MTTTP5, luento 23.10.2018

1 Kokonaisuudet, joihin opintojakso kuuluu

https://www10.uta.fi/opas/opintojakso.htm?id=30277&lang=fi&lvv=2018&uiLang=fi#parents

2 Osaamistavoitteet

- Opiskelija osaa yksinkertaisia todennäköisyyslaskuja sekä kombinatoriikan alkeet.

Esim. Kuinka todennäköistä on saada täysosuma samalla viikolla sekä lotossa että Eurojackpotissa?

- Hän ymmärtää satunnaismuuttujan ja sen jakauman.

Esim. Nopanheitossa silmäluku, diskreetti satunnaismuuttuja, jakauma diskreetti tasajakauma

Esim. Satunnaisesti väliltä (0, 1) valittu reaaliluku, jatkuva satunnaismuuttuja, jakauma jatkuva tasajakauma

- Hän pystyy yksinkertaisissa tilanteissa määrittämään satunnaismuuttujan jakauman.

Esim. Avainnipussa on 5 avainta, joista yksi on kotiavain. Valitset satunnaisesti yhden. Määritä todennäköisyys sille, että saat kotiavaimen yrityskerralla k. Montako kertaa joudut keskimäärin yrittämään saadaksesi kotiavaimen?

- Hän tuntee odotusarvon ja varianssin ominaisuudet.

Esim. Oletetaan, että sijoituskohteista A ja B keskimääräinen tuotto euron sijoituksesta on μ ja varianssi σ^2. Miten 1000 euroa kannattaa sijoittaa kohteisiin A ja B?

- Opiskelija tuntee binomijakauman ja normaalijakauman ja osaa laskea näihin liittyviä todennäköisyyksiä.

Esim. Kuinka todennäköistä on läpäistä väittämistä koostuvan tentti arvaamalla?

Esim. Lentoyhtiöllä on kone, joka voi ottaa kuljetettavaksi 5000 kg. Voiko yhtiö ottaa kuljetettavakseen 100 lammasta? Aiemmin on ollut punnittuna 1000 vastaavanlaista lammasta, joiden keskipaino on ollut 45 kg ja hajonta 3 kg.

- Opiskelija ymmärtää satunnaisotoksen, otossuureen, otossuureen jakauman sekä otossuureiden käytön tilastollisessa päättelyssä.

Esim. Rattaan keskimääräinen pyörimisaika on 150 s ja keskihajonta 10 s. Onko rasvaaminen vaikuttanut keskimääräiseen pyörimisaikaan? Rasvauksen jälkeen viiden rattaan pyörimisaikojen keskiarvo oli 162 s.

Esim. Pyöritetään rulettia 3400 kertaa ja saadaan 140 nollaa, jolloin pelipaikka voittaa. Voitko todistaa oikeudessa, että pelipaikan ruletti toimii väärin?

Esim. Tuottavatko koneet A ja B keskimäärin samanmittaisia tankoja?

- Opiskelija ymmärtää estimoinnin ja hypoteesien testaukseen liittyvän teorian opintojaksolla esitetyssä laajuudessa.

Esim. Populaatiossa π % viallisia. Miten arvioidaan? Onko arvio luotettava?

Esim. Populaation odotusarvon μ arviointi. Miten arvioidaan? Onko arvio luotettava?

Esim. Tarkastellaan kahden populaation odotusarvoja. Miten arvioidaan niiden yhtäsuuruutta? Onko arvio luotettava?

- Hän tunnistaa erilaiset estimointitilanteet, osaa valita tilanteeseen soveltuvan luottamusvälin sekä käyttää sitä tilastollisessa päättelyssä.

Esim. Puolueen kannatuksen arviointi. Kyselyssä kannattajia 15 %, otoskoko 2000.

Esim. Hillopurkkien keskimääräisen painon arviointi. Satunnaisesti valittujen 25 hillopurkin keskipaino 330 g ja keskihajonta 20 g.

Esim. Tuottavatko koneet A ja B keskimäärin samanmittaisia tankoja? Molempien koneiden tuotannosta valittu satunnaisesti 100 tankoja, joiden keskiarvoiksi saadaan 2,5 cm ja 2,7 cm sekä keskihajonnoiksi 0,005 cm ja 0,006 cm.

- Hän ymmärtää tilastollisen testauksen periaatteet ja osaa suorittaa tilastollisen testauksen annetussa empiirisessä tilanteessa.

Esim. Puolue väittää kannatuksensa olevan eli 18 %. Voitko uskoa väitteen?

Esim. Voidaanko uskoa, että hillopurkit painavat keskimäärin 340 g.

Esim. Tuottavatko koneet A ja B keskimäärin samanmittaisia tankoja?

3 Kurssin kotisivu

https://coursepages.uta.fi/mtttp5/syksy-2018/

- Opetus
- Kurssi-info (sisältö, tentit, harjoitushyvitys)
- Luennot, luentorunko (sis. kaavat, taulukot), luentokalvot
- Harjoitukset, tehtävät, ohjeet (Moodle), ratkaisut
- Usein kysyttyä
- Palaute
- Linkkejä
- Oheiskirjallisuutta

Luku 2
Todennäköisyyslaskentaa

2.1 Satunnaisilmiö ja tapahtuma

Satunnaisilmiö

- useita tulosmahdollisuuksia
- epävarmuutta tuloksesta

Esim. 2.1.2 Rahanheitto, nopanheitto, lottoaminen, vakioveikkaus, kortin vetäminen sekoitetusta pakasta.

Kaikki mahdolliset tulokset muodostavat perusjoukon E.

Esim. 2.1.1

Rahanheitto E = {kruuna, klaava}

Nopanheitto E = {1, 2, 3, 4, 5, 6}

Lottoaminen E = {kaikki mahdolliset lottorivit}, joita on 18643560

Vakioveikkaus E = {kaikki mahdolliset rivit}, joita on 1594323

Tapahtuma on perusjoukon osajoukko.

Esim. 2.1.1

Rahanheitto	A = {saadaan kruuna}
Nopanheitto	A = {saadaan parillinen} = {2, 4, 6}
Lottoaminen	A = {saadaan 7 oikein}
	B = {saadaan 6 oikein ja lisänumero }
	C = {saadaan kaikki väärin}
Vakioveikkaus	A = {saadaan 13 oikein}
	B = {saadaan 12 oikein}

2.2 Klassinen todennäköisyys

Perusjoukossa n tulosta, jotka kaikki yhtä todennäköisiä. Tapahtumaan A liittyviä tuloksia k kappaletta. Tällöin A:n todennäköisyys $P(A) = k/n$.

Esim. 2.2.1

Lottoaminen

$A = \{\text{saadaan 7 oikein}\}$, $P(A) = 1/18643560$

Vakioveikkaus

$A = \{\text{saadaan 13 oikein}\}$, $P(A) = 1/1594323$

2.3 Todennäköisyyslaskennan aksioomat ja laskusäännöt

Todennäköisyys on joukkofunktio P, joka liittää jokaiseen satunnaisilmiön tapahtumaan A reaaliluvun P(A). Tätä kutsutaan tapahtuman A todennäköisyydeksi ja se toteuttaa tietyt aksioomat.

Aksiooma 1 Jos A on mikä tahansa satunnaisilmiön tapahtuma, niin $0 \leq P(A) \leq 1$.

Aksiooma 2 $P(E) = 1$ (varma tapahtuma)

Olkoot A ja B saman satunnaisilmiön tapahtumia.

Määritellään yhdiste

$A \cup B$ = {A tai B tai molemmat tapahtuvat}

ja leikkaus

$A \cap B$ = {A ja B molemmat tapahtuvat}.

A ja B ovat erillisiä, jos ne eivät voi tapahtua samanaikaisesti, siis $A \cap B = \emptyset$.

<u>Aksiooma 3</u> Jos A ja B erillisiä, niin $P(A \cup B) = P(A) + P(B)$.

Esim. 2.3.1 Nopanheitto

A = {saadaan parillinen}, $P(A) = 3/6$,

B = {saadaan ykkönen}, $P(B) = 1/6$,

$A \cap B = \emptyset$, joten $P(A \cup B) = P(A) + P(B)$

<u>Laskusääntö 1</u> $P(\emptyset) = 0$.

Tapahtuman A komplementti A^C = {A ei tapahdu}.

<u>Laskusääntö 2</u> $P(A^C) = 1 – P(A)$.

Esim. 2.3.3 Vakioveikkaus

A = {korkeintaan 11 oikein}

$P(A) = 1 – P(A^C)$ = 1 - P{12 tai 13 oikein}

= 1 – (P{12 oikein} + P{13 oikein})

Esim. Heität noppaa kunnes saat numeron 6. Laske todennäköisyys sille, että joudut heittämään ainakin 3 kertaa. Tällöin joudut heittämään 3 tai 4 tai 5 tai ... kertaa. Todennäköisyys lasketaan komplementin kautta,

1 – P{heittokertoja 1 tai 2}

= 1 – (P{heittokertoja 1} +P{heittokertoja 2})

= 1 – P{heittokertoja 1} - P{heittokertoja 2}

$= 1 - 1/6 - (5/6)(1/6) = 25/36$

<u>Laskusääntö 3</u> Jos tapahtumat A_1, A_2, ..., A_k ovat pareittain erilliset, niin

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k).$$

Esim. 2.3.4 Kortin vetäminen korttipakasta
A = {saadaan ruutu}, B = {saadaan hertta}, C = {saadaan risti}. P(saadaan ruutu tai hertta tai risti) = P(A)+P(B)+P(C) = 39/52.

Laskusääntö 4 (yleinen yhteenlaskusääntö)

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

Esim. 2.3.5 Kortin vetäminen korttipakasta

P{kortti pata tai ässä} = P{kortti pata} + P{kortti ässä} - P{kortti pataässä} = 13/52 + 4/52 – 1/52

MTTTP5, luento 25.10.2018

Kertausta

- Satunnaisilmiö (satunnaiskoe), voi syntyä myös useassa eri vaiheessa (yhdistetty satunnaisilmiö)
- Perusjoukko (otosavaruus) E
- Tapahtumat A, B, ... perusjoukon osajoukkoja
- $P(A) = k/n$

  k = tapahtumaan A liittyvien tulosten lukumäärä

  n = kaikki mahdolliset tulokset
- $A \cup B$ = {A tai B tai molemmat tapahtuvat}

- $A \cap B = \{A \text{ ja } B \text{ molemmat tapahtuvat}\}$
- A ja B erillisiä, $A \cap B = \emptyset$
- $0 \leq P(A) \leq 1$, aksiooma 1
- $P(E) = 1$, aksiooma 2
- Jos $A \cap B = \emptyset$, niin $P(A \cup B) = P(A) + P(B)$, aksiooma 3
- $P(\emptyset) = 0$, laskusääntö 1
- $P(A^C) = 1 - P(A)$, laskusääntö 2
- $P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$, kun tapahtumat pareittain erilliset, laskusääntö 3
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, laskusääntö 4 (yleinen yhteenlaskusääntö)

2.3 Todennäköisyyslaskennan aksioomat ja laskusäännöt (jatkoa)

Esim. 2.3.4 Kortin vetäminen korttipakasta

A = {saadaan ruutu}
B = {saadaan hertta}
C = {saadaan risti}

P(saadaan ruutu tai hertta tai risti)
= P(A)+P(B)+P(C) = 13/52 + 13/52 + 13/52 = 39/52.

Esim. 2.3.5 Kortin vetäminen korttipakasta

P{kortti pata tai ässä}
= P{kortti pata} + P{kortti ässä} - P{kortti pataässä}
= 13/52 + 4/52 – 1/52

Ehdollinen todennäköisyys $P(A|B) = P(A \cap B)/P(B)$.

Esim. 2.3.6 On saatu nopanheitossa pariton silmäluku. Mikä on silmäluvun 5 todennäköisyys? $A = \{5\}$, $B = \{1, 3, 5\}$, $P(A|B) = P(A \cap B)/P(B) = (1/6)/(3/6) = 1/3$.

Laskusääntö 5 (yleinen kertolaskusääntö)

Jos $P(B) > 0$, niin

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

Jos A ja B riippumattomia, niin $P(A \cap B) = P(A)P(B)$

Yleistäen: Jos tapahtumat $A_1, A_2, \ldots, A_k$ ovat riippumattomia, niin $P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)P(A_2) \ldots P(A_k)$.

Esim. 2.3.8. Heitetään noppaa kolme kertaa

A = {1. heiton silmäluku pariton},

B = {2. heiton silmäluku pariton},

C = {3. heiton silmäluku pariton}

P{kaikilla heitoilla pariton}

$= P(A)P(B)P(C) = (1/2)(1/2)(1/2) = 1/8$

Esim. Laatikossa on neljä palloa, joista kaksi on mustaa ja kaksi valkoista. Poimitaan satunnaisesti kaksi palloa palauttaen.

P{molemmat pallot valkoisia}

= P{1. pallo valkoinen ja 2. pallo valkoinen}

= P{1. pallo valkoinen}P{2. pallo valkoinen}

= (2/4)(2/4) = ¼

Jos poiminta tehdään palauttamatta, niin

P{molemmat pallot valkoisia}

= P{1. pallo valkoinen}P{2. pallo valkoinen}

= (2/4)(1/3) = 1/6.

Esim. Heität noppaa kunnes saat numeron 6.

P{joudut heittämään ainakin 3 kertaa}

= 1 – P{heittokertoja 1 tai 2}

= 1 – (P{heittokertoja 1} +P{heittokertoja 2})

= 1 – 1/6 – (5/6)(1/6) = 25/36

Voi laskea myös todennäköisyyden, että kahdella ensimmäisellä kerralla ei saada numeroa 6.
Tämä on (5/6)(5/6) = 25/36.

2.4 Kombinatoriikka

Yhdistetyn satunnaisilmiön tulosmahdollisuuksien lukumäärä $n_1 n_2 \ldots n_K$.

Esim. 2.4.1 Vakioveikkauksessa rivien lukumäärä 3^{13} = 1 594 323. Rivejä, joissa ei yhtään oikein $2^{13} = 8192$.

Esim. 2.4.2 Henkilöt A, B ja C voidaan asettaa $3 \cdot 2 \cdot 1 = 6$ erilaiseen jonoon.

Kuinka moneen eri järjestykseen n erilaista alkiota voidaan järjestää? Järjestyksiä eli permutaatioita on $n(n-1)(n-2)\ldots\cdot 2 \cdot 1 = n!$ (n-kertoma)

Esim. 2.4.4 Moneenko erilaiseen jonoon 5 henkilöä voidaan asettaa? Entä kymmenen?

Esim. 2.4.5 Kuinka moneen eri järjestykseen korttipakan 52 korttia voidaan asettaa?
Laskuri http://stattrek.com/online-calculator/combinations-permutations.aspx

Olkoon n erilaista alkiota. Tällöin k:n alkion osajoukkoja eli kombinaatioita voidaan muodostaa

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

(binomikerroin)

Esim. 2.4.3 Erilaisia lottorivejä

$$\binom{40}{7} = \frac{40!}{7!\,(40-7)!} = 18\,643\,560$$

Laskuri http://stattrek.com/online-calculator/combinations-permutations.aspx

Sellaisia lottorivejä, jossa kaikki väärin

$$\binom{33}{7} = \frac{33!}{7!\,(33-7)!} = 4\,272\,048$$

Sellaisia lottorivejä, joissa *k* oikein

$$\binom{7}{k}\binom{40-7}{7-k}$$

Sellaista vakioveikkausriviä, joissa *k* oikein

$$\binom{13}{k} \cdot 2^{13-k}$$

Esim. Laske todennäköisyys sille, että lottorivissä on vähintään kuusi oikein.

P(vähintään 6 oikein)

= P(kuusi oikein tai 7 oikein)

= P(6 oikein) + P(7 oikein)

$$= \frac{\binom{7}{6}\binom{40-7}{7-6}+1}{\binom{40}{7}} = \frac{231+1}{18643560}$$

Esim. 2.4.6 Kahden alkion satunnaisotokset luvuista 1, 2, 3, 4, 5, 6 satunnaisotokset ja niiden suurimmat alkiot

Otos	Max
1, 2	2
1, 3	3
1, 4	4
1, 5	5
1, 6	6
2, 3	3
2, 4	4
2, 5	5
2, 6	6
3, 4	4
3, 5	5
3, 6	6
4, 5	5
4, 6	6
5, 6	6

$P(Max = 2) = 1/15$
$P(Max = 3) = 2/15$
$P(Max = 4) = 3/15$
$P(Max = 5) = 4/15$
$P(Max = 6) = 5/15$

Esim. 2.4.7 Luvuista 1, 2, 3, 4, 5, 6 kahden alkion systemaattisella otannalla tehdyt otokset ja niiden suurimmat alkiot

Otos	Max	
1, 4	4	P(Max = 4) = 1/3
2, 5	5	P(Max = 5) = 1/3
3, 6	6	P(Max = 6) = 1/3

MTTTP5, luento 30.10.2018

Luku 3
Todennäköisyysjakaumia

3.1 Satunnaismuuttuja ja todennäköisyysjakauma

Esim. 3.1.1 Satunnaisilmiö nopanheitto, satunnaismuuttuja X = saatu silmäluku,

$P(X = 1) = P(X = 2) = ... = P(X = 6) = 1/6$

Esim. 3.1.2 Satunnaisilmiö neljän kolikon heitto, satunnaismuuttuja X = klaavojen lukumäärä heittosarjassa,

$P(X = 0) = 1/16$, $P(X = 1) = 4/16$, $P(X = 2) = 6/16$, $P(X = 3) = 4/16$, $P(X = 4) = 1/16$

- Satunnaismuuttuja X on funktio, joka liittää yksikäsitteisen reaaliluvun jokaiseen tarkasteltavan satunnaisilmiön perusjoukon tulokseen.
- Tarkastellaan eri tulosten arvojen todennäköisyyksiä, jolloin saadaan satunnaismuuttujan todennäköisyysjakauma.
- Satunnaismuuttuja voi olla jatkuva tai diskreetti.
- Funktiota, joka määrittää satunnaismuuttujan todennäköisyysjakauman, kutsutaan tiheysfunktioksi, merk. $f(x)$.
- Satunnaismuuttujan X kertymäfunktio $F(x) = P(X \leq x)$.

