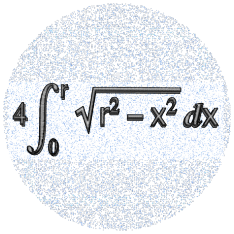


TYYPILLISET JA EPÄTYYPILLISET HAVAINNOT

Tekijät: Pauliina Ilmonen, Hannu Huhtalo, Armi Hopea-Manner, Ulla Mäkilä ja Mika Koskenoja

Laadittu keväällä 2022.

Kohderyhmä: Oppitunti yläkouluun, lukioon (lyhyeen tai pitkään matematiikkaan) tai ammatilliseen oppilaitokseen.

Yläkoulussa oppitunti sopii yhdeksännelle luokalle. Lukion lyhyessä matematiikassa oppitunti sopii lyhyen matematiikan moduliin MAB9. Ammatillisessa koulututuksessa soveltuvuus ja sopiva kurssi riippuvat koulutusohjelmasta. Lukion pitkässä matematiikassa oppitunti sopii moduliin MAA12 tai MAA8 (soveltuvin osin).

Esitiedot: Peruskoulumatematiikka

Oppitunnin kesto: 2×45 minuuttia (lukiossa 75 minuuttia voi riittää)

Oppimistavoitteet: Oppitunnin jälkeen yläkoululainen tuntee yleisimmät lokaatio- ja hajontasuureet aineistoille, ymmärtää mitä ne mittaavat ja osaa laskea ne. Hän tuntee Tsebysevin epäyhtälön aineistosta lasketun version ja osaa käyttää sitä helpoissa esimerkeissä.

Oppitunnin jälkeen lukiolainen (lyhyt matematiikka) ja ammatillisen oppilaitoksen opiskelija tuntee yleisimmät lokaatio- ja hajontasuureet aineistoille, ymmärtää mitä ne mittaavat ja osaa laskea ne suurillekin aineistoille sopivalla ohjelmistolla. Hän tuntee Tsebysevin epäyhtälön aineistosta lasketun version ja osaa käyttää sitä esimerkeissä.

Oppitunnin jälkeen lukiolainen (pitkä matematiikka) tuntee satunnaismuuttujan käsitteen ja osaa laskea odotusarvoja integroimalla. Hän tuntee yleisimmät lokaatio- ja hajontasuureet satunnaismuuttujille, ymmärtää mitä ne mittaavat, tietää miten niitä estimoidaan ja osaa laskea estimaatit suurillekin aineistoille sopivalla ohjelmistolla. Hän tuntee Tsebysevin epäyhtälön sekä teoreettisen että aineistosta lasketun version ja osaa käyttää sitä esimerkeissä.

Muut tavoitteet: Oppijat innostuvat analysoimaan itseään kiinnostavia aineistoja ja ymmärtävät, että tilastotiede on vahva ja tärkeä työkalu tiedon analysointiin.

Sisältö: Satunnaismuuttujat, odotusarvo, varianssi, keskihajonta, mediaani, (kvantiilit), vaihteluväli, otoskeskiarvo, otoskeskihajonta, Tsebysevin epäyhtälö, tilasto-ohjelmistot.

Toteutus:

Ensimmäinen oppitunti (45 minuuttia)

Seuraavasta rungosta opettaja voi valita haluamansa sisällön. Yläkoulussa, lukion lyhyessä matematiikassa ja ammatillisessa koulutuksessa suosittelemme jättämään väliin luvun 1 lopun (kertymäfunktion määritelmästä alkaen) sekä keskittymään luvuissa 2 ja 3 numeerisiin esimerkkeihin.

1. Satunnaismuuttujista

Satunnaismuuttuja. Tilastotieteessä ajatellaan, että havaintojen taustalla on jokin satunnaismuuttuja X. Satunnaismuuttujan mahdolliset lukuarvot muodostavat perusjoukon.

Tilastollinen tutkimus on havaintoaineistojen keräämistä, organisointia, analysointia ja tulkintaa. Tilastollisen tutkimuksen avulla selitetään ympäröivää maailmaa ja erilaisia ilmiöitä ja sitä käytetään apuna monenlaisessa päätöksenteossa.

Tilastotieteen menetelmät ja mallit ovat matemaattisia ja perustuvat todennäköisyyslaskentaan.

Populaatio ja otos.

- Tilastollinen perusjoukko eli *populaatio* koostuu kaikista mahdollisista kiinnostuksen kohteena olevista yksilöistä. (Esim. lukio-opiskelijat Suomessa.)
- Tilastollinen *otos* on populaation osajoukko. (Esim. 200 satunnaisesti valittua lukio-opiskelijaa.)
- Tutkimuksen kohteiksi valittuja perusjoukon alkioita kutsutaan *havaintoyksiköiksi*. (Esim. Maija Kuusankosken lukiosta.)

