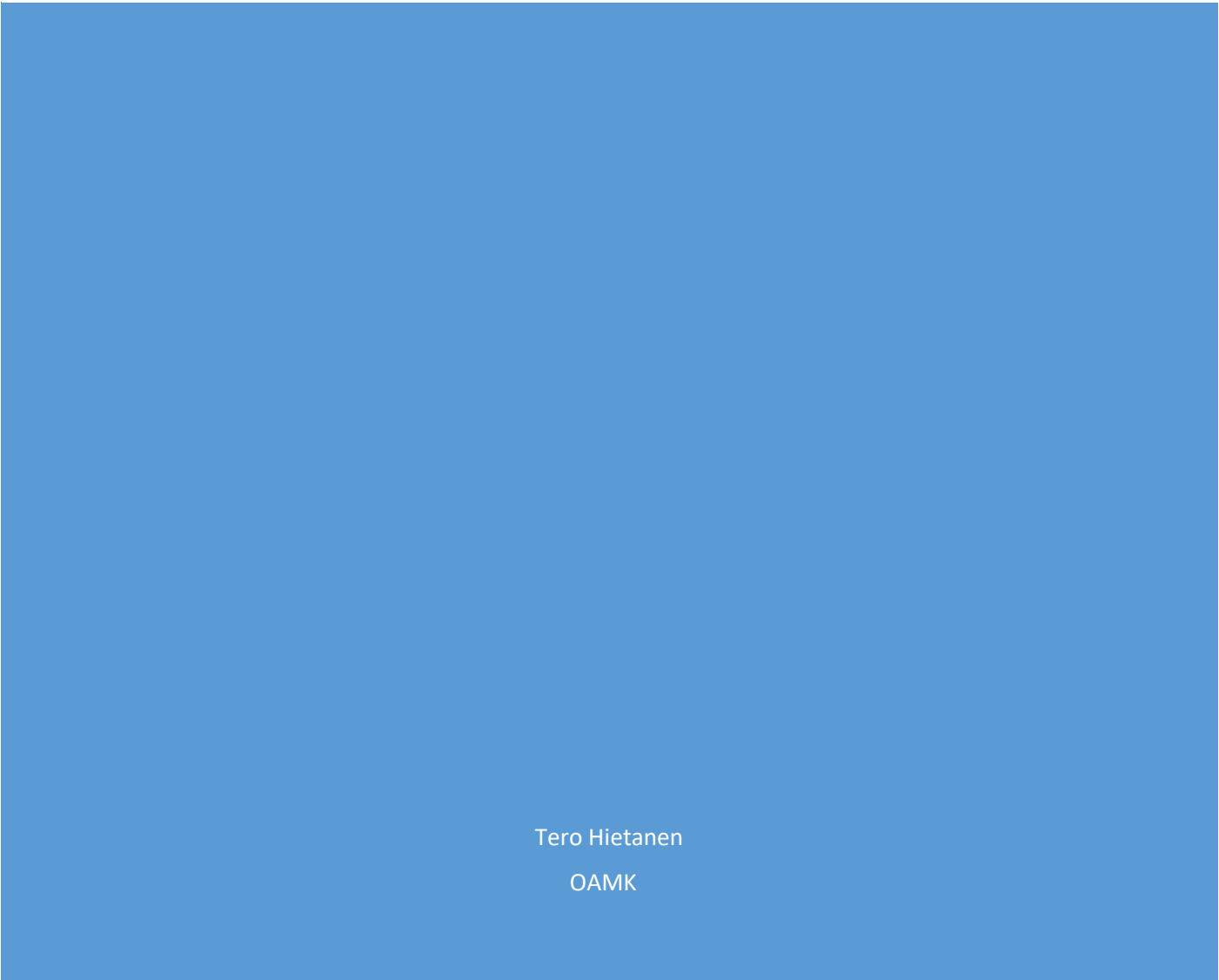




Tilastollinen päättely

2019



Tero Hietanen

OAMK

Sisällys

1	Johdanto	2
2	Normaalijakauma	3
3	Normaalijakaumasta johdettuja jakaumia	5
3.1	Studentin t-jakauma	6
3.2	Khii toiseen (χ^2)-jakauma	7
3.3	Fisherin F- jakauma.....	7
4	Hypoteesien testaus	7
5	Sovellusesimerkki tilastollisesta analyysistä.....	9