Esim. Vakioveikkaus

X = oikein veikattujen kohteiden lukumäärä

$F(0) = P(X \le 0) = P(X = 0)$

$= 2^{13}/3^{13} = 8192/1594323$

$F(5) = P(X \le 5)$

$= P(X = 0) + P(X = 1) + ... + P(X = 5)$

Kertymäfunktion ominaisuuksia

- $F(-\infty) = 0$, $F(\infty) = 1$
- $P(a < X \leq b) = F(b) - F(a)$, $a < b$
- Jos X jatkuva, niin $P(X \leq a) = P(X < a) = F(a)$
- $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
- Jos X jatkuva, niin $F'(x) = f(x)$

3.2 Diskreetti satunnaismuuttuja

Olkoon diskreetin satunnaismuuttujan X arvot x_1, x_2, ..., ja näiden arvojen todennäköisyydet p_1, p_2, ...

Satunnaismuuttujan X todennäköisyysjakauma määritellään pistetodennäköisyyksien

$$P(X = x_i) = \begin{cases} p_i, & i = 1, 2, \ldots, \text{missä} \quad p_1 + p_2 + \cdots = 1 \\ 0, & \text{muulloin} \end{cases}$$

perusteella.

Määritellään odotusarvo

$$E(X) = p_1x_1 + p_2x_2 + \ldots + p_kx_k = \mu$$

ja varianssi

$$\begin{aligned} Var(X) &= E(X - \mu)^2 \\ &= p_1(x_1 - \mu)^2 + p_2(x_2 - \mu)^2 + \ldots + p_k(x_k - \mu)^2 \\ &= \sigma^2 \end{aligned}$$

Varianssin neliöjuuri σ on keskihajonta.

Esim. 3.2.2 Nopanheitto, satunnaismuuttuja X = saatu silmäluku, X:n todennäköisyysjakauma

$P(X = 1) = P(X = 2) = \ldots = P(X = 6) = 1/6$

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3{,}5$$

$$Var(X) = (1 - 3{,}5)^2 \cdot \frac{1}{6} + \cdots + (6 - 3{,}5)^2 \cdot \frac{1}{6} = \frac{35}{12}$$

Esim. X = klaavojen lukumäärä neljän kolikon heitossa

X:n todennäköisyysjakauma:

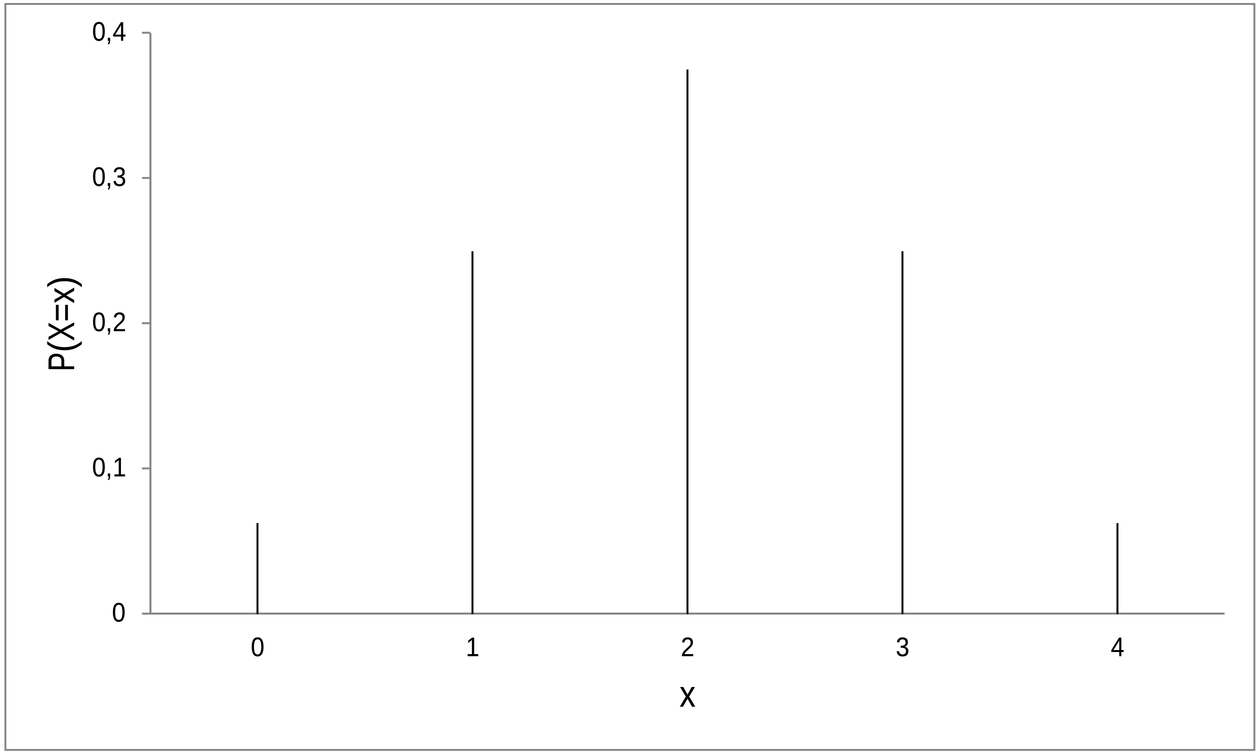
P(X = 0) = 1/16, P(X = 1) = 4/16, P(X = 2) = 6/16,

P(X = 3) = 4/16, P(X= 4) = 1/16

$E(X) = 0 \cdot \frac{1}{16} + 1 \cdot \frac{4}{16} + \cdots + 4 \cdot \frac{1}{16} = 2$

$Var(X) = (0 - 2)^2 \cdot \frac{1}{16} + (1 - 2)^2 \cdot \frac{4}{16} + \cdots + (4 - 2)^2 \cdot \frac{1}{16} = 1$

Jakauma graafisesti:



Esim. Satunnaiskokeessa onnistutaan todennäköisyydellä p ja epäonnistutaan todennäköisyydellä 1 – p. Määritellään satunnaismuuttuja

$$X = \begin{cases} 1, & \text{jos onnistutaan} \\ 0, & \text{jos epäonnistutaan} \end{cases}$$

Siis P(X = 1) = p, P(X = 0) = 1 – p.

$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$

$Var(X) = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p)$

Vrt. esim. 3.2.3.

3.3 Jatkuva satunnaismuuttuja

Olkoon f(x) jatkuvan satunnaismuuttujan X tiheysfunktio. Tällöin f(x)$\geq$ 0 sekä f(x):n ja x-akselin väliin jäävä pinta-ala on yksi.

Määritellään X:n odotusarvo

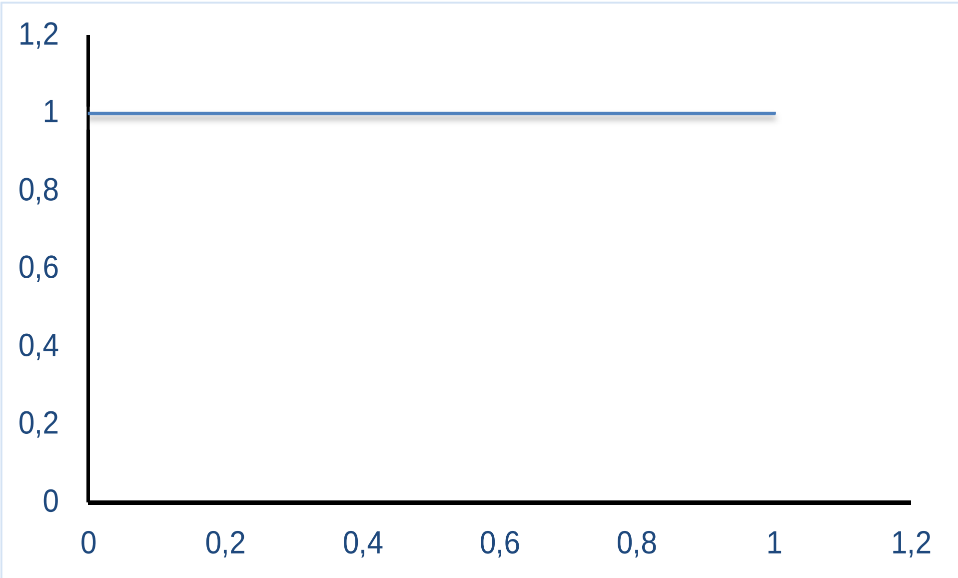
$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \mu$$

ja varianssi

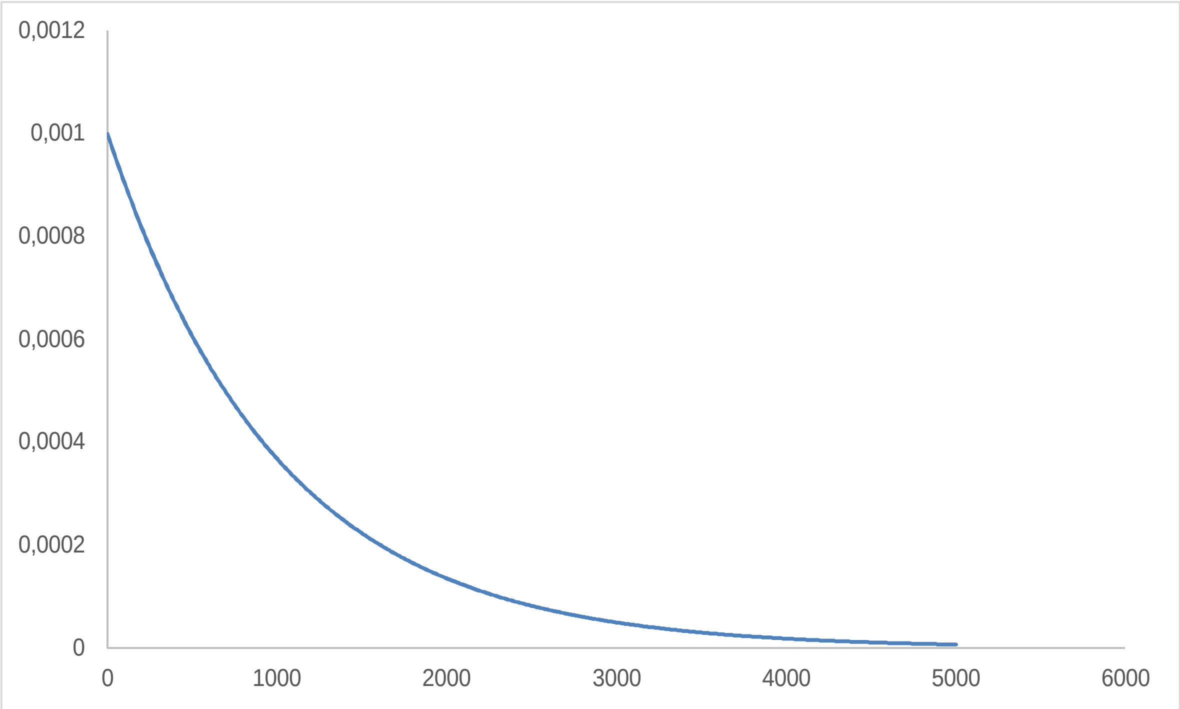
$$Var(X) = E(x - \mu)^2 = \sigma^2$$ ks. kaava (3.4)

X:n keskihajonta on σ.

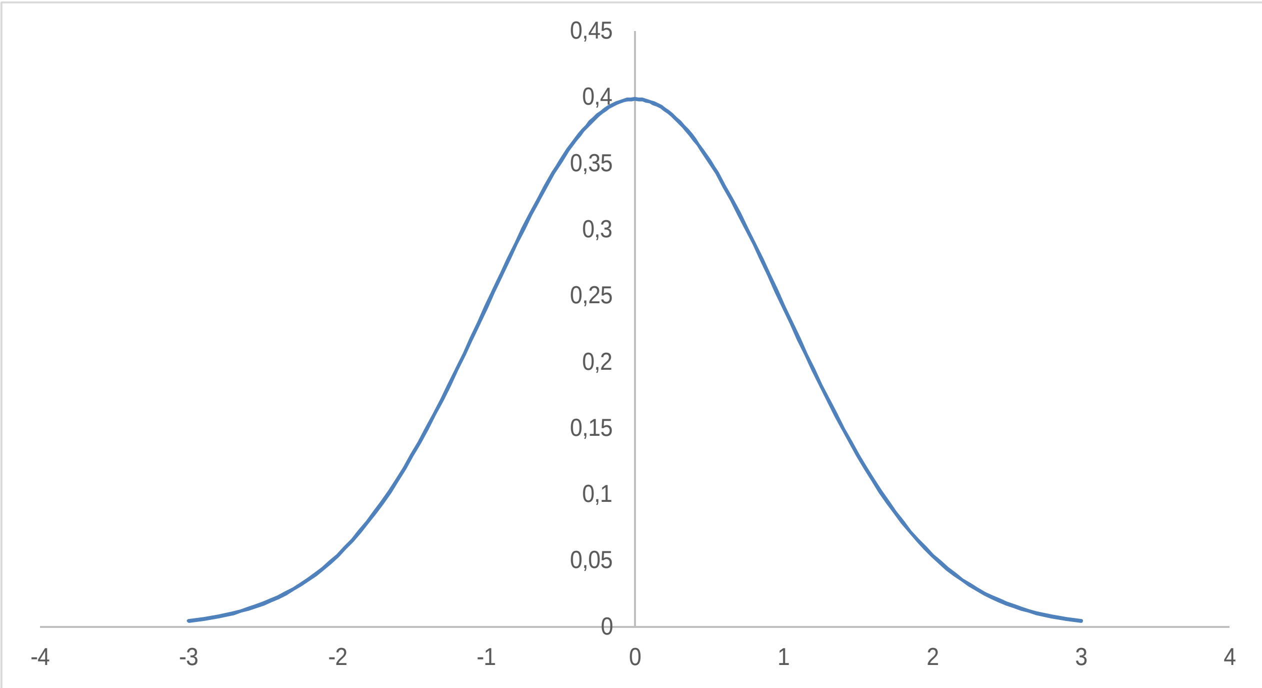
Esim. $f(x) = 1$, kun $0 \leq x \leq 1$



Esim. $f(x) = 0{,}001e^{-0{,}001x}, \; x \geq 0$

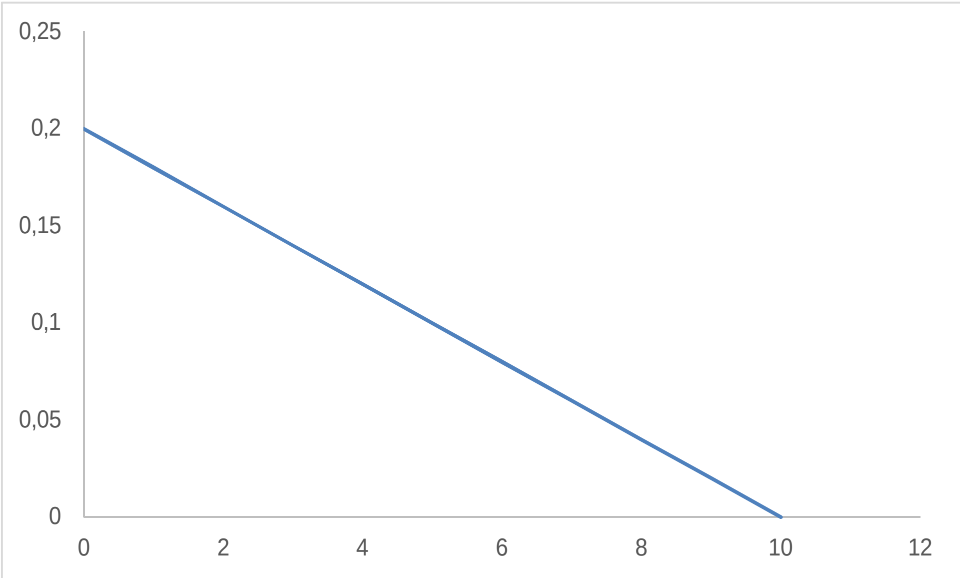


Esim. $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

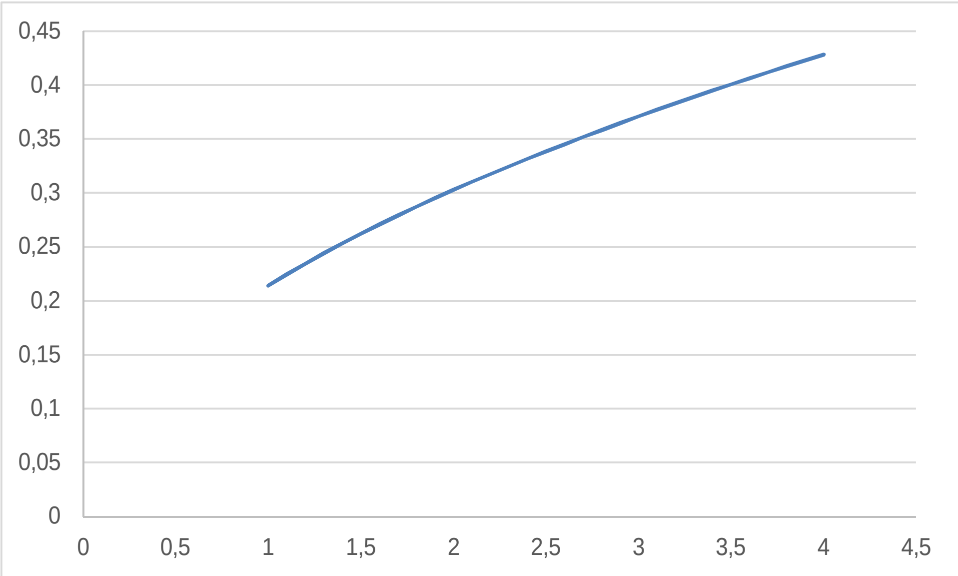


Kyseessä standardoidun normaalijakauman tiheysfunktio (ks. luentomoniste s. 22)

Esim. $f(x) = 0{,}02(10-x),\ 0 \leq x \leq 10$



Esim. $f(x) = \frac{3}{14}\sqrt{x}$, kun $1 \le x \le 4$.



