska pinta-ala käyrän alla kohdan $X = 180$ oikealla puolella on $0,13$, niin $P(X \geq 180) = 0,13 = 13\%$.

Normitettu normaalijakauma

Kertausta

Määritelmä (Normitettu arvo)

Normitetun arvon Z määritelmä on:

$$Z = \frac{x - \bar{x}}{s},$$

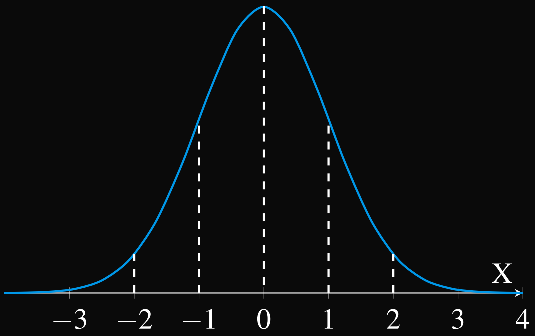
missä x on muuttujan arvo, $\bar{x}$ keskiarvo ja s keskihajonta.

Satunnaismuuttujan normitettu arvo kertoo, kuinka monen keskihajonnan päässä sen arvo on keski-/odotusarvosta.

Normitettu normaalijakauma

Määritelmä

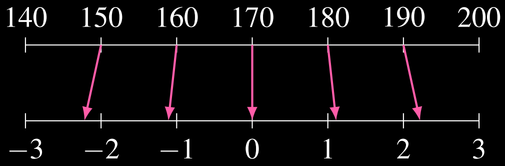
Normitettu normaalijakauma on normaalijaukama, jonka odotusarvo on 0 ja keskihajonta 1 eli $N(0,1)$.



Normaalijakauman normittaminen

Aiemmin olimme tutkineet normaalijakautuneita pituuksia, joiden odotusarvo $\mu = 170$ cm ja keskihajonta $\sigma = 9$ cm. Normitetaan näiden pituuksien arvot, jolloin esimerkiksi:

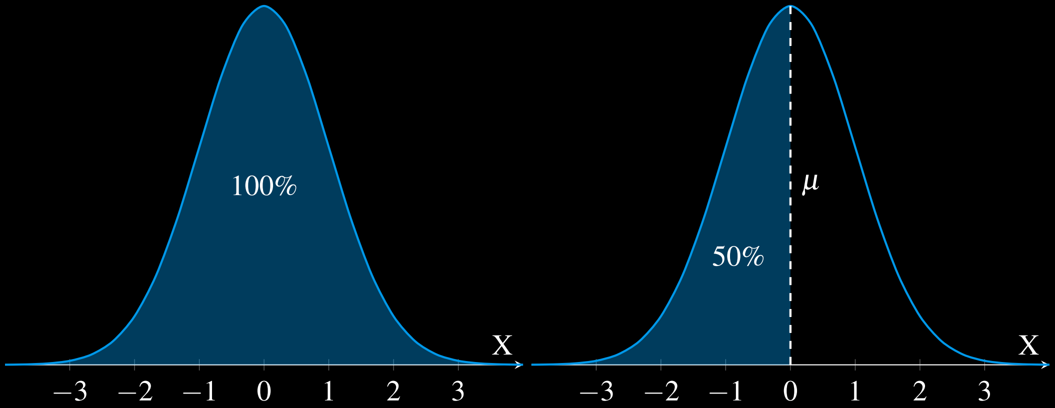
$$z_{180} = \frac{180 - 170}{9} = \frac{10}{9} \approx 1,11.$$



Normitetun normaalijakauman kuvaajassa satunnaismuuttujan arvo $1,11$ vastaa siis pituutta 180 cm. Toisiaan vastaavat arvot sijaitsevat yhtä monen keskihajonnan päässä keskiarvoistaan.

Todennäköisyydet pinta-aloina

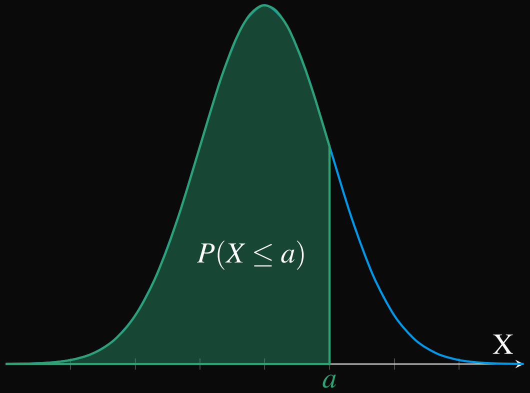
Todennäköisyyden tiheysfunktion käyrän alle jäävä pinta-ala kuvaa todennäköisyyttä.