1 Johdanto

Tilastollisen tutkimuksen kaikkien mahdollisten kohteiden muodostamaa joukkoa tutkimuksen (kohde-) perusjoukoksi ja kutsutaan tutkimuksen kohteiksi valittuja perusjoukon alkioita havaintoyksiköiksi. Tilastollinen aineisto koostuu havaintoyksiköitä kuvaavien muuttujien havaituista arvoista. Olkoon tutkimuksen kohteiksi valittujen havaintoyksiköiden lukumäärä n ja olkoon x_i , $i = 1, 2, \dots, n$ kohdeperusjoukon alkoiden ominaisuutta kuvaavan muuttujan x havaittu arvo havaintoyksikössä i . Kutsumme muuttujan x havaittuja arvoja $x_1, x_2, \dots, x_n$ havaintoarvoiksi tai havainnoiksi. Havaintoarvo x_i saadaan mittaamalla muuttujan x arvo havaintoyksikölle i . Tilastollinen aineisto voi syntyä tilastollisen kokeen tuloksena tai tekemällä suoria havaintoja. Jos tutkimuksen kohteena on koko perusjoukko, tutkimusta kutsutaan kokonaistutkimukseksi, jos vain satunnaisesti valittu osa perusjoukosta tutkitaan, kutsutaan tutkimusta otantatutkimukseksi.

Olkoot $x_1, x_2, \dots, x_n$ muuttujan x havaittuja arvoja. Muuttujan x havaittujen arvojen jakaumaa voidaan kuvailla ja esitellä tiivistämällä havainto-arvoihin sisältyvä informaatio sopivaan muotoon:

1. Jakaumaa kokonaisuutena voidaan kuvata sopivasti valituilla graafisilla esityksillä.
2. Jakauman karakteristisia ominaisuuksia voidaan kuvata sopivasti valituilla tunnusluvuilla.
3. Tunnuslukujen tehtävänä on kuvata havaintoarvojen jakauman keskeisiä karakteristisia ominaisuuksia: Keskimääraisten, tyyppillisten tai yleisten havaintoarvojen sijaintia kuvataan **keskiluvuilla**. Havaintoarvojen hajaantuneisuutta tai keskittyneisyyttä kuvataan **hajontaluvuilla**. /1/

Keskiluvuilla tarkoitetaan tässä yhteydessä erilaisia keskiarvoja ja odotusarvoja. Hajontaluvuilla tarkoitetaan varianssia ja keskihajontaa. Tavallisin keskiarvo aritmeettinen keskiarvo voidaan laskea intervalli- tai suhdeasteikollisesta muuttujasta. Luokittelemattomalle aineistolle keskiarvo saadaan kaavasta.

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

Odotusarvo on satunnaismuuttujan tärkein tunnusluku. Otoskeskiarvolla tarkoitetaan suppean otoksen keskiarvoa suuremmasta populaatiosta. Sen avulla on mahdollista selvittää odotusarvon suuruus likimääräisesti. Keskiarvoa on tällöin pidettävä odotusarvon estimaattina. Kun lasketaan otoskeskiarvojen odotusarvoa, saadaan edellisten päättelysääntöjen avulla: /2/

$$\mu = E(X) = E\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \quad (2)$$

Käytetyimpiä hajontalukuja ovat **varianssi s^2** ja **keskihajonta s** . Varianssi ja keskihajonta voidaan laskea intervalli- tai suhdeasteikollisesta muuttujasta. Keskihajonta on varianssin positiivinen neliöjuuri eli:

$$s = \sqrt{s^2} \quad (3)$$

Varianssi kertoo, kuinka tiiviisti havaintoarvot ovat keskittyneet keskiarvon ympärille. Jos kaikki mittaustulokset ovat samoja, on $s^2 = 0$, muulloin $s^2 > 0$. Keskihajonnassa ja varianssissa muuttujien arvojen vaihtelu ilmaistaan rakenteellisesti samalla tavalla. Keskihajonta on kuvailussa havainnollisempi, koska sillä on sama laatu kuin muuttujan arvoilla, ja se kertoo, kuinka kaukana keskimäärin havainnot ovat keskiarvosta. Varianssi on taas parempi teoreettisissa tarkasteluissa. Luokittelemattoman aineiston varianssi voidaan laskea kaavalla: /3/

$$s^2 = \sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4)$$

Tässä materiaalissa tehdään jatkossa pientä väkivaltaa käyttämällä σ^2 myös varianssin harhattomana estimaattina. Laskennallisestihan on kysymys siitä käytetäänkö yhtälön 4 kertoimena $1/n$ vai $1/(n-1)$, otoskoon kasvaessahan tämän merkitys häviää.

2 Normaalijakauma

Satunnaismuuttujat, merk. $\underline{x}$, $\underline{y}$, ... ovat empiiristen muuttujien matemaattisia vastineita. Käytännössä satunnaismuuttujalla tarkoitetaan jotakin satunnaiskokeen tulokseen liittyvää numeerista suuretta, jonka arvo vaihtelee koetta toistettaessa. Satunnaismuuttuja on satunnainen siinä mielessä, että etukäteen ei voida varmasti tietää, mikä ilmiön tuloksista realisoituu. Esim. Nopanheitossa saatava silmäluku on satunnaismuuttuja. Tiettyinä vuonna tapahtuvien liikenneonnettomuuksien lkm on satunnaismuuttuja. Matemaattisesti määriteltynä satunnaismuuttuja $\underline{x}$ on kuvaus (funktio) alkeistapausten joukolta reaalilukujoukolle. Itse funktio ei ole satunnainen, vaan se on täysin määrätty. Satunnaismuuttujan $\underline{x}$ arvo x voidaan määrätä vasta, kun satunnaiskokeen tulos on selvillä. Empiirisen muuttujan jakauma on selvillä silloin, kun tunnetaan muuttujan saamat arvot (tai luokat, joihin arvot kuuluvat) ja näiden arvojen frekvenssit. Vastaavasti saadaan selville teoreettisen satunnaismuuttujan jakauma ns. todennäköisyysjakauma, mutta frekvenssien sijasta tarkastellaan satunnaismuuttujan arvojen todennäköisyyksiä.

Jatkuvan satunnaismuuttujan jakaumaa ilmaisevaa funktiota f sanotaan tiheysfunktioksi $f(x)$, joka on määriteltä yhtälössä 5.

$$f(x) = \int_{-\infty}^{\infty} f(x) dx = 1 \quad (5)$$

$f(x)$:n rajoittama kokonaispinta-ala on 1. Satunnaismuuttujan $\underline{x}$ kertymäfunktio $F(x)$, määritellään seuraavasti:

$$F(x) = P(\underline{x} \leq a) \quad (6)$$

$F(x)$ ilmaisee todennäköisyyden, että satunnaismuuttuja $\underline{x}$ saa enintään arvon a . Jatkuvalla satunnaismuuttujalle $\underline{x}$ on kertymäfunktio $F(x)$:

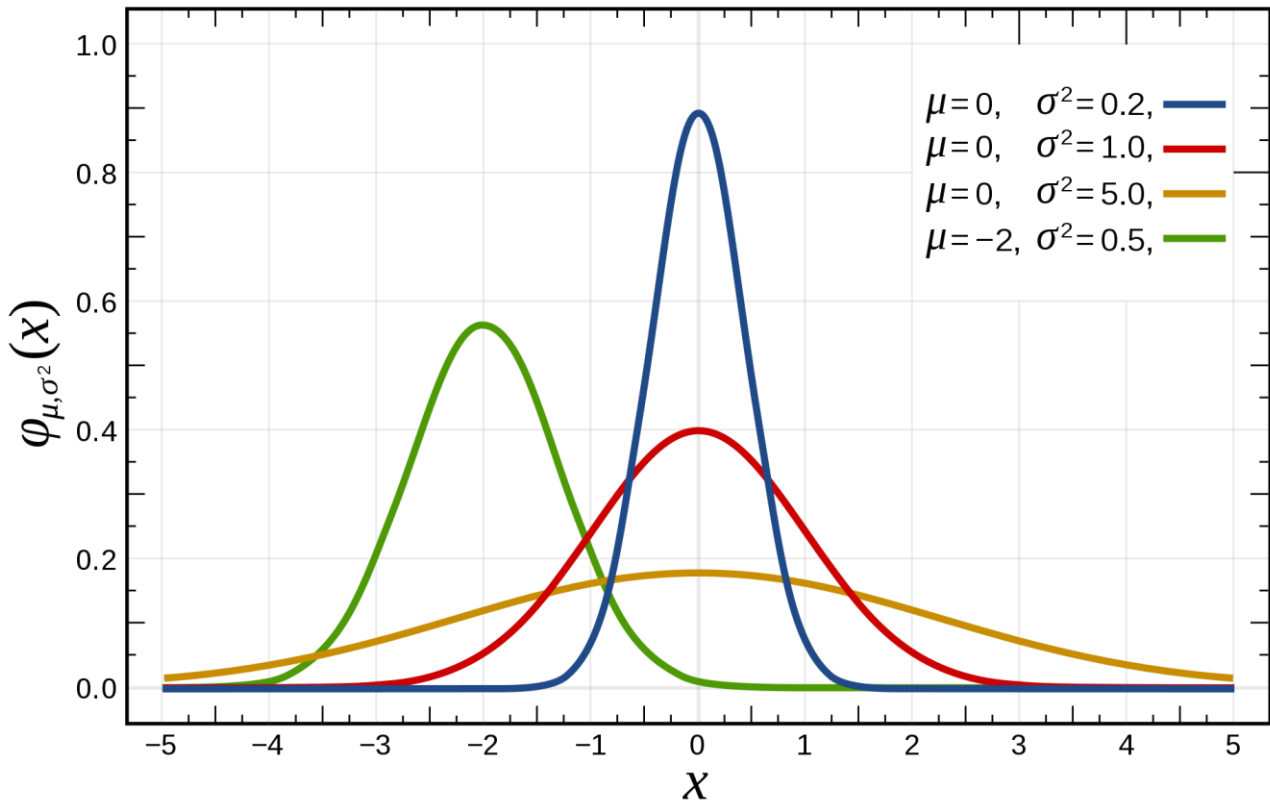
$$F(x) = \int_{-\infty}^a f(x) dx \quad (7)$$

Tilastotieteen kannalta tärkein jatkuva jakauma on normaalijakauma. Jatkuvaa satunnaismuuttujaa $\underline{x}$, joka voi saada arvokseen kaikki reaaliluvut x , sanotaan normaalijakautuneeksi parametrein odotusarvo $E(\underline{x})=\mu$ ja varianssi $\text{var}(\underline{x})=\sigma^2$, $\underline{x} \sim N(\mu, \sigma^2)$, jos sen tiheysfunktio on muotoa:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (8)$$

, missä $-\infty < \mu < \infty$ ja $\sigma > 0$.

Normaalijakauman tiheysfunktion kuvaaja on symmetrinen odotusarvon μ suhteen. Tiheysfunktion $f(x)$ kuvaaja lähestyy x-akselia μ :n molemmilla puolilla. Tiheysfunktio on yksihiippuinen. Kuvaajan sijainti määräytyy odotusarvon μ mukaan ja muoto varianssin σ^2 mukaan, kuva 1. /3/



Kuva 1. Normaalijakauman tiheysfunktion arvoja erilaisilla odotusarvoilla ja variansseilla. Horisontaaliakselilla $\underline{x}=X$ ja vertikaaliakselilla $f(x)=\varphi(x)$. /4/