Olkoot a ja b reaalilukuja ($a \leq b$). Tällöin

- $P(X \leq a) = F(a)$
- $P(X \geq a) = 1 - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a)$

Esim. $f(x) = 1/4$, kun $0 \leq x \leq 4$

$F(x) = P(X \leq x) = x/4$ (suorakulmion pinta-alana)

$P(2 \leq X \leq 3) = F(3) - F(2) = 3/4 - 2/4 = 1/4$

$P(X \geq 1) = 1 - F(1) = 1 - 1/4 = 3/4$

MTTTP5, luento 1.11.2018

Kertausta

- X diskreetti satunnaismuuttuja

$$P(X = x_i) = p_i, i = 1,2,\ldots,k, \sum_{i=1}^{k} p_i = 1$$

- X jatkuva satunnaismuuttuja

f(x) tiheysfunktio, f(x) ≥ 0, $\int_{-\infty}^{\infty} f(x)dx = 1$

- Kertymäfunktio $F(x) = P(X \leq x)$

$P(a < X \leq b) = F(b) - F(a)$, $a < b$

Jos X jatkuva, niin $P(X \leq a) = P(X < a) = F(a)$

$P(X > a) = 1 - P(X \leq a) = 1 - F(a)$

- Odotusarvo

$$\mu = E(X) = \begin{cases} \sum_{i=1}^{k} p_i x_i \text{ , } kun\ X\ diskreetti \\ \int_{-\infty}^{\infty} x f(x) dx \text{ , } kun\ X\ jatkuva \end{cases}$$

- Varianssi

$$\sigma^2 = Var(X) = E(X - \mu)^2 = \begin{cases} \sum_{i=1}^{k} p_i (x_i - \mu)^2 \text{ , } kun\ X\ diskreetti \\ \int_{-\infty}^{\infty} f(x)(x - \mu)^2 dx, kun\ X\ jatkuva \end{cases}$$

Esim.

Satunnaismuuttujan X mahdolliset arvot ovat 1, 2 ja 3 sekä

P(X = 1) = 1-2p, P(X = 2) = P(X = 3) = p, 0≤ p ≤½.

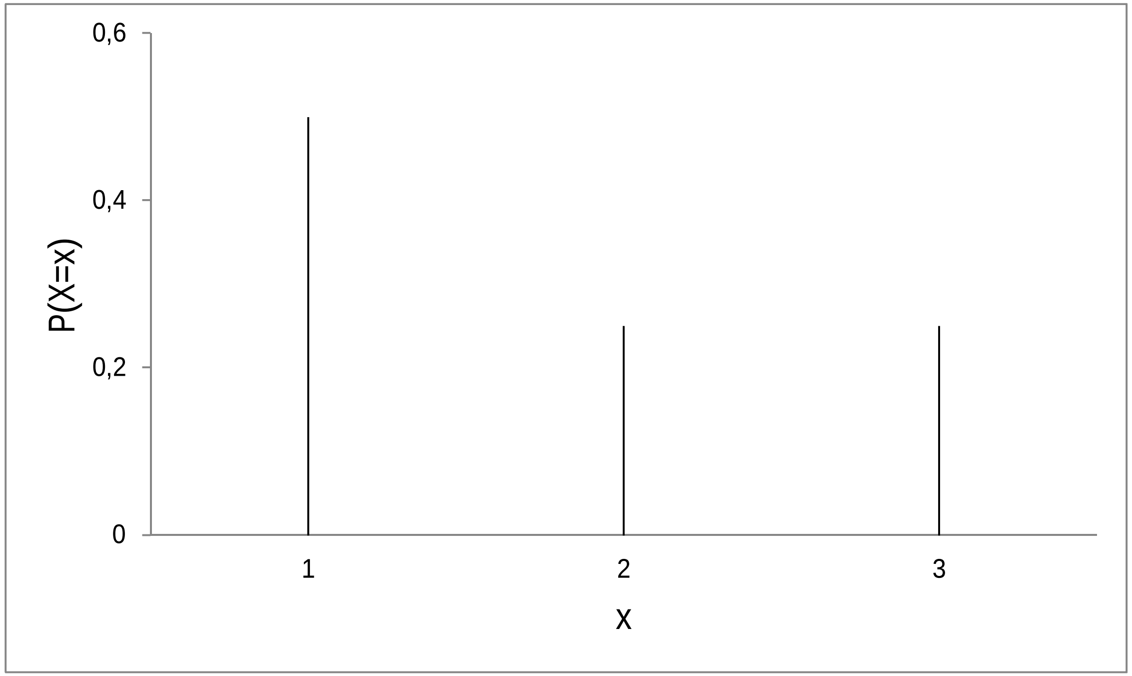
Laske E(X), Var(X). Piirrä X:n todennäköisyysjakauman kuvaaja, kun p = 0,25.

$$E(X) = 1 \cdot (1 - 2p) + 2 \cdot p + 3 \cdot p = 3p + 1$$

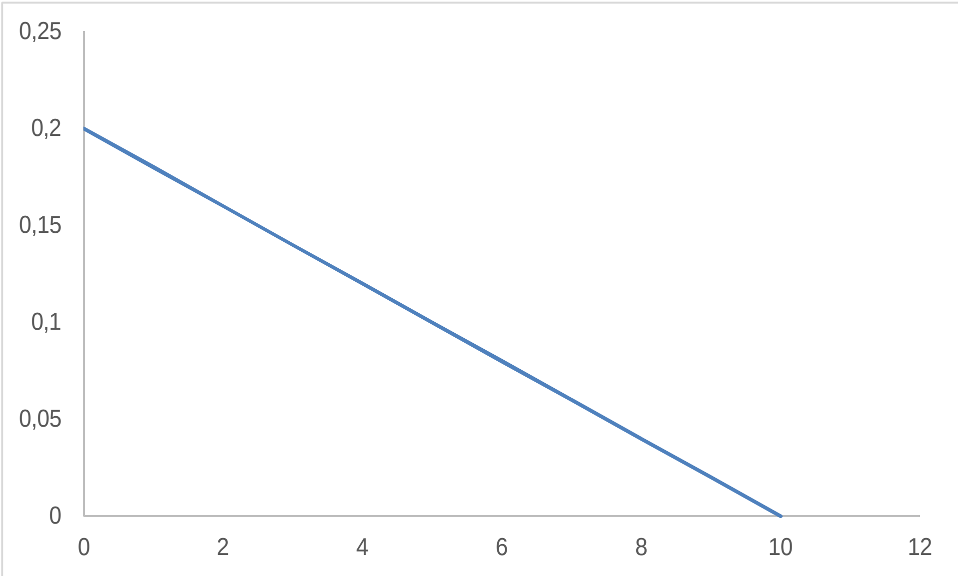
$$\mathrm{Var}(X) = (1 - 3p - 1)^2 \cdot (1 - 2p) + (2 - 3p - 1)^2 \cdot p$$

$$+ (3 - 3p - 1)^2 \cdot p = \cdots = -9p^2 + 5p$$

Jos p = 0,25, niin P(X = 1) = 0,5, P(X = 2) = P(X = 3) = 0,25.
Graafisesti:



Esim. Henkilö A saapuu bussipysäkille. Hän joutuu mahdollisesti odottamaan bussia. Määritellään X = odotusaika minuutteina. Oletetaan, että X:n tiheysfunktio on $f(x) = 0{,}02(10-x)$, $0 \leq x \leq 10$. Tiheysfunktion kuvaaja on laskeva suora, kaksi pistettä suoralta $f(0)=0{,}2$, $f(10)=0$.



Kertymäfunktio $F(x) = P(X \leq x)$. Tämä voidaan laskea kolmion pinta-alaa hyödyntäen

$$
\begin{aligned}
F(x) &= P(X \leq x) \\
&= 1 - P(X > x) \\
&= 1 - (10-x)\, f(10-x)/2 \\
&= -0{,}01x^2 + 0{,}2x.
\end{aligned}
$$

Todennäköisyys sille, että A joutuu odottamaan yli 9 minuuttia, on

$$P(X > 9) = 1 - P(X \leq 9) = 1 - F(9) = 0{,}01.$$

ks. http://www.sis.uta.fi/tilasto/mtttp2/syksy2013/luentoesimerkki_31_10_13.pdf

Jos E(X) = μ ja Var(X) = σ^2, niin X standardoidaan tekemällä muunnos

$$Z = \frac{X - \mu}{\sigma}$$

3.4 Odotusarvon ja varianssin ominaisuuksia

Odotusarvon ominaisuuksia

- $E(a) = a, a\ vakio$
- $E(aX + b) = aE(X) + b, a\ ja\ b\ vakioita$
- $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$

Satunnaismuuttujien riippumattomuus määritellään vastaavalla tavalla kuin tapahtumien riippumattomuus.

Varianssin ominaisuuksia

- $Var(a) = 0, a\ vakio$
- $Var(aX + b) = a^2 Var(X), a\ ja\ b\ vakioita$
- $Hajonta: Sd(aX + b) = |a| Sd(X), a\ ja\ b\ vakioita$
- $Jos\ X_1, X_2, \ldots, X_n\ riippumattomia, niin\ Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$

Esim. 3.4.4 Sijoitat 1000 euroa. Mahdolliset kohteet A ja B, joissa molemmissa pienin sijoitusmäärä 500 euroa.

Merkitään X = tuotto 100 euron sijoituksesta kohteeseen A,
Y = tuotto 100 euron sijoituksesta kohteeseen B

Oletetaan $P(X = -5) = 0,4$, $P(X = 20) = 0,6$
$P(Y = 0) = 0,6$, $P(Y = 25) = 0,4$.

$E(X) = -5 \cdot 0,4 + 20 \cdot 0,6 = 10$
$E(Y) = 0 \cdot 0,6 + 25 \cdot 0,4 = 10$

$Var(X) = (-5 - 10)^2 \cdot 0,4 + (20 - 10)^2 \cdot 0,6 = 150$
$Var(Y) = (0 - 10)^2 \cdot 0,6 + (25 - 10)^2 \cdot 0,4 = 150$

Miten sijoitat?

Mahdolliset vaihtoehdot ja niiden tuotot W

1. 1000 euroa A:han, $W = 10X$

   $E(W) = 10E(X) = 100$, $Var(W) = 10^2 Var(X) = 15000$

2. 1000 euroa B:hen, $W = 10Y$

   $E(W) = 10E(Y) = 100$, $Var(W) = 10^2 Var(Y) = 15000$

3. 500 euroa kumpaankin, $W = 5X + 5Y$

   $E(W) = E(5X + 5Y) = E(5X) + E(5Y) = 100$

   $Var(W) = Var(5X + 5Y) = Var(5X) + Var(5Y) = 5^2 Var(X) + 5^2 Var(Y) = 7500$

Vaihtoehto 3 paras, koska pienin vaihtelu (riski).

Esim. Sijoitat kohteeseen A 500 euroa ja kohteeseen B 1000 euroa.

Olk. X = tuotto euron sijoituksesta kohteeseen A

Y = tuotto euron sijoituksesta kohteeseen B.

Oletetaan tuottojen olevan toisistaan riippumattomia ja $E(X) = E(Y) = \mu$, $Var(X) = Var(Y) = \sigma^2$. Määritä koko 1500 euron tuoton odotusarvo ja varianssi.

Kokonaistuotto $W = 500X + 1000Y$

$E(W) = E(500X + 1000Y) = E(500X) + E(1000Y) = 500E(X) + 1000E(Y) = 500\mu + 1000\mu = 1500\mu$

$Var(W) = Var(500X + 1000Y) = Var(500X) + Var(1000Y) = 500^2Var(X) + 1000^2Var(Y) = 500^2\sigma^2 + 1000^2\sigma^2 = 1250000\sigma^2$

Esim. 3.4.5

Sijoitat 1000 euroa. Mahdolliset kohteet A ja B.

X = 1 euron tuotto sijoituksesta kohteeseen A

Y = 1 euron tuotto sijoituksesta kohteeseen B

X ja Y riippumattomia

$E(X) = E(Y) = \mu$, $Var(X) = Var(Y) = \sigma^2$.

Miten sijoitat?

Sijoitetaan α euroa kohteeseen A ja $(1000-\alpha)$ kohteeseen B

Tuotto $W = \alpha X + (1000-\alpha)Y$

$$E(W) = E(\alpha X + (1000-\alpha)Y)$$

$$= E(\alpha X) + E((1000-\alpha)Y) = \alpha E(X) + (1000-\alpha)E(Y)$$

$$= 1000\mu$$

$$Var(W) = Var(\alpha X + (1000-\alpha)Y)$$

$$= Var(\alpha X) + Var((1000-\alpha)Y)$$

$$= \alpha^2 Var(X) + (1000-\alpha)^2 Var(Y)$$

$$= (2\alpha^2 - 2000\alpha + 1000000)\sigma^2$$

Pienin varianssi, kun $\alpha = 500$

Esim. 3.4.1

$$E(X) = \mu, Var(X) = \sigma^2, Z = \frac{X - \mu}{\sigma}, E(Z) = 0, Var(Z) = 1.$$

Esim. 3.4.2 Olkoot X ja Y riippumattomia,

$E(X) = \mu_X$, $E(Y) = \mu_Y$, $Sd(X) = \sigma_X$, $Sd(Y) = \sigma_Y$.

Määritellään $Z = X – Y$.

$E(Z) = E(X – Y) = E(X) – E(Y) = \mu_X - \mu_Y$

$Var(Z) = Var(X – Y) = Var(X) + Var(-Y)$

$= Var(X) + (-1)^2 Var(Y) = \sigma_X^2 + \sigma_Y^2$

Z:n hajonta $\sqrt{\sigma_X^2 + \sigma_Y^2}$

Esim. 3.4.3

$Olkoot\ X_1, X_2, \ldots, X_n\ riippumattomia\ ja$

$$E(X_i) = \mu, Var(X_i) = \sigma^2, m\text{ää}ritell\text{ää}n\ Y = \frac{X_1 + X_2 + \cdots X_n}{n}.$$

$$E(Y) = \mu, Var(Y) = \frac{\sigma^2}{n}$$

MTTTP5, luento 6.11.2018

3.5 Joitain todennäköisyysjakaumia

3.5.1 Bernoulli-jakauma

Tarkastellaan satunnaisilmiötä, jossa joko onnistutaan (A) tai epäonnistutaan (A^C). Määritellään satunnaismuuttuja X siten, että

$$X = \begin{cases} 1, & \textit{jos onnistutaan} \\ 0, & \textit{jos epäonnistutaan} \end{cases}$$

Olkoon lisäksi P(A) = P(X = 1) = p, P(A^C) = P(X = 0) = 1 - p.

Sanotaan, että X noudattaa Bernoulli-jakaumaa parametrilla p, merk. X ~ Ber(p).

Tällöin E(X) = p, Var(X) = p(1 - p), ks. luento 30.10.

Esim. 3.5.1 Rahanheitto, veikkauksessa yhden kohteen arvaaminen, nopanheitossa silmäluvun 6 saaminen.

Esim. Heitetään noppaa. Määritellään X siten, että

$$X = \begin{cases} 1, & jos\ silmäluku\ 6 \\ 0, & muulloin \end{cases}$$

Siis P(X = 1) = 1/6, P(X = 0) = 5/6, X ~ Ber(1/6).

E(X) = $\frac{1}{6}$, Var(X) = $\frac{1}{6}\left(1 - \frac{1}{6}\right)$

3.5.2 Binomijakauma

Toistetaan n kertaa satunnaisilmiötä, jossa onnistumisen todennäköisyys on p.

Määritellään X = onnistumisten kokonaislukumäärä.

Tällöin sanotaan, että X noudattaa binomijakaumaa parametrein n ja p, merk. X ~ Bin(n, p).

Jos X ~Bin(n, p), niin

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

Graafisesti http://vassarstats.net/

Binomijakautunut satunnaismuuttuja

$X = X_1 + X_2 + ... + X_n$, missä $X_i \sim Ber(p)$

$E(X_i) = p$, $Var(X_i) = p(1-p)$

$$E(X) = E(X_1 + X_2 + ... + X_n)$$
$$= E(X_1) + E(X_2) + ... + E(X_n) = np$$

$$Var(X) = Var(X_1 + X_2 + ... + X_n)$$
$$= Var(X_1) + Var(X_2) + ... + Var(X_n) = np(1 - p)$$

Esim. Heitetään noppaa 10 kertaa. Määritellään
X = silmäluvun 6 kokonaislukumäärä heittosarjassa

X ~ Bin(10, 1/6), $E(X) = 10 \cdot \frac{1}{6}$, $Var(X) = 10 \cdot \frac{1}{6} \cdot \frac{5}{6}$

$$P(X = k) = \binom{10}{k}\left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{10-k}, k = 0, 1, 2, \ldots 10$$

k	P(X = k)
0	0,161506
1	0,323011
2	0,29071
3	0,155045
4	0,054266
5	0,013024
6	0,002171
7	0,000248
8	1,86E-05
9	8,27E-07
10	1,65E-08

Kuvaaja: http://homepage.stat.uiowa.edu/~mbognar/applets/bin.html

Esim. 3.5.2 Vakioveikkaus, X = oikein arvattujen kohteiden kokonaislukumäärä, X ~Bin(13, 1/3), E(X) = $13 \cdot \frac{1}{3}$, Var(X) = $13 \cdot \frac{1}{3} \cdot \frac{2}{3}$

$$P(X = k) = \binom{13}{k}\left(\frac{1}{3}\right)^{k}\left(1 - \frac{1}{3}\right)^{13-k}, k = 0, 1, 2, \ldots 13$$

$$P(X = 0) = \binom{13}{0}\left(\frac{1}{3}\right)^{0}\left(\frac{2}{3}\right)^{13-0} = \left(\frac{2}{3}\right)^{13}$$

$$P(X = 1) = \binom{13}{1}\left(\frac{1}{3}\right)^{1}\left(\frac{2}{3}\right)^{13-1} = 13\left(\frac{1}{3}\right)^{1}\left(\frac{2}{3}\right)^{12}$$

$$P(X = 2) = \binom{13}{2}\left(\frac{1}{3}\right)^{2}\left(\frac{2}{3}\right)^{13-2} = 78\left(\frac{1}{3}\right)^{2}\left(\frac{2}{3}\right)^{11}$$

$$P(X = 3) = \binom{13}{3}\left(\frac{1}{3}\right)^{3}\left(\frac{2}{3}\right)^{13-3} = 286\left(\frac{1}{3}\right)^{3}\left(\frac{2}{3}\right)^{10}$$

$$P(X > 3) = 1 - P(X \le 3)$$

$$= 1 - (P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3))$$

$$\approx 0{,}6776$$

$$P(X = 12) = \binom{13}{12}\left(\frac{1}{3}\right)^{12}\left(\frac{2}{3}\right)^{13-12} = 13\left(\frac{1}{3}\right)^{12}\left(\frac{2}{3}\right)$$

$$P(X = 13) = \binom{13}{13}\left(\frac{1}{3}\right)^{13}\left(\frac{2}{3}\right)^{13-13} = \left(\frac{1}{3}\right)^{13}$$

$$P(X > 11) = P(X = 12) + P(X = 13)$$

$$= 27 \cdot \left(\frac{1}{3}\right)^{13} \approx 0{,}0000169$$

Esim. 3.5.3 Pelaat peliä, jossa heitetään rahaa. Jos tulee klaava, saat ystävältäsi euron, jos tulee kruuna, annat ystävällesi euron. Rahaa on heitetty 20 kertaa, ja olet tappiolla 14 euroa. Onko raha harhaton?