Kertymätodennäköisyys

Määritelmä

Tapahtumaan $X \leq a$ liittyvää todennäköisyyttä $P(X \leq a)$ kutsutaan kertymätodennäköisyydeksi.

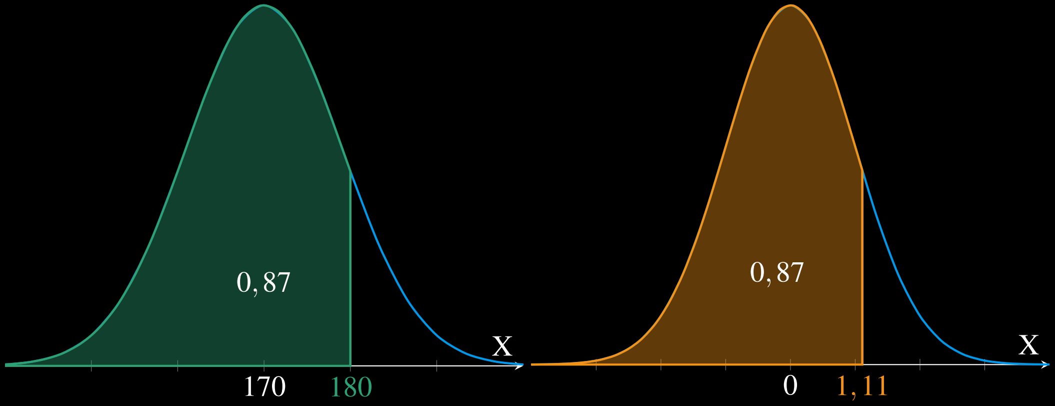


Kertymäfunktio on funktio, joka ilmaisee tämän kertyneen pinta-alan suuruuden.

Normitetun normaalijakauman kertymäfunktio

Aiemmassa esimerkissä totesimme, että pituuksien arvo 180 cm vastaa normitettua arvoa $1,11$. Vastaavasti niiden kertymäfunktioille pätee:

$$P(\text{pituus} \leq 180) = P(X \leq 1,11)$$



Sovelluksia

Normaalijakauma mallina

Lause (Keskeinen raja-arvolause)

Keskiarvo riittävän suuresta määrästä satunnaismuuttujia, joiden odotusarvo ja varianssi ovat hyvin määriteltyjä, jakautuvat likimain normaalisti tietyin edellytyksin.

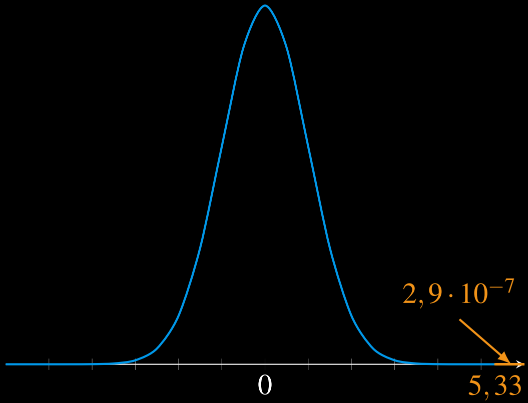
Seuraus

Normaalijakaumaa voi soveltaa hyvin monen satunnaisilmiön tutkimiseen.

Epätodennäköiset arvot

Suomalaisten miesten keskipituus on noin 181 cm ja pituuden keskihajonta noin 6 cm. Pituuksien voidaan olettaa noudattavan normaalijakaumaa $N(181, 6)$. Lauri Markkanen on 213 cm pitkä suomalainen koripalloilija. Hänen pituutensa normitettu arvo on

$$Z_{213} = \frac{213 - 181}{6} \approx 5,33, \text{ joten } P(X \geq 213) \approx 2,9 \cdot 10^{-7}$$



Tuntematon tunnusluku

Suomalaisten miesten mittauksessa keskiarvoksi saatiin 181 cm. Miehistä noin $6,7\%$ olivat yli 190 cm pitkiä. Mikä oli pituuksien keskihajonta?