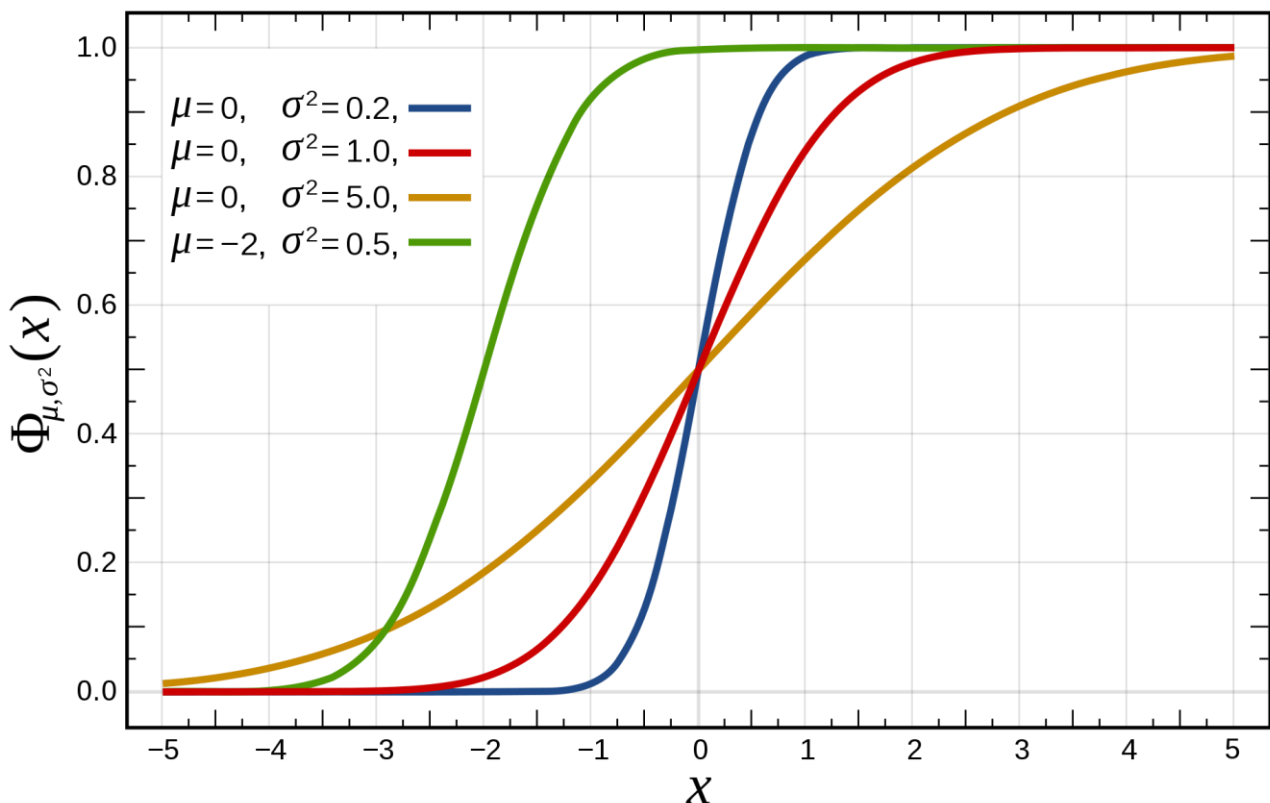
Normaalijakauman tärkein erikoistapaus on ns. standardoitu normaalijakauma. Nyt $\mu = 0$ ja $\sigma^2 = 1$, jolloin satunnaismuuttajalle $z \sim N(0, 1)$. Tiheysfunktio on esitetty yhtälössä 9.

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (9)$$

ja kertymäfunktio

$$\Phi(z) = P(\underline{z} \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad (10)$$

Kuvassa on esitetty kertymäfunktio erilaisilla hajonnan ja varianssin arvoilla.



Kuva 2. Kuvassa 2 on esitetty kuvan 1 mukaisten satunnaisjakaumien kertymäfunktioiden kuvaajat. /4/

Kertymäfunktioiden Φ arvoja on taulukoitu, koska tiheysfunktiolla φ ei ole integraalifunktiota alkeisfunktioiden joukossa.