X = klaavojen lukumäärä heittosarjassa, X ~Bin(20, ½)

$$P(X = k) = \binom{20}{k}\left(\frac{1}{2}\right)^{k}\left(1 - \frac{1}{2}\right)^{20-k} = \binom{20}{k}\left(\frac{1}{2}\right)^{20},$$

$$k = 0, 1, 2, \ldots, 20$$

X – (20 –X) = -14, josta X = 3. Klaavoja on tullut 3.

$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$$= \left[\binom{20}{0} + \binom{20}{1} + \binom{20}{2} + \binom{20}{3}\right]\left(\frac{1}{2}\right)^{20} = 1351\left(\frac{1}{2}\right)^{20} \approx 0{,}0013$$

Tämä on harvinaista, joten voidaan päätellä, että raha ei ole harhaton.

Ks. laskuri http://stattrek.com/online-calculator/binomial.aspx

3.5.3 Diskreetti tasajakauma

Esim. Nopanheitossa X = silmäluku, P(X = 1) = P(X = 2) =...= P(X = 6) = 1/6. Sanotaan, että X noudattaa diskreettiä tasajakaumaa välillä (1, 6), merk. X ~ Tasd(1, 6).

Olkoon satunnaismuuttujan X arvot

a, a + 1, a + 2, ..., a + (n-1) = b ja kukin n:stä arvosta yhtä todennäköinen. Sanotaan, että X noudattaa diskreettiä tasajakaumaa välillä (a, b), merk. X ~ Tasd(a, b).

Tällöin $E(X) = \frac{a+b}{2}, Var(X) = \frac{n^2-1}{12}$.

Esim. 3.5.4 Olkoon X yksinumeroinen satunnaisluku. Mahdolliset arvot 0, 1, 2, ..., 9 sekä jokaisen arvon todennäköisyys 1/10, siis $X \sim \text{Tasd}(0, 9)$. $E(X) = (0+9)/2$, $\text{Var}(X) = (10^2-1)/12$.

3.5.4 Jatkuva tasajakauma

Satunnaismuuttuja X noudattaa jatkuvaa tasajakaumaa välillä (a, b), jos sen tiheysfunktio on

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

Merk. X ~ Tas(a, b)

Tällöin $E(X) = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12}$.

Esim. $X \sim \mathrm{Tas}(1, 3)$

$f(x) = 1/(3-1) = 1/2$

$F(x) = P(X \leq x) = (x-1)/2$

$E(X) = (1+3)/2 = 2$

$\mathrm{Var}(X) = (3-1)^2/12 = 1/3$

Esim. $X \sim \mathrm{Tas}(0, 1)$, luento 30.10.

$X \sim \mathrm{Tas}(0, 4)$, luento 30.10.

$X \sim \mathrm{Tas}(-1, 1)$, harj. 2 teht. 2.

MTTTP5, luento 8.11.2018

3.5.5 Normaalijakauma

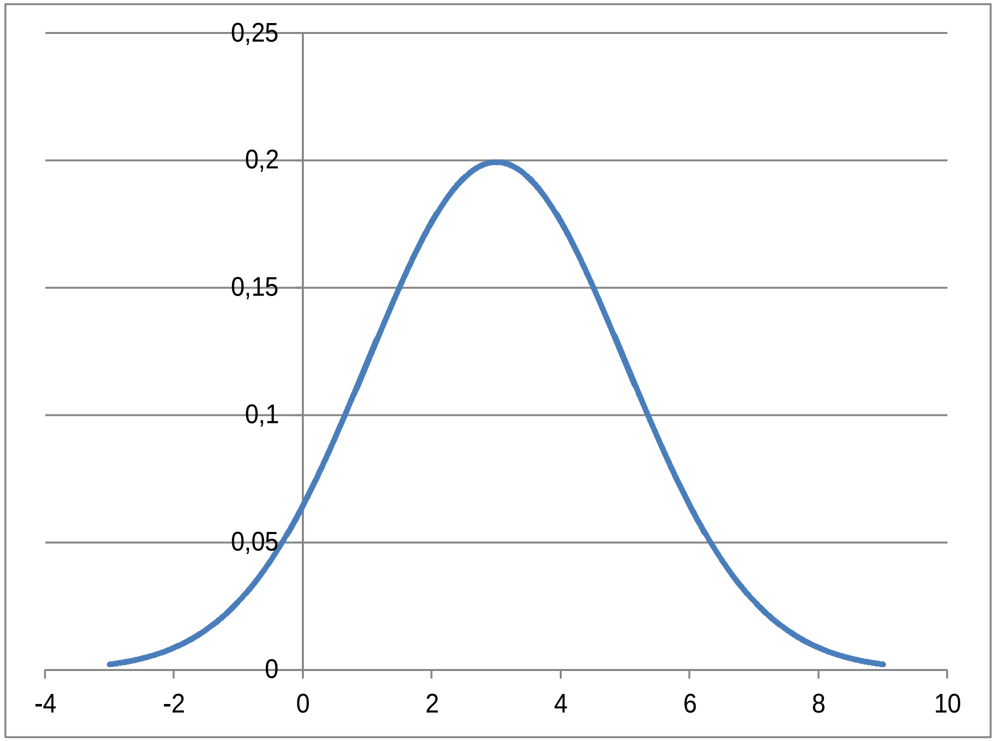
Satunnaismuuttuja X noudattaa normaalijakaumaa parametrein μ ja σ^2, jos sen tiheysfunktio on

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

Merk. X ~ N(μ, σ^2)

$$E(X) = \mu, Var(X) = \sigma^2$$

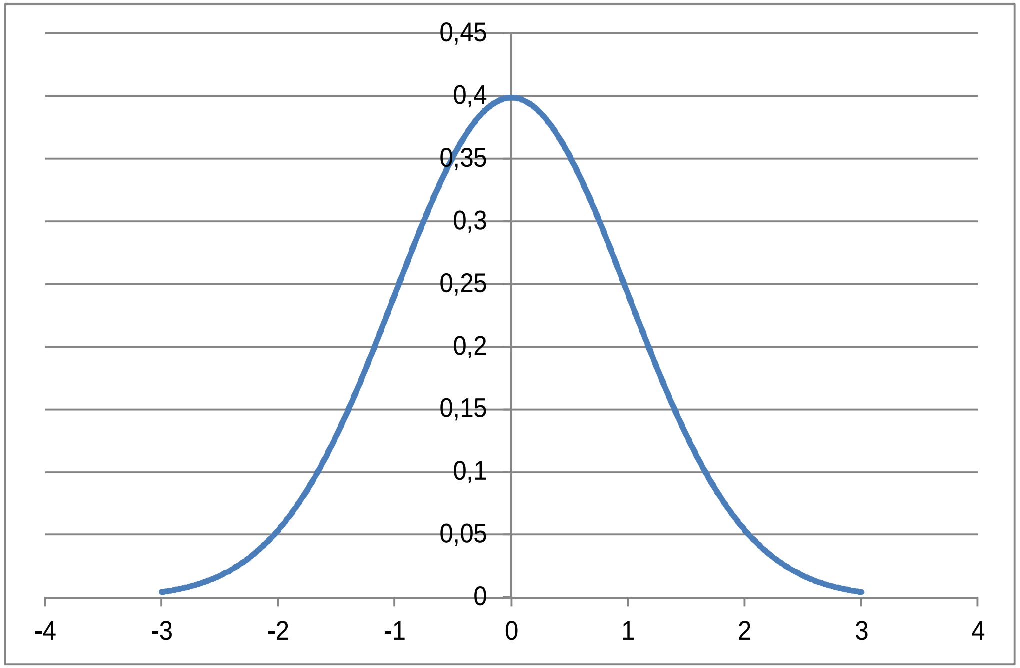
Esim. N(3, 4)



Esim. Parametrien vaikutus jakauman muotiin,
http://vassarstats.net/zsamp.html

Jos X ~ N(0, 1), niin kyse standardoidusta normaalijakaumasta. Tällöin

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



Merkitään $Z \sim N(0, 1)$

- tiheysfunktio $\phi(z)$
- kertymäfunktio $\Phi(z)$

$$P(Z \leq z) = \Phi(z)$$

$$\Phi(z) = 1 - \Phi(-z)$$

Kertymäfunktion arvoja on taulukoitu, ks. taulukko

http://www.sis.uta.fi/tilasto/mtttp5/syksy2018/N(0_1).pdf

Esim. 3.5.7 $Z \sim N(0, 1)$

$P(Z \leq 1) = \Phi(1) = 0{,}8413$

$P(Z \leq 1{,}1) = \Phi(1{,}1) = 0{,}8643$

$P(Z \leq 1{,}14) = \Phi(1{,}14) = 0{,}8729$

$P(Z \leq -1) = \Phi(-1) = 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587$

$P(Z \geq 2{,}4) = 1 - P(Z \leq 2{,}4) = 1 - \Phi(2{,}4) = 1 - 0{,}9918$

$= 0{,}0082$

$$P(Z \geq -1{,}14) = 1 - P(Z \leq -1{,}14) = 1 - \Phi(-1{,}14)$$

$$= 1 - (1 - \Phi(1{,}14)) = \Phi(1{,}14)$$

$$= 0{,}8729$$

$$P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = \ldots = 0{,}6826$$

$$P(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) = \ldots = 0{,}9544$$

$$P(-3 \leq Z \leq 3) = \Phi(3) - \Phi(-3) = \ldots = 0{,}9974$$

Esim. 3.5.8 $Z \sim N(0, 1)$

Jos $\Phi(z) = 0{,}75$, niin $z \approx 0{,}67$, koska $\Phi(0{,}67) = 0{,}7486$

Jos $\Phi(z) = 0{,}26$, niin $\Phi(-z) = 1 - 0{,}26 = 0{,}74$, $-z \approx 0{,}64$,

koska $\Phi(0{,}64) = 0{,}7389$, $z = -0{,}64$.

Ks. http://www.sis.uta.fi/tilasto/mtttp2/syksy2012/norm_graaf.pdf

Jos $X \sim N(\mu, \sigma^2)$, niin $Z = (X - \mu)/\sigma \sim N(0, 1)$.

- $P(X \leq a) = P((X - \mu)/\sigma \leq (a - \mu)/\sigma)$

  $= \Phi((a - \mu)/\sigma)$

- $P(X \geq a) = 1 - P(X \leq a)$

  $= 1 - P((X - \mu)/\sigma \leq (a - \mu)/\sigma)$

  $= 1 - \Phi((a - \mu)/\sigma)$

- $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$

  $= \Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma)$

Esim. 3.5.9 Sinulla on sijoitusvaihtoehdot A ja B. Oletat, että sijoitusten tuottoprosentit noudattavat normaalijakaumaa odotusarvoina 10,4 ja 11,0 sekä hajontoina 1,2 ja 4,0. Haluat tehdä sijoituksen, jolla on todennäköisempää saada vähintään 10 %:n tuotto. Kumman sijoitusvaihtoehdon valitset ja miksi?

Merkitään

X = tuotto sijoituksesta A, $X \sim N(10{,}4,\ 1{,}2^2)$

Y = tuotto sijoituksesta B, $Y \sim N(11{,}0,\ 4{,}0^2)$

$$
\begin{aligned}
P(X > 10) &= 1 - P(X \leq 10) \\
&= 1 - P((X - 10{,}4)/1{,}2 \leq (10 - 10{,}4)/1{,}2) \\
&= 1 - \Phi(-0{,}33) \\
&= 1 - (1 - \Phi(0{,}33)) \\
&= 0{,}6293
\end{aligned}
$$

$$
\begin{aligned}
P(Y > 10) &= 1 - P(Y \leq 10) \\
&= 1 - P((Y - 11)/4 \leq (10 - 11)/4) \\
&= 1 - \Phi(-0{,}25) = 1 - (1 - \Phi(0{,}25)) = 0{,}5987.
\end{aligned}
$$

Valitset sijoitusvaihtoehto A, koska siinä suurempi todennäköisyys saada vähintään 10 % tuotto.

Esim. Matti valmistaa tehtaassa erästä komponenttia. Ilman häiriötekijöitä Matin tekemien komponenttien pituus vaihtelee normaalijakauman, jonka odotusarvo on 2,500 cm ja keskihajonta 0,005 cm, mukaisesti. Eräänä päivänä Matti oli hieman väsynyt. Työpäivän lopussa hän valitsi satunnaisesti yhden tämän päivän aikana tekemänsä komponentin, jonka pituus oli 2,493 cm.

a) Laske todennäköisyys sille, että ko. päivän aikana Matin tekemien komponenttien joukosta satunnaisesti valitun komponentin pituus on pienempi kuin Matin valitseman, jos oletetaan, että päivän aikana tehtyjen komponenttien pituuden vaihtelussa ei ole tapahtunut tavanomaisesta poikkeavaa muutosta.

Olkoon X = komponentin pituus, joka tavanomaisessa tilanteessa noudattaa normaalijakaumaa odotusarvona 2,500 ja keskihajontana 0,005.

Tällöin

$$P(X \leq 2{,}493) = \Phi((2{,}493 - 2{,}500)/0{,}005)$$

$$= \Phi(-1{,}4) = 1 - \Phi(1{,}4)$$

$$= 1 - 0{,}9192 = 0{,}0808.$$

b) Oliko Matin väsymys vaikuttanut työn laatuun?
Valittaessa satunnaisesti yksi Matin tekemistä komponenteista, niin tavanomaisessa tilanteessa on siis n. 8,1 % todennäköisyys saada komponentti, joka on 2,493 cm lyhyempi. Ei siis ole mitenkään harvinaista, että saadaan komponentti, joka on pituudeltaan alle Matin valitseman komponentin. Näin päätellään, että Matin väsymys ei ole vaikuttanut työn laatuun. (Jos kuitenkin pitää laskettua todennäköisyyttä pienenä, niin tekee päinvastaisen päättely, mutta tällöin kiinnitetään riskitaso, joka on suurempi kuin 0,0808!)

Esim. Oletetaan, että sähkölamppujen käyttöikä X (tunteina) noudattaa normaalijakaumaa parametrein 800 ja 1600.

a) Laske todennäköisyys sille, että satunnaisesti valitun lampun käyttöikä on alle 850 mutta yli 700.

$P(700 \leq X \leq 850)$

$= \Phi((850 – 800)/40) - \Phi((700 – 800)/40)$

$= \Phi(5/4) - \Phi(-10/4) = \Phi(1,25) – (1 - \Phi(2,5))$

$= 0,8944 – (1 – 0,9938) = 0,8882$

b) 25 % valmistajan lampuista kestää yli a tuntia eli $P(X \geq a) = 0{,}25$. Määritä a.

Nyt $P(X < a) = 0{,}75$, joten $\Phi((a - 800)/40) = 0{,}75$. Taulukosta $\Phi(0{,}67) = 0{,}7486$, joten $(a - 800)/40 \approx 0{,}67$. Tästä $a = 826{,}8$.

MTTTP5, luento 13.11.2018

Kertausta

Jos $X \sim N(\mu, \sigma^2)$, niin $Z = (X - \mu)/\sigma \sim N(0, 1)$.

- $P(X \leq a) = P((X - \mu)/\sigma \leq (a - \mu)/\sigma)$
  $= \Phi((a - \mu)/\sigma)$

- $P(X \geq a) = 1 – P(X\leq a)$
  $= 1 - P((X - \mu)/\sigma \leq (a - \mu)/\sigma)$
  $= 1 - \Phi((a - \mu)/\sigma)$

- $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$
  $= \Phi((b - \mu)/\sigma) - \Phi((a - \mu)/\sigma)$

3.5.5 Normaalijakauma (jatkuu)

Normaalijakaumaan liittyviä keskeisiä tuloksia

- $Jos\ X \sim N(\mu, \sigma^2),\ niin\ aX + b \sim N(a\mu + b, a^2\sigma^2)$

- $Jos\ X_1, X_2, \ldots, X_n\ ovat\ riippumattomia\ ja\ X_i \sim N(\mu_i,\ \sigma_i^2), niin$

$$X_1 + X_2 + \cdots + X_n \sim N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$$

Esim. Lentomatkustajien tavaroiden painon oletetaan vaihtelevan siten, että ne painavat keskimäärin 20 kg keskihajonnan ollessa 5 kg. Oletetaan lisäksi painon vaihtelevan normaalijakauman mukaisesti. Eräs lentokonetyyppi kuljettaa 100 matkustajaa. Millä todennäköisyydellä matkatavaroiden yhteispaino ylittää 2150 kg?