1. Todennäköisyys $P(\text{pituus } \geq 190) = 0,067$
2. Normitetussa jakaumassa $N(0,1)$ pätee $P(X \geq x) = 0,067$, kun $x = 1,5$.
3. Koska $P(\text{pituus } \geq 190) = P(X \geq 1,5)$, niin

$$\begin{aligned} Z_{190} &= 1,5 \\ \frac{190-181}{\sigma} &= 1,5 \\ 1,5\sigma &= 190-181 = 9 \\ \sigma &= 6 \end{aligned}$$

Tilastollinen päättely

Keskiarvon keskivirhe ja luottamusväli

Johdanto

Aiemmassa esimerkissä todettiin, että suomalaisten miesten keskipituus on 181 cm ja sen keskihajonta 6 cm.

- Miten tulos on saatu?
- Onko tulos luotettava?
- Mistä tuloksen luotettavuus riippuu?

Useiden otoksien avulla saaduista keskiarvoista $\overline{X}$ voidaan muodostaa normaalijakauma, jossa keskiarvona on perusjoukon keskiarvo. Otosten keskiarvojen $\overline{X}$ poikkeamat kuvaavat niiden virheitä.

Seuraus

Perusjoukon keskiarvoa ja -hajontaa voidaan arvioida suorittamalla hyvin monta satunnaisotostutkimusta.

Keskiarvon keskivirhe

Määritelmä

Keskiarvon keskivirhe $s_{\overline{x}}$ on

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

missä n on otoksen koko ja s perusjoukon keskihajonta.

Keskivirhe kuvaa otoksen perusteella lasketun arvion tarkkuutta.

- Pieni keskivirhe $\Longrightarrow$ pieni vaihtelu
- Suuri keskivirhe $\Longrightarrow$ suuri vaihtelu

Esimerkki

Jos miesten pituuden tutkimisessa otoskoko $n = 1$, niin $s_{\overline{x}} = \frac{6}{\sqrt{1}} = 6$.

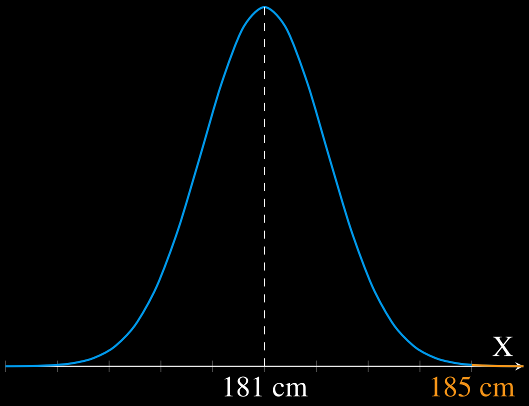
Toisaalta, jos $n = 100$, niin $s_{\overline{x}} = \frac{6}{\sqrt{100}} = 0,6$.

Keskivirhe käytännössä

Tutkitaan miesten pituutta 25 miehen otoksella. Aiemmin on todettu, että perusjoukon odotusarvo on 181 cm ja keskihajonta 6 cm. Tämän 25 miehen otoksen keskiarvon keskivirhe on

$$s_{\overline{x}} = \frac{6\text{cm}}{\sqrt{25}} = \frac{6\text{cm}}{5} = 1,2\text{cm}$$

Käytännössä:
sattumanvaraisesti tehdyn 25 miehen otosten keskipituudet jakautuvat normaalijakauman $N(181;1,2)$ mukaisesti. Esimerkiksi satunnaisen 25 miehen ryhmän keskipituus olisi yli 185 cm noin $0,048\%$ todennäköisyydellä.



Keskiarvon luottamusväli

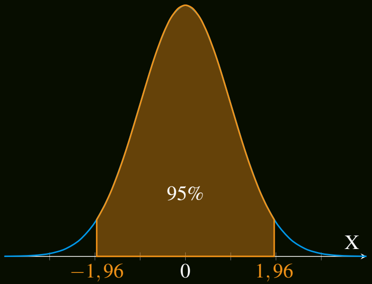
Otoksesta laskettu tunnusluku usein poikkeaa perusjoukon vastaavista tunnusluvuista. $\Longrightarrow$

Määritelmä (Luottamusväli)

- Luottamusväli on väli, jolla perusjoukon tunnusluku todennäköisesti sijaitsee.
- Luottamustaso kertoo, kuinka suurella todennäköisyydellä otoksen tunnusluku osuu tiettyyn luottamusväliin.