Useimmissa taulukoissa on kertymäfunktion arvoja esitetty vain z :n positiivisilla arvoilla, joten negatiivisilla arvoilla nojaututaan normaalijakauman symmetrisyyteen ($z > 0$):

$$P(\underline{z} \leq -z) = \Phi(-z) = 1 - \Phi(z) \quad (11)$$

Standardoidun normaalijakauman $N(0, 1)$ käyttökelpoisuus perustuu seuraavaan tulokseen. Jos $\underline{x} \sim N(\mu, \sigma^2)$, niin muunnoksella:

$$\underline{z} = \frac{\underline{x} - \mu}{\sigma} \quad (12)$$

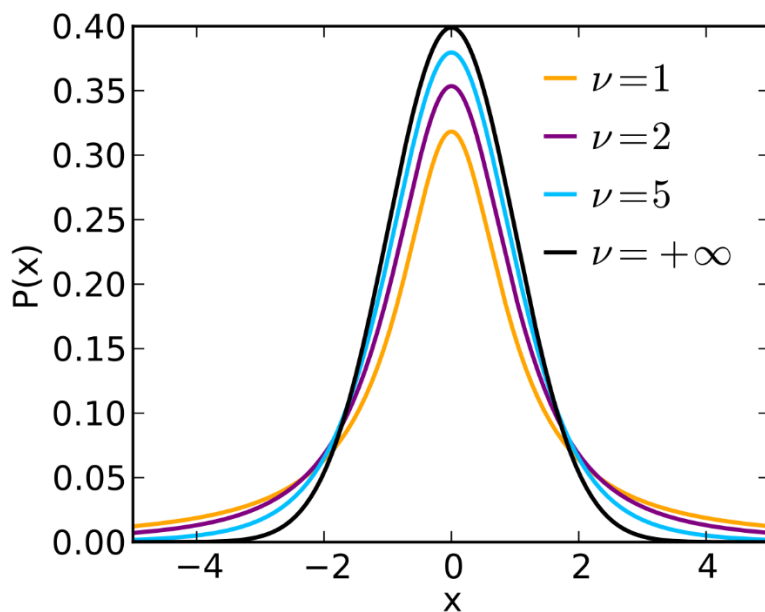
,jota kutsutaan standardoinniksi, saadaan $E(\underline{z}) = 0$ ja $\text{Var}(\underline{z}) = 1$. Tällöin $\underline{z} \sim N(0, 1)$. Normaalijakauman keskeinen asema tilastotieteessä perustuu siihen, että useat empiiriset jakaumat ovat likimain normaalisia. Keskeisen raja-arvolauseen mukaan mielivaltaisten riippumattomien satunnaismuuttujien summan jakauma on likimain normaalin, jos yhteenlaskettavien lukumäärä on suuri, ja eräät hyvin yleisluontoiset ehdot ovat voimassa. /3/

3 Normaalijakaumasta johdettuja jakaumia

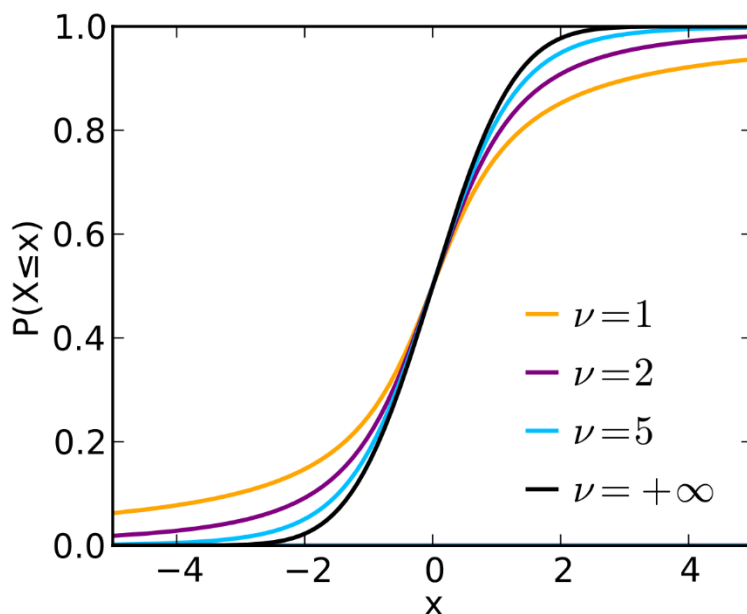
Luvussa tarkastellaan keskiarvojen, odotusarvojen ja varianssien testaamisen kannalta merkittäviä tilastollisia jakaumia. Näitä ovat Studentin t-jakauma, khii toiseen jakauma sekä Fisherin F-jakauma.