Havaintoaineisto. Tilastollinen aineisto koostuu tutkimuksen kohteita kuvaavien muuttujien havaituista arvoista. Esimerkkejä:

- lämpötila, pituus, rahamäärä, korko (jatkuvia kvantitatiivisia muuttujia)
- sukupuoli, asuinpaikka, väri, viallisuus (diskreettejä luokitteluasteikollisia kvalitatiivisia muuttujia)
- kouluarvosanat, vaatteiden koko (s, m, l), koulutustaso (diskreettejä järjestysasteikollisia muuttujia)

Satunnaismuuttujan arvon yleisyys ilmaistaan todennäköisyydellä ja kaikkien mahdollisten arvojen todennäköisyydet muodostavat todennäköisyysjakauman, joka määrittää satunnaismuuttujan täysin.

Määritelmä, kertymäfunktio. Olkoon X satunnaismuuttuja. Todennäköisyysmitan P avulla määritellään satunnaismuuttujalle X *kertymäfunktio*

$$F(x) = P(X \leq x).$$

Määritelmä, tiheysfunktio ja pistetodennäköisyysfunktio. Jatkuvan satunnaismuuttujan X *tiheysfunktio* $f(x)$, on sen kertymäfunktion derivaatta,

$$f(x) = F'(x).$$

(Huom. tiheysfunktio ei ole aina olemassa.)

Esimerkki. Koulussa kysyttiin uusilta lukion aloittavilta oppilailta, kuinka paljon heidän koulumatkansa muuttui viimeisimmästä matkasta yläkouluun verrattuna (itseisarvo, kilometreinä, voi sisältää desimaaleja). Huomattiin, että tiheysfunktio tälle oli likipitäen muotoa

$$f(x) = \begin{cases} \sin(x)/2, & \text{kun } 0 \leq x \leq \pi \\ 0 & \text{muualla.} \end{cases}$$

(Eli kenenkään koulumatka ei muuttunut yli $\pi \approx 3{,}14$ kilometriä.) Näin ollen koulumatkan muutoksen kertymäfunktio, eli todennäköisyys, että koulumatka muuttui korkeintaan x kilometriä, oli

$$\int_{-\infty}^{x} f(y)\,dy = \int_{0}^{x} \frac{\sin(y)}{2}\,dy = -\frac{\cos(x)}{2} + \frac{\cos(0)}{2} = -\frac{\cos(x)}{2} + \frac{1}{2}.$$

Esimerkiksi todennäköisyys, että matka muuttui korkeintaan 1,5 kilometriä oli

$$P(X \leq 1{,}5) = -\frac{\cos(1{,}5)}{2} + \frac{1}{2} \approx 46\%.$$

Diskreetin satunnaismuuttujan tiheysfunktiota vastaa *pistetodennäköisyysfunktio*

$$p(x) = P(X = x),$$

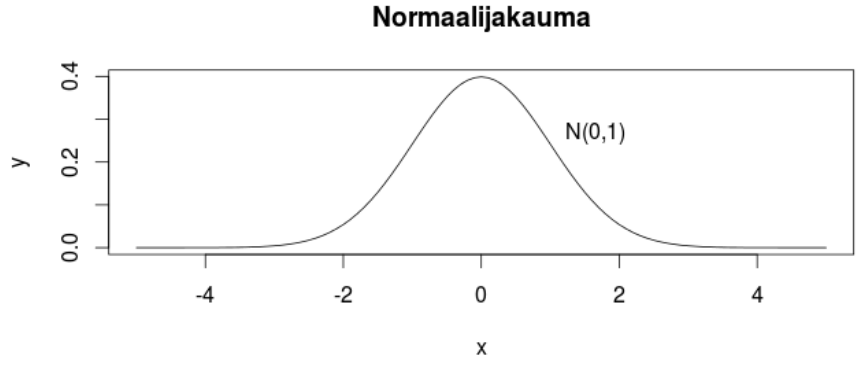
joka kertoo diskreetin satunnaismuuttujan todennäköisyyden saada arvo x.

Usein satunnaismuuttujat määritellään suoraan määrittelemällä niiden tiheys- ja/tai kertymäfunktio.

Tiheysfunktio, normaalijakauma. Yksi tunnettu jakauma on *normaalijakauma*. Normaalijakautuneen muuttujan tiheysfunktio

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Normaalijakauman parametrit ovat odotusarvo μ ja varianssi σ. Odotusarvon määritelmään tutustutaan seuraavaksi ja varianssiin luvussa 3.