Yhteispaino $Y = X_1 + X_2 + \ldots + X_{100}$, missä $X_i \sim N(20,25)$.

$E(Y) = E(X_1) + E(X_2) + \ldots + E(X_{100}) = 100 \cdot 20 = 2000$

$Var(Y) = Var(X_1) + Var(X_2) + \ldots + Var(X_{100}) = 100 \cdot 25$

$= 2500$

$Y \sim N(2000, 2500)$

$P(Y > 2150) \quad = 1 - P(Y \leq 2150)$

$= 1 – \Phi((2150-2000)/50)$

$= 1 – \Phi(3) = 1 – 0{,}9987 = 0{,}0013.$

- $Jos\ X_1, X_2, \ldots, X_n\ ovat\ riippumattomia\ \ ja$

$$E(X_i) = \mu_i$$

$$Var(X_i) = \sigma_i^2,$$

$niin$

$$X_1 + X_2 + \cdots + X_n \overset{likimain}{\sim} N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$$

Esim. Lentoyhtiötä pyydetään kuljettamaan 100 lammasta. Yhtiöllä on käytössä kone, joka voi ottaa kuljetettavakseen 5000 kg. Aiemmin on punnittu 1000 vastaavanlaista lammasta ja saatu keskiarvoksi 45 kg, hajonnaksi 3 kg ja painot ovat vaihdelleet välillä 37 kg – 56 kg. Voiko yhtiö ottaa pyydetyn 100 lampaan lastin kuljetettavakseen?

Yhteispaino $Y = X_1 + X_2 + \ldots + X_{100}$, missä $E(X_i) = 45$, $Var(X_i) = 9$

$E(Y) = E(X_1) + E(X_2) + \ldots + E(X_{100}) = 100 \cdot 45 = 4500$

$Var(Y) = Var(X_1) + Var(X_2) + \ldots + Var(X_{100}) = 100 \cdot 9$
$= 900$

$Y \sim N(4500, 900)$, likimain

$P(Y > 5000) = 1 - P(Y \leq 5000)$
$\approx 1 - \Phi((5000-4500)/30) = 1 - \Phi(16{,}67) \approx 0$

On siis lähes mahdotonta, että raja ylittyisi. Lampaat voi hyvin ottaa kuljetettavaksi. Liian varovainen arvio olisi $100 \cdot 56 = 5600$.

Edellisten tulosten perusteella saadaan otoskeskiarvoon liittyvät tulokset

- $Jos\ X_i \sim N(\mu,\ \sigma^2)\ ja\ X_i{:}t\ riippumattomia, niin$

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)\ \ ja$$

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)\ \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Esim. 3.5.14 GMAT-testiä käytetään useiden yliopistojen pääsykokeena. Kokeen tuloksen on todettu noudattavan normaalijakaumaa odotusarvona 525 ja keskihajontana 100. Sadan pyrkijän ryhmä osallistui ennen pääsykoetta valmennuskurssille. Pääsykokeessa heidän GMAT-testin keskiarvo oli 541,4. Menestyivätkö he pääsykokeessa muita paremmin?

$$\overline{X} \sim N\left(525, \frac{100^2}{100}\right)$$

$$P\left(\overline{X} \geq 541{,}4\right) = 1 - P\left(\overline{X} < 541{,}4\right)$$

$$= 1 - \Phi\left(\frac{541{,}4 - 525}{10}\right) = 1 - \Phi(1{,}64) = 0{,}0505$$

Eivät menestyneet paremmin kuin muut, koska ei ole harvinaista saada otoskeskiarvoa, joka suurempi kuin 541,4 silloin, kun menestyminen tavanomaista.

Esim. Auton sytytystulppien valmistaja väittää, että tulpat kestävät keskimäärin 60 000 km keskihajonnan ollessa 6 000 km sekä vaihtelu luonnehdittavissa normaalijakaumalla. Tutkit väitettä ja valitset satunnaisesti 4 tulppaa, joiden keskimääräiseksi kestoksi saat 55 500 km. Voitko uskoa valmistajan väitteen?

Jos valmistajan väite tosi, niin

$$\overline{X} \sim N\left(60000, \frac{6000^2}{4}\right).$$

$$P\left(\overline{X} \le 55500\right) = \Phi\left(\frac{55500 - 60000}{\frac{6000}{2}}\right) = \Phi(-1{,}5)$$

$$= 1 - \Phi(1{,}5) = 1 - 0{,}9332 = 0{,}0668$$

Uskotaan valmistajan väite, koska väitteen ollessa tosi ei ole harvinaista saada otosta, jonka keskiarvo alle 55500.

- $Jos\ X_1, X_2, \ldots X_n\ ovat\ riippumattomia, E(X_i) = \mu, Var(X_i) = \sigma^2,$
$niin$

$$X_1 + X_2 + \cdots + X_n \overset{likimain}{\sim} N(n\mu, n\sigma^2)$$

ja

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n) \overset{likimain}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Binomijakaumaa voidaan approksimoida normaalijakaumalla

- $Jos\ X \sim Bin(n,p), niin\ X \overset{likimain}{\sim} N\big(np, np(1-p)\big), kun\ n\ suuri.$

Esim. Tutkittiin uuden menetelmän käyttökelpoisuutta ihosairauden hoidossa. Vanhan menetelmän avulla 60 % potilasta parani. Uudella menetelmällä 72 potilasta sadasta parani. Onko uusi menetelmä vanhaa parempi?

Olkoon X = parantuneiden lukumäärä.
Jos uusi menetelmä toimii vanhan tavoin, niin
$X \sim \mathrm{Bin}(100,\ 0{,}6)$, $E(X) = 60$, $\mathrm{Var}(X) = 24$, joten
$X \sim N(60,\ 24)$ likimain.

Tällöin $P(X \geq 72) = 1 - P(X \leq 71)$

$\approx 1 - \Phi((71-60)/\sqrt{24})$

$= 1 - \Phi(2{,}26) = 0{,}0119.$

Binomijakaumasta laskettuna $P(X \geq 72) = 0{,}00843$.

Jos toimisi vanhan tavoin, niin olisi harvinaista saada parantuneita enemmän kuin 71. Päätellään uuden olevan parempi.

Esim. 3.5.12 Tentissä on 100 väittämää, jotka ovat tosia tai epätosia. Vastataan kaikkiin kysymyksiin arvaamalla.

Olkoon X = oikeiden vastausten lukumäärä.

X ~ Bin(100,1/2), joten

$$P(X = k) = \binom{100}{k}\left(\frac{1}{2}\right)^{100}, \qquad k = 0, 1, 2, \ldots 100$$

$$P(X \leq 60) = \sum_{k=0}^{60}\binom{100}{k}\left(\frac{1}{2}\right)^{100} = 0{,}9824$$

http://homepage.stat.uiowa.edu/~mbognar/applets/bin.html

E(X) = 100/2 = 50, Var(X) = 100/4 = 25, joten

$$X \overset{likimain}{\sim} N(50, 25)$$

$$P(X \leq 60) \approx \Phi\left(\frac{60-50}{5}\right) = \Phi(2) = 0{,}9772$$

MTTTP5, luento 15.11.2018

Luku 4
Satunnaisotos, otossuure ja otosjakauma

4.1. Satunnaisotos

$X_1, X_2, \ldots, X_n$ on satunnaisotos, jos X_i:t ovat riippumattomia ja noudattavat samaa jakaumaa.

Sanonta "$X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta" tarkoittaa, että jokainen $X_i \sim N(\mu, \sigma^2)$ ja X_i:t ovat riippumattomia.

4.2. Otossuureet ja otosjakaumat

- Otossuure

  satunnaisotoksen avulla määritelty funktio

- Otosjakauma

  otossuureen todennäköisyysjakauma

Otossuureita ja niiden jakaumia

1) Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta. Tällöin

$$\boxed{\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).}$$

2) Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos jakaumasta, jonka odotusarvo μ ja varianssi σ^2.

Tällöin

$$\boxed{\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), likimain.}$$

Esim. Erään tilastotoimiston (The National Center for Health Statistics) mukaan väestössä keski-ikäisten miesten verenpaineen keskiarvo on 128 ja keskihajonta 15. Haluttiin selvittää, poikkeaako keski-ikäisten yritysjohtajien verenpaineen keskiarvo koko väestön vastaavasta keskiarvosta. Mitattiin 72 yritysjohtajan verenpaineet ja saatiin keskiarvoksi 130,5. Onko eroja?

Olkoon X = verenpaine. Nyt

$$\overline{X} \sim N\left(128, \frac{15^2}{72}\right), likimain,$$

jos otos koko väestöstä.

$$P(\overline{X} \geq 130{,}5) = 1 - P(\overline{X} \leq 130{,}5) \approx 1 - \Phi\left(\frac{130{,}5 - 128}{\frac{15}{\sqrt{72}}}\right)$$

$$= 1 - \Phi(1{,}41) = 1 - 0{,}9207 = 0{,}0793$$

Ei voida ajatella, että yritysjohtajien verenpaineen keskiarvo olisi korkeampi kuin koko väestön, koska ei ole koko väestöstä tehdyssä 72 alkion otoksessa harvinaista saada otoskeskiarvoa, joka on yli yritysjohtajilta mitatun.

3) Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos populaatiosta, jossa π % viallisia. Määritellään $X_i = \begin{cases} 1, & jos\ viallinen \\ 0, & muulloin \end{cases}$

Viallisten kokonaislukumäärä otoksessa

$X = X_1 + X_2 + ... + X_n \sim \text{Bin}(n, \pi/100)$

Lisäksi

$$\boxed{X \overset{likimain}{\sim} N\left(\frac{n\pi}{100}, n(\frac{\pi}{100})\left(1 - \frac{\pi}{100}\right)\right), kun\ n\ suuri.}$$

Viallisten prosenttiosuus otoksessa p = 100X/n.

E(p) = π, Var(p) = $\pi(100-\pi)$/n, ks. esim. 5.1.1.

Koska X:n jakauma on likimain normaalijakauma, niin

$$\boxed{p \overset{likimain}{\sim} N\left(\pi, \frac{\pi(100-\pi)}{n}\right), kun\ n\ suuri.}$$

Esim. 4.2.3 Olet todistamassa oikeudessa, jossa väitetään erään pelipaikan ruletin toimivan väärin. Rulettissa on 37 numeroa, joiden kaikkien pitäisi olla yhtä todennäköisiä. Pelipaikka voittaa numerolla nolla. Olet saanut selville, että 3700 kertaa rulettia pyöritettäessä nolla tuli 140 kertaa. Millaisen todistuksen annat oikeudessa?

Olkoon X = nollien lukumäärä.

Jos ruletti toimii oikein, niin X ~ Bin(3700, 1/37).

E(X) = 3700·(1/37) = 100,

Var(X) = 3700·(1/37)·(36/37) = 3600/37.

Tällöin X ~ N(100, 3600/37), likimain.

$P(X \geq 140) = 1 - P(X \leq 139) \approx 1 - \Phi(\frac{139-100}{\sqrt{3600/37}})$

$= 1 - \Phi(3{,}95) \approx 0$. Tämä on siis lähes mahdotonta.

Todistan, että pelipaikan ruletti toimii väärin.

Esim. Yritys tekee tiettyä komponenttia, jota käytetään auton moottorissa. Tämä komponentti hajoaa joskus heti, kun se on otettu käyttöön. Yritys valvoo tuotantoaan siten, että virheellisten komponenttien osuus ei saisi olla suurempi kuin 4 %. Laaduntarkkailussa tehtiin 500 komponentin otos, joista 28 komponenttia osoittautui virheelliseksi. Onko tuotanto keskeytettävä?

Ratkaisu 1

Olkoon X = virheellisten komponenttien lukumäärä 500 alkion otoksessa.

Jos tuotannossa virheellisiä 4 %, niin

$X \sim \mathrm{Bin}(500,\ 0{,}04)$, jolloin $E(X) = 500 \cdot 0{,}04 = 20$,

$\mathrm{Var}(X) = 500 \cdot 0{,}04 \cdot 0{,}96 = 19{,}2$.

Lisäksi $X \sim N(20,\ 19{,}2)$, likimain.

$$
\begin{aligned}
P(X \geq 28) &= 1 - P(X \leq 27) \\
&\approx 1 - \Phi((27 - 20)/\sqrt{19{,}2}) \\
&= 1 - \Phi(1{,}60) = 0{,}0548.
\end{aligned}
$$

Tämä ei harvinaista, tuotantoa voidaan jatkaa.

Binomijakaumasta laskettuna $P(X \geq 28) = 0{,}0489$,

ks. http://homepage.stat.uiowa.edu/~mbognar/applets/bin.html

Ratkaisu 2

Olkoon p = virheellisten komponenttien prosenttiosuus 500 alkion otoksessa

Jos tuotannossa virheellisiä 4 %, niin

$$p \overset{likimain}{\sim} N(4, \underbrace{\frac{4(100-4)}{500}}_{0{,}768})$$

P(p ≥ 5,6) = 1 – P(p ≤ 5,6)

≈ 1 – Φ((5,6 – 4)/√0,768)

= 1 – Φ(1,83) = 1 – 0,9664 = 0,0336

Sama tulos ratkaisusta 1, jos lasketaan

$P(X \geq 28) \approx 1 - \Phi((28 - 20)/\sqrt{19{,}2}) = 1 - \Phi(1{,}83)$.

4) Olkoot $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \ \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2, \ \sigma_2^2)$:sta sekä otokset riippumattomia.

Tällöin

$$\boxed{\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)}$$

Esim. Tarkastellaan lapsen syntymäpainoa grammoina. Oletetaan, että tytöillä syntymäpaino $X \sim N(3450, 520^2)$ ja pojilla syntymäpaino $Y \sim N(3640, 440^2)$. Tarkastellaan tyttöpopulaatiosta 100 alkion ja poikapopulaatiosta 200 alkion satunnaisotoksia. Määritä otoskeskiarvojen jakaumat sekä otoskeskiarvojen erotuksen jakauma. Laske todennäköisyys sille, että tyttöjen otoskeskiarvo on suurempi kuin poikien.

$$\overline{X} \sim N\left(3450, \frac{520^2}{100}\right)$$

$$\overline{Y} \sim N\left(3640, \frac{440^2}{200}\right)$$

$$\overline{X} - \overline{Y} \sim N\left(3450 - 3640, \frac{100^2}{100} + \frac{440^2}{200}\right)$$

$$\overline{X} - \overline{Y} \sim N(-190, 3672), joten\ P\left(\overline{X} - \overline{Y} > 0\right)$$

$$= 1 - \Phi\left(\frac{0 - (-190)}{60{,}6}\right) = 1 - \Phi(3{,}14)$$

$$= 0{,}0008$$

Esim. 4.2.4 Koneiden A ja B pitäisi valmistaa keskimäärin samanmittaisia tankoja. Molempien koneiden tuottamien tankojen pituuksissa X ja Y (cm) on jonkin verran vaihtelua, jota voidaan luonnehtia normaalijakaumalla, jonka varianssi on 0,20 cm^2. Laadunvalvonnassa seurataan koneiden toimintaa ja valitaan satunnaisesti koneen A tuotannosta 20 ja koneen B tuotannosta 10 tankoa. Koneen A tuotannosta valittujen tankojen keskipituus on 40,0 cm ja koneelta B valittujen 39,5 cm. Tuottavatko koneet keskimäärin samanmittaisia tankoja?

$$\overline{X} - \overline{Y} \sim N\left(0, \underbrace{\frac{0{,}2}{20} + \frac{0{,}2}{10}}_{0{,}03}\right), \textit{jos tuottavat samanmittaisia}$$

$$P(\left|\overline{X} - \overline{Y}\right| > 0{,}5 = 1 - P\left(-0{,}5 \leq \overline{X} - \overline{Y} \leq 0{,}5\right)$$

$$= 1 - \left(\Phi\left(\frac{0{,}5 - 0}{\sqrt{0{,}03}}\right) - \Phi\left(\frac{-0{,}5 - 0}{\sqrt{0{,}03}}\right)\right)$$

$$= 1 - \left(\Phi(2{,}89) - \Phi(-2{,}89)\right) = 2 - 2\Phi(2{,}89)$$

$$= 0{,}0038$$

Eivät tuota keskimäärin samanmittaisia tankoja. Jos tuottaisivat, niin olisi harvinaista saada otokset, joiden keskiarvojen erotuksen itseisarvo olisi suurempi kuin 0,5 cm.

MTTTP5, luento 20.11.2018

Otossuureita ja niiden jakaumia (jatkuu)

4) Olkoot $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta sekä otokset riippumattomia.

Tällöin

$$\boxed{\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)}$$

Esim. 4.2.4 Koneiden A ja B pitäisi valmistaa keskimäärin samanmittaisia tankoja. Molempien koneiden tuottamien tankojen pituuksissa X ja Y (cm) on jonkin verran vaihtelua, jota voidaan luonnehtia normaalijakaumalla, jonka varianssi on 0,20 cm^2. Laadunvalvonnassa seurataan koneiden toimintaa ja valitaan satunnaisesti koneen A tuotannosta 20 ja koneen B tuotannosta 10 tankoa. Koneen A tuotannosta valittujen tankojen keskipituus on 40,0 cm ja koneelta B valittujen 39,5 cm. Tuottavatko koneet keskimäärin samanmittaisia tankoja?