Esimerkki

Normitetussa normaalijakaumassa satunnaismuuttuja sijaitsee 95% todennäköisyydellä (luottamustaso) välillä $[-1,96; 1,96]$ (luottamusväli).
Lukua $1,96$ kutsutaan luottamustason 95% kriittiseksi arvoksi z^*.



Virhemarginaali

Määritelmä

Otoskeskiarvon virhemarginaali on kriittisen arvon ja keskivirheen tulo:

$$z^* \cdot \frac{s}{\sqrt{n}}$$

Esimerkki

Määritetään virhemarginaalit tutkimukselle, jossa mitattiin 100 suomalaisen miehen pituus. Otoskeskiarvoksi saatiin 179 cm ja keskihajonnaksi 7cm. Asetetaan luottamustasoksi 95%, jolloin kriittinen arvo on $1,96$ ja virhemarginaalina

$$1,96 \cdot \frac{7}{\sqrt{100}} \approx 1,3$$

Luottamusväliksi saadaan tällöin

$$[179 - 1,3\ ;\ 179 + 1,3] = [177,7\ ;\ 180,3].$$

Suuri otoskoko

Jos perusjoukon keskihajonta s tunnetaan tai otoskoko n on suuri, niin

1. keskiarvon keskivirhe on $\frac{s}{\sqrt{n}}$
2. luottamustason kriittinen arvo z^* määritellään normaalijakaumasta
3. luottamusväli on

$$\left[\bar{x} - z^* \cdot \frac{s}{\sqrt{n}}, \bar{x} + z^* \cdot \frac{s}{\sqrt{n}}\right]$$

Pieni otoskoko

Jos perusjoukon keskihajonta on tuntematon ja otoskoko on pieni (≤ 30), niin

1. keskihajontana käytetään otoskeskihajontaa s_{n-1}
2. luottamustason kriittinen arvo määritellään t-jakaumalla
3. luottamusväli eli t-väli muodostetaan ohjelmistolla (Geogebrassa "keskiarvon t-estimaatti")

Suhteellisen osuuden virhemarginaali ja luottamusväli

Suhteellinen osuus

Yle ja Helsingin sanomat julkaisevat säännöllisesti puolueiden kannatusmittauksia. Alla on taulukoituna syyskuun 2023 galluppien tulokset. Ylen otoksen koko oli 2 490 ja HS:n 2 456.

Puolue	Yle (%)	HS (%)
Kokoomus	21,5	21,0
SDP	20,8	21,6
PS	17,5	18,4
Keskusta	10,7	9,8
Vasemmistoliitto	9,8	8,3
Vihreät	9,1	8,6
RKP	4,0	4,2
KD	3,7	4,1
muut	1,7	2.2
Liike Nyt	1,2	1,8

Virhemarginaaliksi Yle ilmoitti 1,9 ja HS 2,0 %-yksikköä.

Suhteellisen osuuden otosjakauman keskihajonta

Määritelmä

Suhteellisen osuuden keskihajonta s voidaan laskea kaavasta

$$s = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}},$$

missä

- $\hat{p}$ on arvon suhteellinen osuus.
- n on otoksen koko.

Esimerkki

Ylen gallupin Kokoomuksen suhteellisen otoksen $21,5\%$ keskihajonnaksi saadaan

$$s_{YKok} = \sqrt{\frac{0,215 \cdot (1 - 0,215)}{2490}} \approx 0,0082 = 0,82\%$$

Suhteellisen osuuden virhemarginaali

Määritelmä

Suhteellisen osuuden virhemarginaali voidaan laskea samaan tapaan kuin keskiarvon virhemarginaali.

$$z^* \cdot s$$

missä

- z^* on luottamustason kriittinen arvo.
- s on suhteellisen otoksen keskihajonta.

Esimerkki

Luottamustasolla 95% ($z^* = 1,96$) Ylen gallupin Kokoomuksen tuloksen $21,5\%$ virhemarginaali on

$$1,96 \cdot 0,0082 = 0,01672 \approx 1,6\%,$$

jolloin luottamusväliksi saadaan

$$[21,5\% - 1,6\%; 21,5\% + 1,6\%] = [19,9\%; 23,1\%].$$