Määritelmä, odotusarvo. Olkoon X jatkuva satunnaismuuttuja. Jos $\int_{-\infty}^{\infty} |h(x)|f(x)dx < \infty$, missä $f(x)$ on muuttujan X tiheysfunktio, niin satunnaismuuttujan $h(X)$ *odotusarvo* on (reaaliluku)

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

Olkoon X diskreetti satunnaismuuttuja, jonka arvojoukko on I. Jos $\sum_{x\in I} |h(x)|p(x) < \infty$, missä $p(x)$ on muuttujan X pistetodennäköisyysfunktio, niin satunnaismuuttujan $h(X)$ *odotusarvo* on

$$E[h(X)] = \sum_{x\in I} h(x)p(x).$$

Esimerkki.

- X:n odotusarvo $E[X]$ saadaan asettamalla $h(X) = X$.
- X:n *varianssi* $\text{var}[X]$ saadaan asettamalla $h(X) = (X - E[X])^2$.

Numeerinen esimerkki odotusarvoista. Olkoon X jatkuva satunnaismuuttuja, jolla on tiheysfunktio

$$f(x) = \begin{cases} 1, & 0 \le x \le 1, \\ 0, & \text{muualla.} \end{cases}$$

Halutaan odotusarvo $E[X]$, joten asetetaan $h(X) = X$ ja sijoitetaan

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_0^1 x \cdot 1dx = \frac{1}{2}.$$

Olkoon X diskreetti satunnaimuuttuja, jonka pistetodennälöisyysfunktio

$$p(x) = P(X = x) = \frac{1}{30} \cdot x^2, x \in \{0, 1, 2, 3, 4\}$$

Asettamalla $h(X) = X$ ja sijoittamalla, saadaan satunnaismuuttujan odotusarvo:

$$E[h(X)] = \sum x \cdot p(x) = 0 \cdot 0 + 1 \cdot \frac{1}{30} + 2 \cdot \frac{4}{30} + 3 \cdot \frac{9}{30} + 4 \cdot \frac{16}{30} = \frac{10}{3}.$$

2. Lokaatio

Lokaatio. Yleisimmin käytettyjä lokaatiolukuja ovat keskiarvo, mediaani ja moodi.

Keskiarvo. Olkoot $x_1, x_2, ..., x_n$ satunnaismuuttujan X toisistaan riippumattomat havaitut arvot. Tällöin *otoskeskiarvo*

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

estimoi muuttujan x odotusarvoa $E[X] = \mu$.

Mediaani. Olkoot $x_1, x_2, ..., x_n$ satunnaismuuttujan X toisistaan riippumattomat havaitut arvot. Olkoot $y_1 < y_2 < ... < y_n$ arvot asetettuna suuruusjärjestykseen. Tällöin otoksen *mediaani* on suuruusjärjestykseen asetettujen havaintojen keskimmäinen arvo. Jos havaintoja on parillinen määrä, otetaan kaksi keskimmäistä ja lasketaan näiden keskiarvo. Otosmediaani estimoi populaatiomediaania, joka määritellään seuraavasti: Satunnaismuuttujan X mediaani m on luku joka toteuttaa ehdot

$$P(X < m) \leq \frac{1}{2}, \text{ ja } P(X \leq m) \geq \frac{1}{2}.$$

Numeerinen esimerkki lokaatioluvuista. Olkoot $\{3, 1, 2, 3, 7, 8, 3, 4, 4, 6\}$ satunnaismuuttujan X toisistaan riippumattomat havainnot. Tällöin otoksen keskiarvo on

$$\bar{x} = \frac{1}{10} \cdot (3 + 1 + 2 + 3 + 7 + 8 + 3 + 4 + 4 + 6) = \frac{41}{10} = 4{,}1,$$

Otoksen mediaani on kahden suuruusjärjestyksessä keskimmäisen luvun keskiarvo tai suuruusjärjestyksessä keskimmäinen luku.

$$m = \frac{3+4}{2} = \frac{7}{2} = 3{,}5.$$

Kvantiili. Satunnaismuuttujan X β-kvantiili $k_\beta, 0 < \beta < 1$, on luku joka toteuttaa ehdot

$$P(X < k_\beta) \leq \beta, \text{ ja } P(X \leq k_\beta) \geq \beta.$$

Lokaatio. Muita lokaatiolukuja ovat esim. painotettu keskiarvo, vaihteluvälin keskipiste (midrange), ...

3. Hajonta

Hajonta. Yleisimmin käytettyjä hajontalukuja ovat varianssi, keskihajonta ja vaihteluväli.