Jos tuottaisivat keskimäärin samanmittaisia tankoja, niin

$$\overline{X} \sim N\left(\mu, \frac{0{,}20}{20}\right) \text{ ja } \overline{Y} \sim N\left(\mu, \frac{0{,}20}{10}\right), \text{ joten}$$

$$\overline{X} - \overline{Y} \sim N\left(0, \underbrace{\frac{0{,}2}{20} + \frac{0{,}2}{10}}_{0{,}03}\right).$$

Tällöin

$$P(|\overline{X}-\overline{Y}| > 0{,}5) = 1 - P\left(-0{,}5 \le \overline{X}-\overline{Y} \le 0{,}5\right)$$

$$= 1 - \left(\Phi\left(\frac{0{,}5-0}{\sqrt{0{,}03}}\right) - \Phi\left(\frac{-0{,}5-0}{\sqrt{0{,}03}}\right)\right)$$

$$= 1 - \left(\Phi(2{,}89) - \Phi(-2{,}89)\right) = 2 - 2\Phi(2{,}89)$$

$$= 0{,}0038$$

Eivät tuota keskimäärin samanmittaisia tankoja. Jos tuottaisivat, niin olisi harvinaista saada otokset, joiden keskiarvojen erotuksen itseisarvo olisi suurempi kuin 0,5 cm.

Luku 5
Parametrien estimointi

5.1 Piste-estimointi

- Estimointi

  populaation tuntemattoman parametrin arviointia otossuureen avulla

- Otossuure

  satunnaisotoksen avulla määritelty funktio

- Otosjakauma

  otossuureen todennäköisyysjakauma

- Estimaattori

  otossuure, jolla estimoidaan populaation tuntematonta parametria

- Estimaattorin keskivirhe

  estimaattorin keskihajonta

- Estimaatti

  estimaattorin arvo (tehdyn otoksen perusteella laskettu)

- Miten estimaattori valitaan?
- Mitä estimaattorista on tiedettävä?
- Mikä on hyvä estimaattori?
- Miksi $\bar{X}$ on hyvä μ:n estimaattori?
- Miksi p on hyvä π:n estimaattori?

Esim. Jos populaatiossa viallisia π %, niin viallisten prosenttiosuus otoksessa

$$p \sim N\left(\pi, \frac{\pi(100-\pi)}{n}\right), likimain.$$

Siis E(p) = π eli estimaattorin odotusarvo on estimoitava parametri. Tätä ominaisuutta kutsutaan harhattomuudeksi, p on π:n harhaton estimaattori.

Estimaattorin p:n hajonta eli keskivirhe on $\sqrt{\frac{\pi(100-\pi)}{n}}$.

Esim. 5.1.2 $E(\bar{X}) = \mu$, joten $\bar{X}$ on μ:n harhaton estimaattori, $\bar{X}$:n keskivirhe on $\frac{\sigma}{\sqrt{n}}$.

Estimaattorilta vaadittavia ominaisuuksia

- harhattomuus
- mahdollisimman pieni varianssi (tehokkain estimaattori)
- tarkentuvuus eli otoskoon kasvaessa rajatta estimaattorin varianssi lähenee nollaa

Esim. 5.1.2 Voidaan osoittaa, että normaalijakauman tapauksessa $\bar{X}$ on tehokkain μ:n estimaattori eli $\bar{X}$:lla on pienin varianssi harhattomien estimaattoreiden joukossa.

Esim. 5.1.3 $E(S^2) = \sigma^2$

5.2 Luottamusvälejä

Parametria väliestimoidaan nk. luottamusvälin avulla.

Olkoon $Z \sim N(0, 1)$. Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$. Vastaavalla tavalla $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$.

Esim. $z_{0,05} = 1,64$, koska $\Phi(1,64) = 0,9495$

$z_{0,025} = 1,96$, koska $\Phi(1,96) = 0,9750$

$z_{0,005} = 2,58$, koska $\Phi(2,58) = 0,9951$

Ks.

http://www.sis.uta.fi/tilasto/mtttp3/kevat2015/zalpha.pdf

5.2.1 Populaation odotusarvon luottamusväli

Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 tunnettu.

Tällöin

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ ja}$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

joten

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1 - \alpha.$$

Tästä saadaan

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Satunnaisväli $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ sisältää μ:n todennäköisyydellä 1-α. Tätä väliä kutsutaan populaation odotusarvon μ 100(1 - α) %:n luottamusväliksi (kaava 4.1). Varmuus eli luottamustaso on 1 - α.

Esim. 5.2.2 Sokerin pussituskone tuottaa pusseja, joiden paino vaihtelee normaalijakauman mukaisesti keskihajontana 2,5 g. Koneeseen tehdään säätöjä ja punnitaan 20 pussia. Näiden keskipainoksi saadaan 1002 g. Voidaanko päätellä, että kone tuottaa keskimäärin kilon pusseja?

95 %:n luottamusväli μ:lle, kun σ tunnettu,

$$\bar{X} \pm z_{0,05/2}\ \frac{\sigma}{\sqrt{n}}.$$

Saadaan

$$1002 \pm 1{,}96 \cdot 2{,}5/\sqrt{20}$$

$$1002 \pm 1{,}1 \text{ eli } (1000{,}9,\ 1003{,}1)$$

Luottamusväli ei sisällä kiloa. Päätellään, että kone ei tuota keskimäärin kilon pusseja.

Sama päättely 99 %:n luottamusvälin

$$1002 \pm 2{,}58 \cdot 2{,}5/\sqrt{20}$$

$$(1000{,}6,\ 1003{,}4)$$

perusteella.

Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 <u>tuntematon</u>. Tällöin

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Olkoon t_{df} Studentin t-jakaumaa noudattava satunnaismuuttuja.

Määritellään $t_{\alpha,df}$ siten, että $P(t_{df} \geq t_{\alpha,df}) = \alpha$ ja $t_{\alpha/2,df}$ siten, että $P(t_{df} \geq t_{\alpha/2,df}) = \alpha/2$.

Sivut 19-24 seuraavalle luennolle

Esim. 5.2.4

$t_{0,05,\ 10}$ = 1,812, $t_{0,05,\ 30}$ = 1,697

$t_{0,01,\ 10}$ = 2,764, $t_{0,01,\ 30}$ = 2,457

Satunnaismuuttujan $t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ perusteella voidaan

johtaa (kuten kaava 4.1) populaation odotusarvon μ

100(1 - α) %:n luottamusväli, kun σ^2 tuntematon.

Saadaan

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}. \qquad \text{kaava (4.2)}$$

Esim. 5.2.5 Poikien keskimääräinen syntymäpituus (SAIDIT-aineisto) $\bar{x} = 50{,}95$, $s = 1{,}97$, $n = 65$, $t_{0,05/2,64} \approx 2{,}000$. Nyt 95 %:n luottamusväli on

$$50{,}95 \pm t_{\frac{0{,}05}{2};65-1} \frac{1{,}97}{\sqrt{65}}$$

$$50{,}95 \pm 2 \cdot \frac{1{,}97}{\sqrt{65}}.$$

Poikien keskipituuden arvellaan olevan välillä

(50,46, 51,44).

Esim. Cooperin testin tulos (CTESTI-aineisto), 15-vuotiaat, $\bar{x}$ = 2534, s = 255, n = 28, $t_{0,05/2,27}$ = 2,052, 95 %:n luottamusväli odotusarvolle

$$2534 \pm 2{,}052 \cdot \frac{255}{\sqrt{28}}$$ eli väli (2435, 2633).

SPSS-tulos

T-Test

luottamusväli μ:lle $\bar{x} \pm t_{\alpha/2;n-1} \cdot s/\sqrt{n}$

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
cooper	$n =$ 28	2533,93 $= \bar{x}$	255,460 $= s$	48,277 $= s/\sqrt{n}$

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference: Lower	95% Confidence Interval of the Difference: Upper
cooper	52,487	27	,000	2533,929	2434,87	2632,99

$t_{0{,}025;27} = 2{,}052$

$2533{,}93 - 2{,}052 \cdot 48{,}277$

$2533{,}93 + 2{,}052 \cdot 48{,}277$

Esim. 5.2.6 Keskimääräiset neliövuokrat Tampereen Hervannassa (2011), aineisto Tre_vuokra-asunnot_2011.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

$\bar{x}$ = 12,32, s = 2,25, n = 26,

95 %:n luottamusväli odotusarvolle (11,41, 13,23)

SPSS-tulos

Neliövuokra Hervannassa, aineisto Tre_vuokra-asunnot_2011.sav sivulla http://www.uta.fi/sis/mtt/mtttp1/aineistoja.html.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Neliövuokra	26	12,3190	2,25281	,44181

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Neliövuokra	27,883	25	,000	12,31904	11,4091	13,2290

$t_{0.025,25} = 2,060$, luottamusväli $12,3190 \pm 2,060 \cdot 2,25281/\sqrt{26}$

MTTTP5, luento 22.11.2018

Luottamusväli, määritelmä

Olkoot A ja B satunnaisotoksen perusteella määriteltyjä satunnaismuuttujia. Väli (A, B) on parametrin θ 100(1 - α) %:n luottamusväli, jos

$P(A \leq \theta \leq B) = 1 - \alpha.$

5.2.1 Populaation odotusarvon luottamusväli (jatkoa)

Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 tunnettu. Tällöin populaation odotusarvon μ $100(1 - \alpha)$ %:n luottamusväli on

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}. \qquad \text{Kaava 4.1.}$$

Olkoon $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 tuntematon. Tällöin

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Olkoon t_{df} Studentin t-jakaumaa noudattava satunnaismuuttuja.

Määritellään $t_{\alpha,df}$ siten, että $P(t_{df} \geq t_{\alpha,df}) = \alpha$ ja $t_{\alpha/2,df}$ siten, että $P(t_{df} \geq t_{\alpha/2,df}) = \alpha/2$.

Esim. 5.2.4

$t_{0,05,\ 10} = 1{,}812, \quad t_{0,05,\ 30} = 1{,}697$

$t_{0,01,\ 10} = 2{,}764, \quad t_{0,01,\ 30} = 2{,}457$

Satunnaismuuttujan $t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ perusteella voidaan johtaa (kuten kaava 4.1) populaation odotusarvon μ 100(1 - α) %:n luottamusväli, kun σ^2 tuntematon. Saadaan

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}. \qquad \text{kaava (4.2)}$$

Esim. 5.2.5 Poikien keskimääräinen syntymäpituus (SAIDIT-aineisto) $\bar{x}$ = 50,95, s = 1,97, n = 65, $t_{0,05/2,64}$ ≈ 2,000. Nyt 95 %:n luottamusväli on

$$50{,}95 \pm t_{\frac{0{,}05}{2};65-1} \frac{1{,}97}{\sqrt{65}}$$

$$50{,}95 \pm 2 \cdot \frac{1{,}97}{\sqrt{65}}.$$

Poikien keskipituuden arvellaan olevan välillä

(50,46, 51,44).

Esim. 5.2.6 Keskimääräiset neliövuokrat Tampereen Hervannassa (2011), aineisto Tre_vuokra-asunnot_2011.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

$\bar{x}$ = 12,32, s = 2,25, n = 26,

95 %:n luottamusväli odotusarvolle (11,41, 13,23)

SPSS-tulos

Neliövuokra Hervannassa, aineisto Tre_vuokra-asunnot_2011.sav sivulla http://www.uta.fi/sis/mtt/mtttp1/aineistoja.html.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Neliövuokra	26	12,3190	2,25281	,44181

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Neliövuokra	27,883	25	,000	12,31904	11,4091	13,2290

$t_{0.025,25} = 2,060$, luottamusväli $12,3190 \pm 2,060 \cdot 2,25281/\sqrt{26}$

SPSS-ohjeet:

Neliövuokra: Transform -> Compute -> Neliövuokra = Vuokra/Neliöt

Vain Hervanta analyyseihin: Data -> Select Cases -> If condition is satisfied -> Alue=8 (tai Kaupunginosa='Hervanta')

Luottamusväli: Analyze -> Compare Means -> One-Sample T Test -> Test Variable Neliövuokra

Esim. Eräs yritys tarjoaa valmennuskurssia yliopistoon pyrkijöille. Yritys haluaa tutkia kurssinsa tehokkuutta. Tutkitaan pareja, joilla on samanlaiset lähtötiedot. Toinen osallistuu valmennuskurssille, toinen ei. Saadaan aineisto, jossa pyrkijöiden valintakoepisteet.

	Osallistui	Ei osallistunut	Erotus D
Pari 1	82	75	7
Pari 2	73	71	2
Pari 3	59	52	7
Pari 4	48	46	2
Pari 5	69	70	-1
Pari 6	93	83	10

Tarkastellaan erotusta D, muodostetaan sen odotusarvolle luottamusväli, joka on

$$\bar{D} \pm t_{\alpha/2,n-1} \frac{s_D}{\sqrt{n}}$$

$\bar{d}$=(7+2+7+2-1+10)/6=4,5

$$s_D^2 = \frac{(7^2 + 2^2 + 7^2 + 2^2 + (-1)^2 + 10^2) - 6 \cdot 4{,}5^2}{6-1} = 17{,}1$$

95 %:n luottamusväli on $4{,}5 \pm 2{,}571 \cdot \frac{\sqrt{17{,}1}}{\sqrt{6}}$ eli väli (0,16, 8,84). Erotuksen odotusarvon ei siis ajatella olevan nolla, joten valmennuskurssilla on vaikutusta.

http://vassarstats.net/-> t-Tests & Procedures

VassarStats Printable Report
0.95 and 0.99 Confidence Intervals for the Estimated Mean of a Population
Thu Jan 15 2015 16:07:22 GMT+0200 (FLE Standard Time)

Values entered:
X = {7,2,7,2,-1,10}

Summary Values:

$N = 6$
$\Sigma X = 27$
$\Sigma X^2 = 207$
mean = 4.5
variance = 17.1
std. dev. = 4.1352
std. error = 1.6882
df = 5
$t_{crit(.05)} = 2.57$
$t_{crit(.01)} = 4.03$

Confidence Intervals for Estimated Mean of Population
For .95 CI: 4.5±4.3387
For .99 CI: 4.5±6.8034

SPSS-tulosteet

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Erotus	$n =$ 6	$\bar{x} =$ 4,5000	$s =$ 4,13521	1,68819 $= s/\sqrt{n}$

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference: Lower	95% Confidence Interval of the Difference: Upper
Erotus	2,666	5	,045	4,50000	,1604	8,8396

$\bar{x} \pm t_{0,05/2;5} \cdot s/\sqrt{n}$

$t_{0,025;5} = 2,571$

5.2.2 Prosentuaalisen osuuden luottamusväli

Jos populaatiossa viallisia π %, niin viallisten prosenttiosuus otoksessa

$$p \sim N\left(\pi, \frac{\pi(100-\pi)}{n}\right), likimain.$$

Tällöin

$$Z = \frac{p - \pi}{\sqrt{\pi(100 - \pi)/n}} \sim N(0,1), likimain.$$

Tämän perustella saadaan π:n 100(1 - α) %:n luottamusväli

$$p \pm z_{\alpha/2}\sqrt{\frac{p(100-p)}{n}}.$$

Kaava 4.3.

Esim. Ruletissa 37 numeroa, joista pyöritettäessä jokaisella pitäisi olla sama todennäköisyys tulla tulokseksi. Pelipaikka voittaa numerolla nolla. Rulettia pyöritetään 3700 kertaa. Saadaan nollia 140 eli 3,78 %. Toimiiko ruletti oikein?

Lasketaan 99 %:n luottamusväli nollien % -osuudelle. Nyt α = 0,01, $z_{0,01/2} = z_{0,005} = 2,57$, koska $\Phi(2,57) = 0,9949$, p = (140/3700)·100 = 3,78, luottamusväli

$$3{,}78 \pm 2{,}57\sqrt{\frac{3{,}78(100-3{,}78)}{3700}}.$$

Nollien prosenttiosuuden arvellaan olevan välillä 2,97–4,59.

Jos ruletti toimisi oikein, niin nollia pitäisi tulla $(1/37)\cdot 100$ % = 2,70 %. Tämä ei kuulu luottamusvälille, joten päätellään ruletin toimivan väärin.

Jos laskettaisiin 95 %:n luottamusväli, saataisiin väli (3,17, 4,39).

Esim. Hyväkuntoisten osuus myydyistä kolmioista, aineisto Tre_myydyt_kolmiot_2010.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Kunto

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hyvä	120	69,0	**69,0**	69,0
	Tyydyttävä	51	29,3	29,3	98,3
	Huono	3	1,7	1,7	100,0
	Total	**174**	100,0	100,0	

95 %:n luottamusväli on $69 \pm 1{,}96\sqrt{\dfrac{69(100-69)}{174}}$

Hyväkuntoisten % -osuuden arvellaan olevan välillä 62,1 – 75,9.

Esim. Hyväkuntoisten osuus myydyistä kolmioista sijainnin mukaan tarkasteltuna

Kunto * Sijainti Crosstabulation

			Sijainti				
			Keskusta	Kaleva, Amuri, Pyynikki	Hatanpää, Nekala, Epilä, Kissanmaa	Lentävänniemi	Hervanta
Kunto	Hyvä	Count	23	21	25	27	24
		% within Sijainti	**76,7%**	**56,8%**	75,8%	71,1%	66,7%
	Tyydyttävä	Count	7	14	7	11	12
		% within Sijainti	23,3%	37,8%	21,2%	28,9%	33,3%
	Huono	Count	0	2	1	0	0
		% within Sijainti	0,0%	5,4%	3,0%	0,0%	0,0%
Total		Count	**30**	**37**	33	38	36
		% within Sijainti	100,0%	100,0%	100,0%	100,0%	100,0%

95 %:n luottamusväli

- Keskusta: $76{,}7 \pm 1{,}96\sqrt{\frac{76{,}7(100-76{,}7)}{30}}$ eli 61,5 – 91,9
- Kaleva, Amuri, Pyynikki: $56{,}8 \pm 1{,}96\sqrt{\frac{56{,}8(100-56{,}8)}{37}}$

  eli 40,8 – 72,8.