3.1 Studentin t-jakauma

Studentin t-jakauma on todennäköisyysjakauma, jota hyödynnetään normaalijakautuneiden populaatioiden keskiarvon tarkastelussa kun otoskoko on pieni. Yleisesti käytetty t-testi perustuu t-jakaumaan. Jakauman arvojen muodostuksessa on hyödynnetty Γ -funktioita. [Gammafunktio](#) Gamma-funktio pienentää odotusarvon todennäköisyyttä ja leventää jakauman arvoaluetta. Otskoon kasvaessa t-jakauma lähestyy normitettua normaalijakaumaa. Tiheys- ja kertymäfunktion arvoja tietyillä vapausasteilla voidaan esim. hypoteesien yhteydessä lukea taulukoiduista arvoista, [linkki](#). Kuvisa 3 ja 4 on esitetty t-jakauman tiheys- ja kertymäfunktiot erilaisilla vapausasteilla.



Kuva 3. t-jakauman tiheysfunktio.



Kuva 4. t-jakauman kertymäfunktio. /5/

Olkoot $\underline{x}$, $z_1, \dots, z_n$ $N(0, 1)$ -jakautuneita riippumattomia satunnaismuuttujia. Silloin satunnaismuuttuja $\underline{t}$:

$$\underline{t} = \frac{\underline{x}}{\sqrt{\frac{1}{n} \sum_{i=1}^n z_i^2}} \quad (13)$$

on Studentin t-jakautunut vapausasteella n . Odotusarvona on $E(\underline{t}) = 0$ ja varianssina $\text{Var}(\underline{t}) = n/(n-2)$.

3.2 Khii toiseen (χ^2)-jakauma

Olkoot $z_1, \dots, z_n$ $N(0, 1)$ -jakautuneita riippumattomia satunnaismuuttujia. Silloin satunnaismuuttuja:

$$\underline{\chi^2} = \sum_{i=1}^n z_i^2 \quad (14)$$

noudattaa χ^2 -jakaumaa, jonka muoto riippuu vapausasteluvusta n . Odotusarvona on $E(\underline{\chi^2}) = n$ ja varianssina $\text{Var}(\underline{\chi^2}) = 2n$. Linkistä löytyy jakauman Tiheys- ja kertymäfunktion [kuvaajat](#). Jakauman muodostamisessa on hyödynnetty eulerin gammafunktia. χ^2 -jakaumaa voidaan käyttää riippumattomuus- ja yhteensopivuustesteissä. Jakauman taulukoituja arvoja [linkissä](#).

3.3 Fisherin F- jakauma

Olkoot $\chi^2(m)$ ja $\chi^2(n)$ riippumattomia χ^2 -jakautuneita satunnaismuuttujia vapausasteina m ja n . Silloin satunnaismuuttuja:

$$\underline{F} = \frac{\frac{\chi^2(m)}{m}}{\frac{\chi^2(n)}{n}} \quad (15)$$

noudattaa Fisherin F-jakaumaa vapausastein m ja n . Odotusarvona on $E(\underline{\chi^2}) = n/(n-2)$, $n > 2$ ja varianssina $\text{Var}(\underline{\chi^2}) = n^2(2m+2n-4)/(m(n-2)^2(n-4))$, $n > 4$. /6/ Linkistä löytyy jakauman Tiheys- ja kertymäfunktion [kuvaajat](#). F-jakauman taulukoituja arvoja on [linkissä](#). F-jakaumaa voidaan käyttää mm. varianssitesteissä.