Varianssi. Olkoot $x_1, x_2, ..., x_n$ satunnaismuuttujan X toisistaan riippumattomat havaitut arvot. Tällöin *otosvarianssi*

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

estimoi muuttujan X populaatiovarianssia $E[(X - E[X])^2] = \sigma^2$.

Keskihajonta. Otoksen $x_1, x_2, ..., x_n$ *keskihajonta* on sen otosvarianssin neliöjuuri

$$s = \sqrt{s^2}.$$

Vaihteluväli. Olkoot $x_1, x_2, ..., x_n$ satunnaismuuttujan X toisistaan riippumattomat ja samoin jakautuneet havaitut arvot. Olkoon max otoksen suurin arvo ja olkoon min otoksen pienin arvo. Tällöin otoksen vaihteluväli on väli $[\min, \max]$ ja vaihteluvälin pituus on $\max - \min$.

Numeerinen esimerkki hajontaluvuista. Olkoot $\{3, 1, 2, 3, 7, 8, 3, 4, 4, 6\}$ satunnaismuuttujan X toisistaan riippumattomat ja samoin jakautuneet havainnoit. Otoskeskiarvo on 4,1. Tällöin otoksen varianssi

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - 4{,}1)^2 = 4{,}9888...$$

ja keskihajonta $\sqrt{s^2} = \sqrt{4{,}9888...} = 2{,}233....$

Vaihteluväli saadaan havaintojen minimin ja maksimin avulla:

$$[\min, \max] = [1, 8].$$

Vaihteluvälin pituus on $8 - 1 = 7$.

Tsebysevin epäyhtälö. Olkoon X satunnaismuuttuja, jolla on äärellinen odotusarvo $E[X] = \mu$ ja äärellinen varianssi $E[(X - E[X])^2] = \sigma^2$. Olkoon $k > 1$. Tällöin

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Tsebysevin epäyhtälö. Kun $k = 2$, niin $1 - \frac{1}{k^2} = 75\%$. Eli havainto sijaitsee kahden keskihajonnan sisällä odotusarvosta ainakin 75% todennäköisyydellä.

Kun $k = 3$, niin $1 - \frac{1}{k^2} \approx 88{,}9\%$. Eli havainto sijaitsee kahden keskihajonnan sisällä odotusarvosta vähintään noin 88,9%. todennäköisyydellä.

Käytännössä odotusarvo ja varianssi joudutaan estimoimaan.

Tsebysevin epäyhtälö, otosversio. Olkoon $x_1, \ldots, x_n$ jonkin satunnaismuuttujan X toisistaan riippumattomat ja samoin jakautuneet havaitut arvot. Olkoot $\bar{x}$ otoksesta laskettu keskiarvo ja s otoksesta laskettu keskihajonta. Tällöin kun otoskoko on suuri pätee suunnilleen, että samasta jakaumasta tulevalla satunnaisella x on

$$P(|x - \bar{x}| \geq ks) \leq \frac{1}{k^2}$$

kaikilla $k > 1$.

Tsebysevin epäyhtälöä käytetään, kun arvioidaan yksittäisen havaintoarvon yleisyyttä. Yleisesti havaintoja, jotka sijaitsee yli kahden keskihajonnan päässä keskiarvosta, pidetään harvinaisina ja havaintoja, jotka sijaitsevat yli kolmen keskihajonnan etäisyydellä keskiarvosta pidetään hyvin harvinaisina.

Jos tiedetään, että havainnot tulevat normaalijakaumasta, niin saadaan Tsebysevin epäyhtälöä tarkemmat rajat. Tällöin tiedetään, että todennäköisyys sille, että havainto sijaitsee yhden keskihajonnan sisällä odotusarvosta on noin 68 %. Todennäköisyys sille, että havainto sijaitsee kahden keskihajonnan sisällä odotusarvosta on noin 95 % ja todennäköisyys sille, että havainto sijaitsee kolmen keskihajonnan sisällä odotusarvosta on noin 99,7 %.

Toinen oppitunti (45 minuuttia)

Toisella oppitunnilla sovelletaan ensimmäisellä oppitunnilla opittuja asioita. Käytetään esimerkiksi Ilmatieteen laitoksen tarjoamia säähavaintoaineistoja
`https://www.ilmatieteenlaitos.fi/havaintojen-lataus`.
Valitaan halutut suureet (esim. lämpötila tai tuulen nopeus), aikaväli ja paikkakunta. Voidaan myös valita useita paikkakuntia ja vain yksi ajanhetki.

Yläkoulussa toisen oppitunnin esimerkin voi halutessaan korvata helpommalla esimerkillä, jonka pystyy laskemaan ilman tilasto-ohjelmaa.