SPSS-ohjeet:

Frekvenssijakauma:

Analyze -> Descriptive Statistics -> Frequencies -> Kunto

Ristiintaulukko:

Analyze -> Descriptive Statistics -> Crosstabs -> Row(s) -> Kunto, Column(s) -> Sijainti -> Cells... prosenttijakaumat

5.2.3 Kahden populaation odotusarvojen erotuksen luottamusväli

Esim. Kolmioiden keskineliöhinnat ja keskineliöhintojen luottamusvälit sijainnin mukaan tarkasteltuna

One-Sample Statistics

Sijainti		N	Mean	Std. Deviation	Std. Error Mean
Keskusta	Neliöhinta	**30**	**2598,7142**	517,63351	**94,50652**
Kaleva, Amuri, Pyynikki	Neliöhinta	37	2353,2405	368,52174	60,58460
Hatanpää, Nekala, Epilä, Kissanmaa	Neliöhinta	33	1963,3219	538,43268	93,72910
Lentävänniemi	Neliöhinta	38	1630,4808	429,16341	69,61950
Hervanta	Neliöhinta	36	1371,3906	136,52228	22,75371

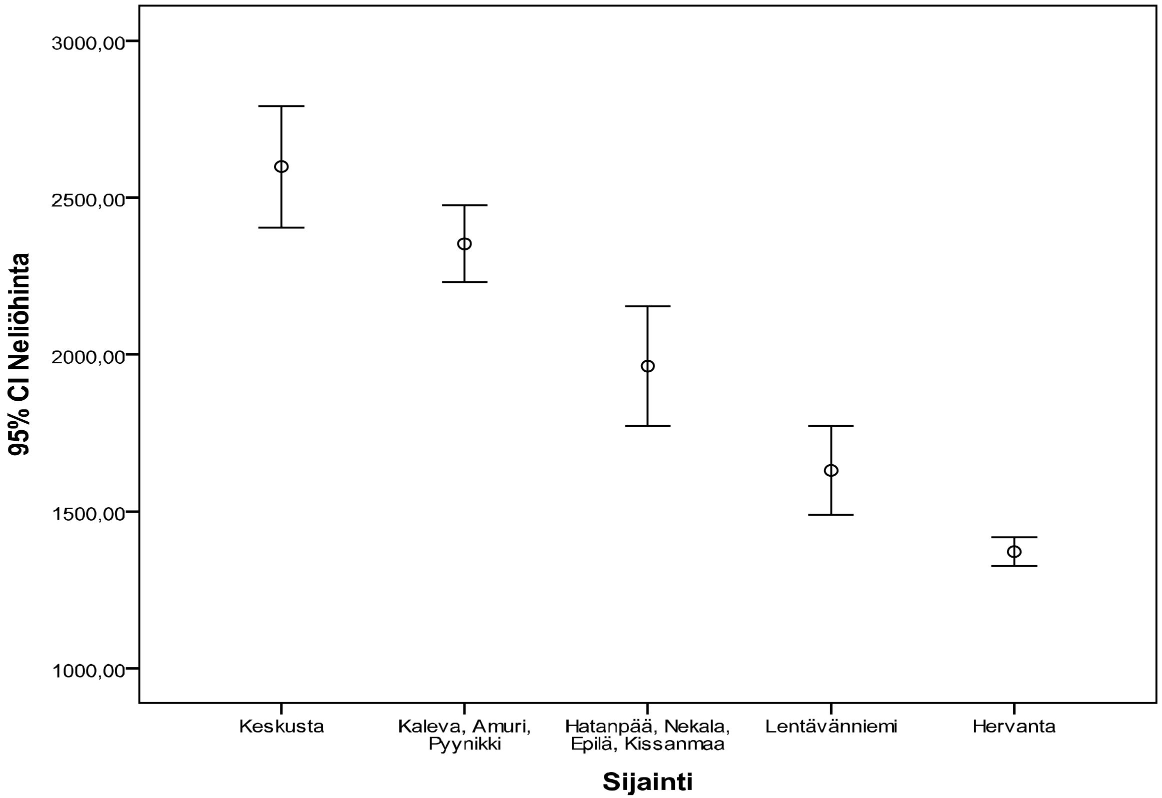
95 %:n luottamusvälit odotusarvolle

Sijainti		Lower	Upper
Keskusta	Neliöhinta	**2405,4267**	**2792,0017**
Kaleva, Amuri, Pyynikki	Neliöhinta	2230,3692	2476,1118
Hatanpää, Nekala, Epilä, Kissanmaa	Neliöhinta	1772,4019	2154,2418
Lentävänniemi	Neliöhinta	1489,4183	1771,5433
Hervanta	Neliöhinta	1325,1981	1417,5831

Luottamusväli keskustassa

$$2598{,}71 \pm t_{0{,}05/2;30-1} \frac{517{,}63}{\sqrt{30}} = 2598{,}71 \pm 2{,}045 \cdot 94{,}51.$$

Luottamusvälit graafisesti:



Esim. Kolmioiden keskineliöhinnat, kahden sijainnin vertailu, luottamusväli odotusarvojen erotukselle

Group Statistics

	Sijainti	N	Mean	Std. Deviation	Std. Error Mean
Neliöhinta	Keskusta	30	2598,7142	517,63351	94,50652
	Kaleva, Amuri, Pyynikki	37	2353,2405	368,52174	60,58460

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means			
						95% Confidence Interval of the Difference	
		F	Sig.	Mean Difference	Std. Error Difference	Lower	Upper
Neliöhinta	Equal variances assumed	4,462	,039	245,47371	108,42446	**28,93510**	**462,01231**
	Equal variances not assumed			245,47371	112,25852	20,08589	470,86152

SPSS-ohjeet:

Neliöhinta: Transform -> Compute -> Neliöhinta = Hinta/Neliöt

Luottamusväli: Analyze -> Compare Means -> Independent-Samples T Test -> Test Variable(s): Neliöhinta, Grouping Variable: Sijainti: Group 1: 1 (Keskusta), Group 2: 2 (Kaleva, Amuri, Pyynikki)

$X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu_1, \sigma_1^2)$:sta,

$Y_1, Y_2, \ldots, Y_m$ on satunnaisotos $N(\mu_2, \sigma_2^2)$:sta.

Oletetaan, että varianssit tunnettuja ja satunnaisotokset riippumattomia. Tällöin otoskeskiarvojen erotus

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right), \text{ joten}$$

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

Tästä voidaan johtaa 100(1 - α) %:n luottamusväli erotukselle (μ_1- μ_2). Luottamusväliksi saadaan

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}. \qquad \text{Kaava 4.4.}$$

Jos varianssit tuntemattomia, mutta voidaan olettaa, että $\sigma_1^2 = \sigma_2^2$, niin 100(1 - α) %:n luottamusväli erotukselle (μ_1- μ_2) on

$$\overline{X} - \overline{Y} \pm t_{\frac{\alpha}{2}, n+m-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}$$ Kaava 4.5.

Tuntematonta varianssia estimoidaan otosvarianssien avulla

$$\hat{\sigma}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} = s^2.$$

Esim. Kolmioiden keskineliöhinnat, kahden sijainnin vertailu

Group Statistics

	Sijainti	N	Mean	Std. Deviation	Std. Error Mean
Neliöhinta	Keskusta	$n =$ 30	2598,7142	$s_x =$ 517,63351	94,50652
	Kaleva, Amuri, Pyynikki	$m =$ 37	2353,2405	$s_y =$ 368,52174	60,58460

$$s^2 = \frac{(30-1)\,517{,}63351^2 + (37-1)\,368{,}52174^2}{30+37-2} \approx 441{,}32^2$$

$$s\sqrt{\frac{1}{n}+\frac{1}{m}}$$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means			
						95% Confidence Interval of the Difference	
		F	Sig.	Mean Difference	Std. Error Difference	Lower	Upper
Neliöhinta	Equal variances assumed	4,462	,039	245,47371	108,42446	**28,93510**	**462,01231**
	Equal variances not assumed			245,47371	112,25852	20,08589	470,86152

$t_{0,05/2,\,65} \approx 2$

luottamusväli:

$245{,}47371 \pm 2 \cdot 108{,}42446$

MTTTP5, luento 27.11.2018

5.2.3 Kahden populaation odotusarvojen erotuksen luottamusväli (kertausta)

Kun $\sigma_1^2 = \sigma_2^2$ tuntemattomia, niin 100(1 - α) %:n

luottamusväli odotusarvojen erotukselle (μ_1- μ_2), on

$$\overline{X} - \overline{Y} \pm t_{\alpha/2;n+m-2}\, s\sqrt{\frac{1}{n} + \frac{1}{m}}$$ Kaava 4.5.

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

Esim. 5.2.11 Miesten ja naisten musikaalisuus.

Group Statistics

	Sukupuoli	N	Mean	Std. Deviation	Std. Error Mean
Musikaalisuus	Mies	$n =$ 20	$\bar{x}$ 43,90	s_x 7,860	1,758
	Nainen	$m =$ 25	$\bar{y}$ 42,20	s_y 6,982	1,396

$$s^2 = \frac{(20-1)7,86^2+(25-1)6,982^2}{20+25-2} = 54,51$$

$$s = 7,38$$

$$s\sqrt{\frac{1}{n}+\frac{1}{m}} = 7,38\sqrt{\frac{1}{20}+\frac{1}{25}}$$

Independent Samples Test

evene's Test for Equality of Variances		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
,185	,669	,768	43	,447	$\bar{x}-\bar{y}$ 1,700	2,215	-2,767	6,167
		,757	38,435	,453	1,700	2,245	-2,843	6,243

$t_{0,05/2;43} \approx 2,021$

luottamusväli $1,7 \pm 2,021 \cdot 2,215$

Luku 6

Hypoteesien testaus

Tutkimusongelmia ja tilastollisia hypoteeseja:

- Perunalastupussien keskimääräinen paino?

$H_0 : \mu = \mu_0$	Nollahypoteesi
$H_1 : \mu < \mu_0$	Vaihtoehtoinen hypoteesi (yksisuuntainen)

- Virheellisten komponenttien osuus tuotannossa?

$H_0 : \pi = \pi_0$	Nollahypoteesi
$H_1 : \pi > \pi_0$	Vaihtoehtoinen hypoteesi (yksisuuntainen)

- Asuntojen keskimääräisen neliöhinnat keskustassa ja lähiössä?

$H_0 : \mu_1 = \mu_2$ Nollahypoteesi

$H_1 : \mu_1 > \mu_2$ Vaihtoehtoinen hypoteesi (yksisuuntainen)

Esim. Perunalastupussien valmistaja ilmoittaa pussien keskipainoksi 340 g. Oletetaan painon vaihtelun olevan normaalijakautunut hajontana 10 g. Tutkitaan väitettä ja tehdään 9 alkion satunnaisotos. Otoskeskiarvoksi saadaan 336 g.

$H_0 : \mu = 340$ g

$H_1 : \mu < 340$ g

Tilastollinen hypoteesi

Jos H_0 on tosi, niin $\overline{X} \sim N\left(340, \frac{10^2}{9}\right)$.

Tällöin

$$Z = \frac{\bar{X} - 340}{10/3} \sim N(0,1)$$

$\boxed{\text{Otossuure}}$, jonka jakauma tunnetaan, kun H_0 tosi. Otossuureesta käytetään nimitystä $\boxed{\text{testisuure}}$.

$$z_{havaittu} = \frac{336-340}{10/3} = -1{,}2$$

Testisuureen arvo

otoksesta laskettuna,

päättely tämän

perusteella

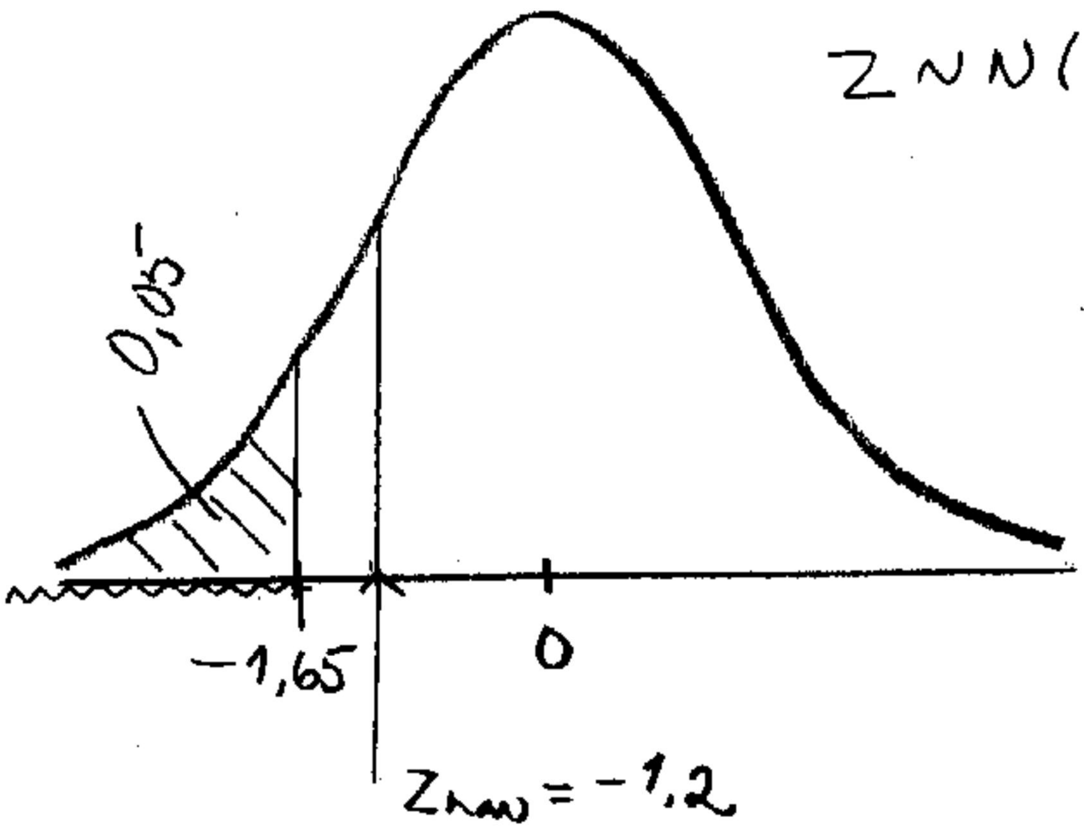
$\boxed{\text{Hyväksytäänkö vai hylätäänkö nollahypoteesi } H_0\text{?}}$

Hyväksytään H_0, jos otoksesta laskettu testisuureen arvo kuuluu tavanomaisiin arvoihin. Jos otoksesta laskettu testisuureen arvo kuuluu harvinaisiin arvoihin, niin H_0 hylätään ja H_1 hyväksytään.

Mikä on harvinaista?

Testisuure noudattaa H_0:n ollessa tosi standardoitua normaalijakaumaa, joten harvinaisina arvoina voidaan pitää esimerkiksi

$-1,65 = -z_{0,05}$ pienempiä arvoja. Jos tehdään näin, niin suoritetaan testaus 5 %:n merkitsevyys- eli riskitasolla, ja hyväksytään H_0.



$Z \sim N(0,1)$, kun H_0 tosi

Usein $\boxed{\text{riskitaso } \alpha}$ = 0,05, 0,025, 0,01, 0,001

Voidaan määrittää myös pienin riskitaso, jolla H_0 voidaan hylätä. Tätä kutsutaan $\boxed{\text{p–arvoksi}}$.

Nyt jos H_0 tosi, niin $P(Z \leq -1,2) = 1 - \Phi(1,2) = 0,1151$ = p-arvo.

Tätä suuremmilla riskeillä H_0 voidaan hylätä. Ei oteta näin suurta riskiä!

Testaukseen liittyvät virhetodennäköisyydet

	Todellinen tilanne	
	H_0 tosi	H_0 epätosi
H_0 hyväksytään	$1 - \alpha$	β 2. lajin virhe
H_0 hylätään	α 1. lajin virhe	$1 - \beta$ testin voimakkuus

6.1 Erilaisia testejä

6.1.1 Yhden populaation odotusarvoa koskeva päättely

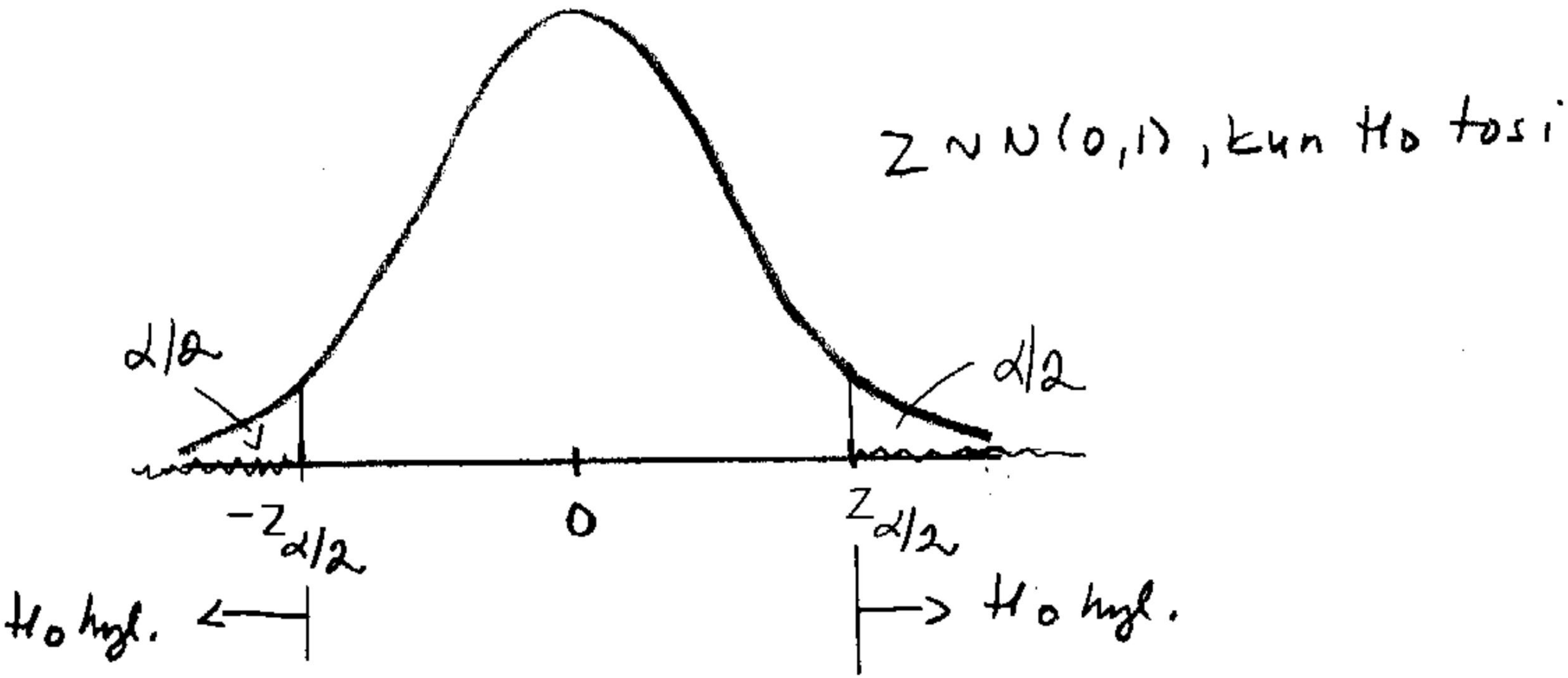
$$\boxed{H_0 : \mu = \mu_0}$$

Olk. $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 tunnettu.