4 Hypoteesien testaus

Kaikkien tilastoyksiköiden joukko muodostaa populaation eli perusjoukon. Perusjoukon osajoukkoa kutsutaan satunnaisotokseksi, jos jokaisella perusjoukon alkiolla on tunnettu positiivinen todennäköisyys tulla valituksi otokseen. Jos ko. ehto ei ole voimassa, on kyseessä näyte. Mikäli

tiedot kerätään jokaisesta perusjoukon alkioista, on kyseessä kokonaistutkimus. Jos tiedot kerätään otoksen avulla, on kyseessä otantatutkimus. /3/

Otantatutkimuksen suunnittelun vaiheet:

1. Tavoitteiden määrittely

- täsmentäminen
- tietojen käyttötarkoitus
- tulosten täsmällisyysvaatimus

2. Ongelman muuntaminen tilastolliseksi

- perusjoukon täsmentäminen
- tilastoyksiköt / otantayksiköt
- muuttujat
- kohdepopulaatio/kehikkopopulaatio
- mittausmenetelmät
- kyselykaavakkeet

3. Otoksen poimiminen

- otantamenetelmän valinta
- otoskoon määrittäminen

4. Aineiston kerääminen

- kysely tai muunlainen mittaaminen

5. Aineiston käsittely

- tietojen muokkaus ja analysointi

6. Raportointi

- tulosten esittäminen selväkielisesti

Kootun havaintoaineiston perusteella pyritään päättämään, pitääkö perusjoukon parametreista tms. tehty oletamus eli hypoteesi paikkansa. Tilastollinen hypoteesi on väite, joka tehdään populaation jakaumasta ja se koskee jotakin populaation parametria. Tällaista väitettä kutsutaan myös nollahypoteesiksi H_0 . Yleensä nollahypoteesi on oletusarvo, joka edustaa aikaisempaa tietämystä. Testin avulla H_0 joko hylätään tai hyväksytään. H_0 :n vaihtoehtoa sanotaan vastahypoteesiksi tai vaihtoehtoiseksi hypoteesiksi H_1 , joka voi esim. edustaa uutta teoriaa, josta varsinaisesti ollaan kiinnostuneita.

Esim. Maalipurkkeja täyttävän koneen on tarkoitus täyttää 1000 g:n purkkeja. Aikaisempien tutkimusten mukaan painon on havaittu vaihtelevan keskihajontana 10 g. Koneen toiminnan tutkimiseksi asetetaan hypoteesit $H_0: \mu = 1000$ g ja $H_1: \mu \neq 1000$ g. Edellä vastahypoteesi H_1 on

kaksisuuntainen. Yksisuuntainen hypoteesi olisi kyseessä ollut silloin, jos olisi asetettu $H_1: \mu > 1000$ g tai $H_1: \mu < 1000$ g.

Kaksisuuntaisuus tässä on avainasia tarkasteltaessa vertailulukua jakauman kertymäfunktioista. Käytännössä tässä kaksisuuntaisuus syntyy siitä, että maalipurkin paino voi olla joka pienempi tai suurempi kuin odotusarvo. Kaksisuuntaisessa tapauksessa pitää siis huomioida jakauman molemmat puolet eli jos testisuure katsotaan 95% luottamusvälillä pitää se katsoa taulukosta 97.5% prosenttien kohdalta.

Koska valinta H_0 :n ja H_1 :n välillä tapahtuu otoksen antaman informaation avulla, on muodostettava testisuure, jonka avulla päätös tehdään. Usein testisuure perustuu testattavan parametrin estimaattoriin. Jotta testisuureen otantajakauma voidaan määrätä, on edellytettävä tiettyjen jakaumaa koskevien oletusten olevan voimassa. /3/

5 Sovellusesimerkki tilastollisesta analyysistä

Kirjallisuusluettelo:

1. <https://math.aalto.fi/opetus/sovtoda/oppikirja/Johdanto.pdf> 12.6.2019
2. <https://fi.wikipedia.org/wiki/Odotusarvo> 12.6.2019
3. <http://lipas.uwasa.fi/~chg/Tilastotieteen%20perusteet%20luentorunko.pdf> 12.6.2019
4. https://fi.wikipedia.org/wiki/Normaalijakauma#/media/File:Normal_Distribution_PDF.svg
12.6.2019
5. https://fi.wikipedia.org/wiki/Studentin_t-jakauma 12.6.2019
6. https://math.aalto.fi/opetus/sovtodb/luennot_syksy2007/l08.pdf 12.6.2019