Esimerkki. Valitse paikkakunnaksi Kouvolan Anjalan mittauspiste. Lataa 1970-luvun (vuodet 1970–79) jokaisen heinäkuun päivittäiset maksimilämpötilat. Tällöin saat 10 Excel-tiedostoa. Yhdistä nämä yhdeksi taulukoksi. Lisäksi lataa vuoden 2021 heinäkuun päivittäiset maksimilämpötilat. Laske 1970-luvun aineistosta keskiarvo, mediaani, vaihteluväli, varianssi ja keskihajonta. Tee sama vuoden 2021 aineistolle. Vertaa laskettuja suureita. Valitse satunnainen päivä vuoden 2021 aineistosta. Käytä Tsebysevin epäyhtälössä 1970-luvun aineiston keskiarvoa ja keskihajontaa. Katso, olisiko vuoden 2021 aineiston havainto poikkeava, jos oletettaisiin, että jakauma ei ole muuttunut.

Katso ohessa oleva Excel-tiedosto, joka on muodostettu hakemalla 24.4.2022 Ilmatieteen laitoksen kotisivuilta [3] Kouvolan Anjalan säätiedot 1970-luvulta ja laskemalla siihen halutut tiedot. Saadaan seuraavat tulokset: 1970-luvun aineiston heinäkuiden päivittäisten maksimilämpötilojen keskiarvo on 21,85968 °C $\approx$ 21,9 °C, mediaani on 22,0 °C, vaihteluväli on [11,8 °C; 32,5 °C], vaihteluvälin pituus on 20,7 °C, varianssi on 13,51348 °C^2 $\approx$ 13,5 °C^2 ja keskihajonta on 3,676069 °C $\approx 3,7$ °C. Vuoden 2021 aineiston heinäkuun päivittäisten maksimilämpötilojen keskiarvo on 27,2 °C, mediaani on 28,6 °C, vaihteluväli on [21,0 °C; 32,7 °C], vaihteluvälin pituus on 11,7 °C, varianssi on 12,5 °C^2 ja keskihajonta on 3,5 °C.

Valitaan vuoden 2021 aineistosta satunnaiseksi päiväksi 13.7. Tällöin maksimilämpötila oli 31,4 °C. Pitää laskea tämän ja 1970-luvun aineiston keskiarvon erotuksen itseisarvo $|31{,}4 - 21{,}85968| = 9{,}54032$ (°C). Sen jälkeen tämä itseisarvo jaetaan 1970-luvun aineiston keskihajonnalla

$$\frac{9{,}54032\ ^\circ\text{C}}{3{,}676069\ ^\circ\text{C}} \approx 2{,}6.$$

Nyt tiedämme, että lämpötila oli noin 2,6 keskihajonnan päässä 1970-luvun heinäkuiden päivittäisten maksimilämpötilojen keskiarvosta. Sitä voidaan siis pitää harvinaisena, mutta ei hyvin harvinaisena

Huomataan, että Anjalassa oli kesällä 2021 tosi kuumaa.

Toki esimerkissä Kouvolan Anjalan voi korvata haluamallaan paikkakunnalla.

G	H	I	J	K	L	M
Ylin lämpötila (degC)		Heinäkuun päivien maksimilämpötilojen keskiarvo:				
26,2		21,85968		”=KESKIARVO(G2:G311)”		
23,8						
20,2		Keskihajonta:		”=KESKIHAJONTA.S(G2:G311)”		
18,6		3,676069				
19,8						
20,1		Varianssi:		”=VAR.S(G2:G311)”		
18,4		13,51348				
22,2						
24,8		Mediaani:		”=MEDIAANI(G2:G311)”		
25		22				
21,7						
17		Vaihteluväli:		”=MIN(G2:G311)”		
19,4			minimi:			
18,5			11,8			
16,7						
16,6			maksimi:	”=MAKS(G2:G311)”		
19,9			32,5			
20						
31,2			siis väliksi	saadaan		
24,9			[11,8;32,5]			
20,1						
16,9			välin pituus:			
14,4			20,7			

Viitteet

[1] J. Crawshaw, J. Chambers: A Concise Course in Advanced Level Statistics, Nelson Thornes Ltd 2013.
[2] R. V. Hogg, J. W. McKean, A. T. Craig: Introduction to Mathematical Statistics, Pearson Education 2005.
[3] Ilmatieteen laitos, https://www.ilmatieteenlaitos.fi/havaintojen-lataus.
[4] J. S. Milton, J. C. Arnold: Introduction to Probability and Statistics, McGraw-Hill Inc 1995.