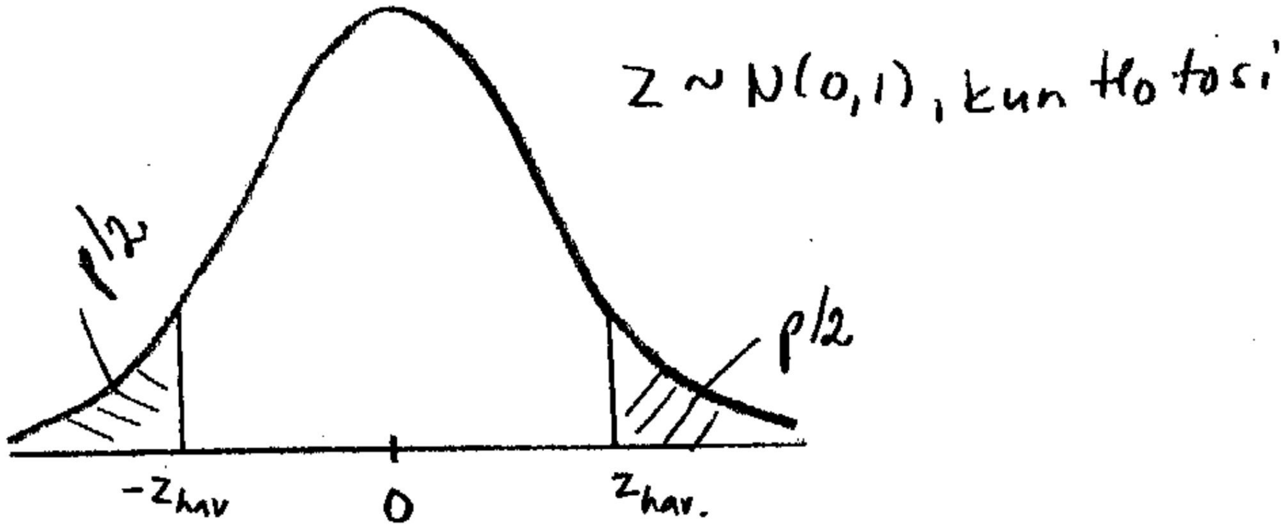
Jos H_0 on tosi, niin

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1).$$

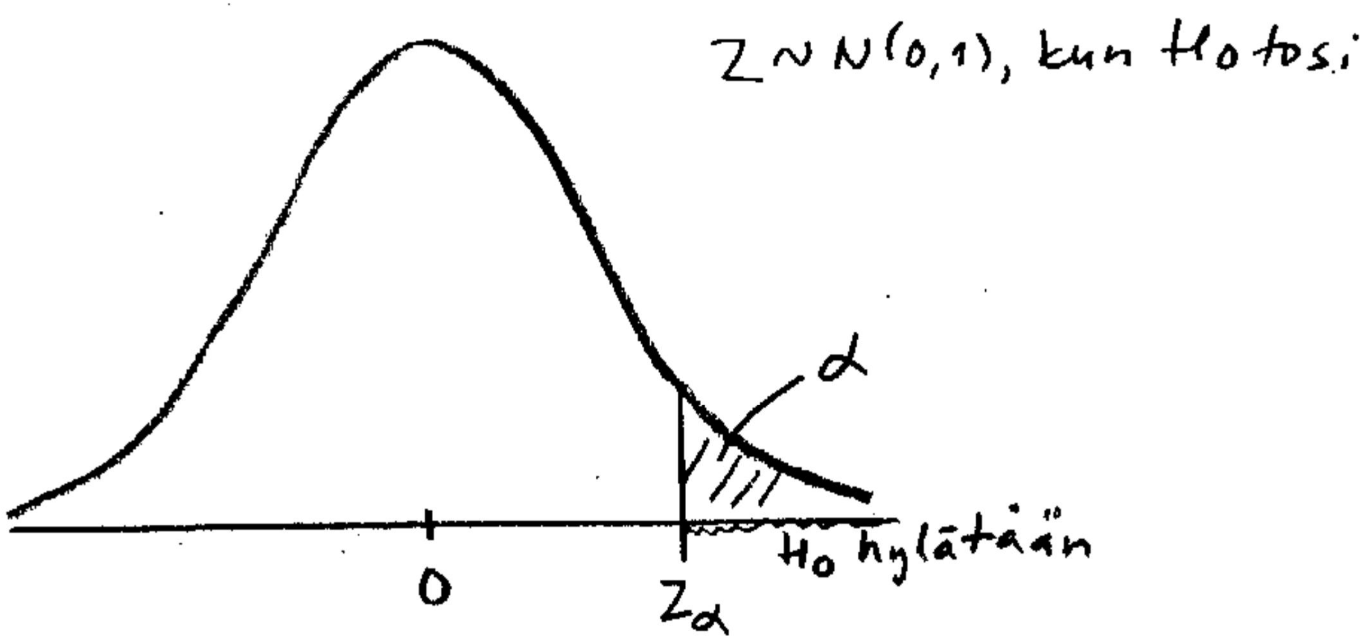
- Jos $H_1 : \mu \neq \mu_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $|z_{hav}| > z_{\alpha/2}$.



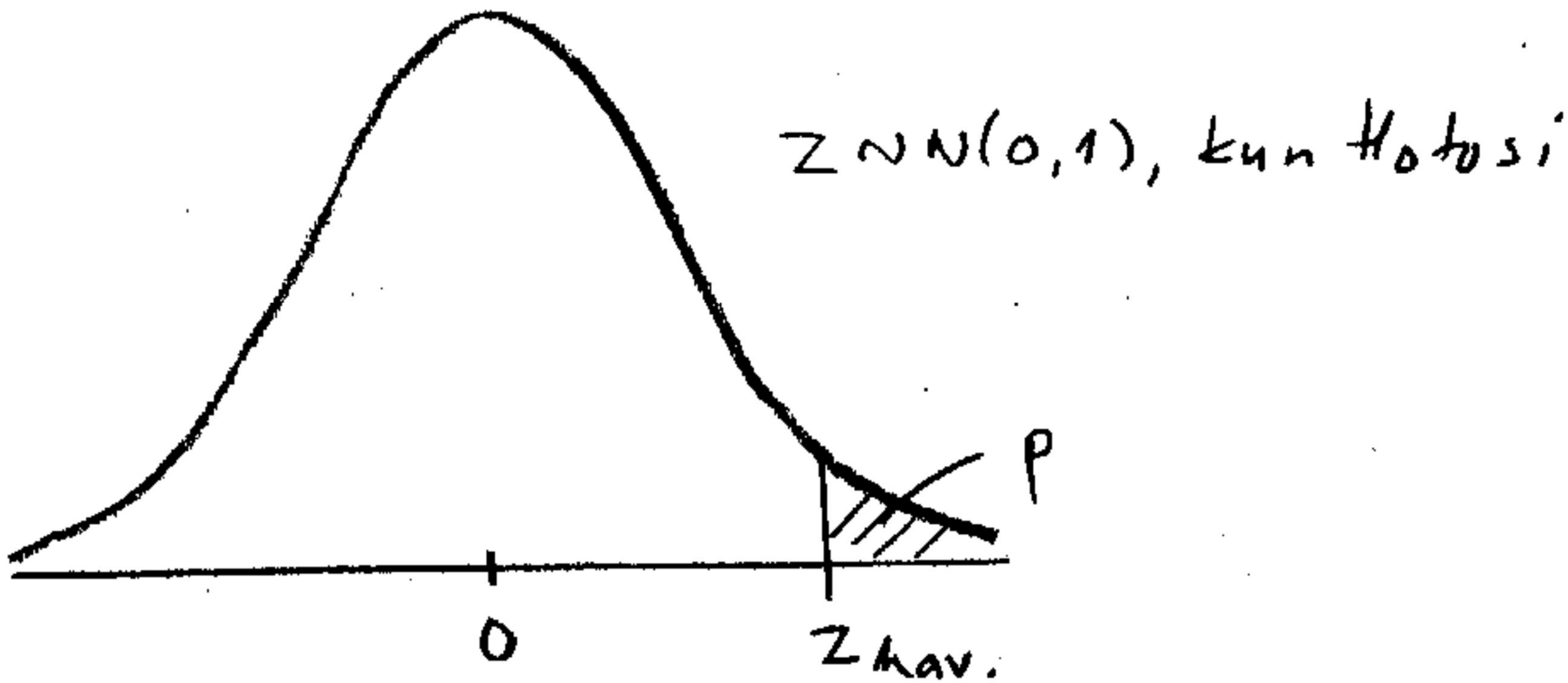
Pienin riskitaso p, jolla H_0 voidaan hylätä, on $2P(Z > |z_{hav}|)$.



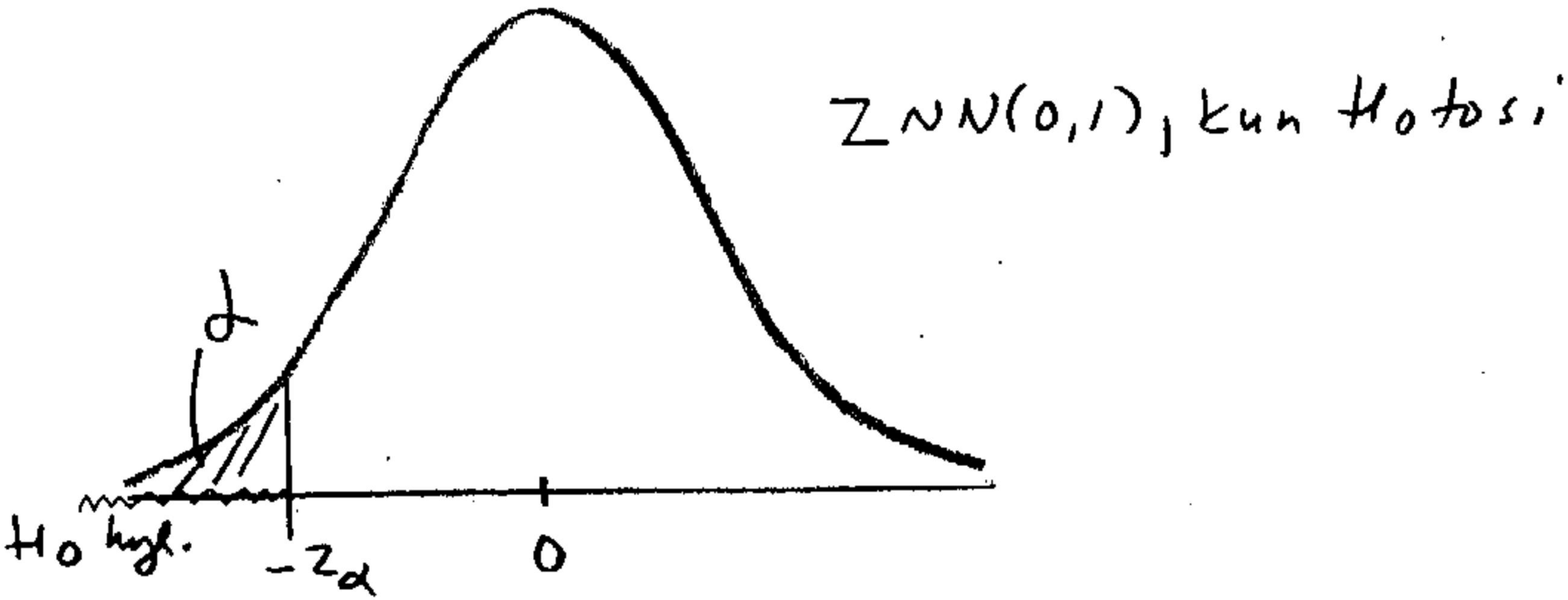
- Jos $H_1 : \mu > \mu_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $z_{hav} > z_\alpha$.



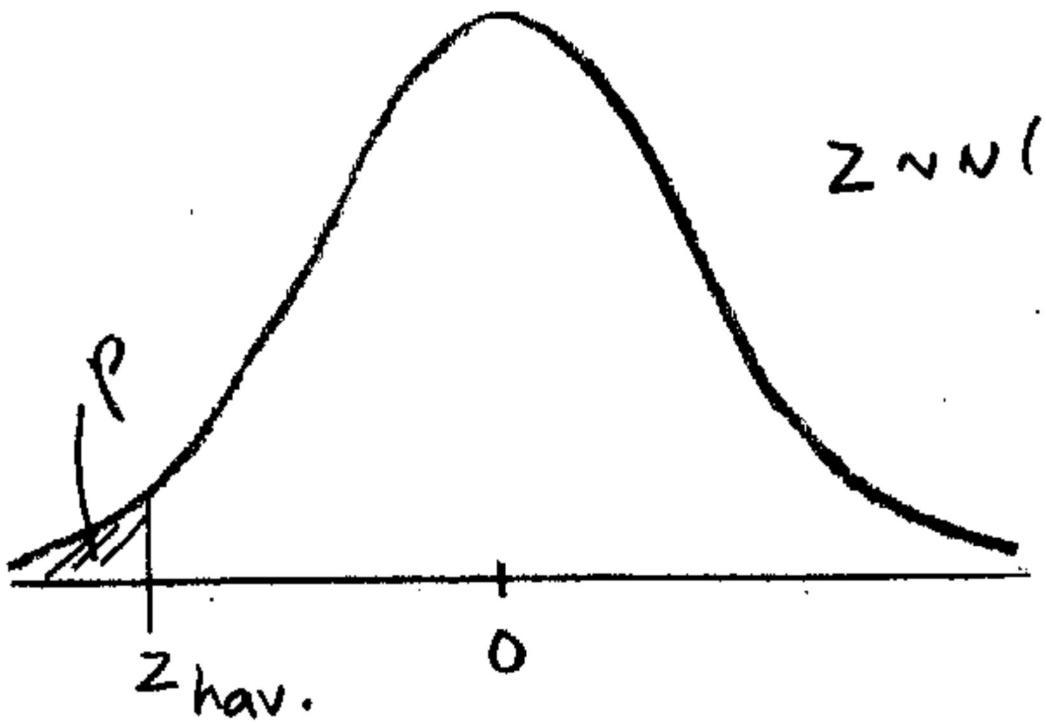
Pienin riskitaso p, jolla H_0 voidaan hylätä, on $P(Z > z_{hav})$.



- Jos $H_1 : \mu < \mu_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $z_{hav} < -z_{\alpha}$.



Pienin riskitaso p, jolla H_0 voidaan hylätä, on $P(Z < z_{hav})$.



$Z \sim N(0,1)$, kun H_0 tosi

Esim. Valtakunnallisessa matematiikan kokeessa tulospistemäärä on noudattanut normaalijakaumaa parametrein 64 ja 64. Eräänä vuonna erään koulun 54 oppilaan keskiarvo oli 68. Voidaanko koulua pitää poikkeavana?

$H_0 : \mu = 64$

$H_1 : \mu > 64$

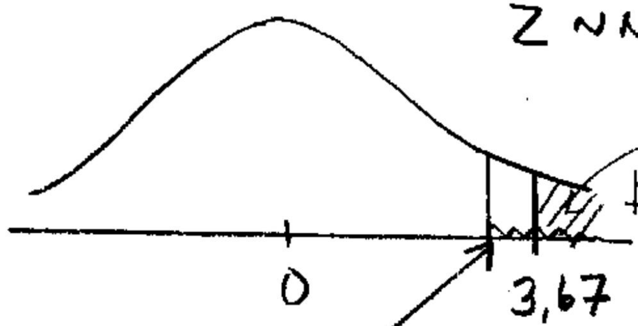
Jos H_0 on tosi, niin

$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$. Tässä siis $Z = \frac{\bar{X}-64}{\sqrt{64}/\sqrt{n}} \sim N(0,1)$.

Otoksesta saadaan

$$z_{hav.} = \frac{68-64}{\sqrt{64}/\sqrt{54}} = 3{,}67.$$

$Z \sim N(0,1)$



p-arvo

Hylätään 0,1 %:n riskitasolla.

$z_{0.001} = 3{,}09$

$$\text{p-arvo} = P(Z > 3{,}67) = 1 - \underbrace{\Phi(3{,}67)}_{\approx 1} \approx 0$$

Esim. Tutkija olettaa, että reagointiaika erääseen ärsykkeeseen on keskimäärin alle 6 sekuntia. Mitattiin 25 henkilön reagointiajat ja saatiin keskiarvoksi 5,2 s. Oletetaan, että reagointiaika on normaalisti jakautunut hajontana 2 s. Onko tutkija oikeassa?

$H_0 : \mu = 6$

$H_1 : \mu < 6$

Jos H_0 on tosi, niin

$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$. Tässä $\mu_0 = 6$, $\sigma = 2$.

Saadaan

$$z_{hav.} = \frac{5{,}2-6}{2/\sqrt{25}} = -2.$$

Pienin riskitaso, jolla H_0 voidaan hylätä, on $P(Z<-2) = 1-\Phi(2) = 1 - 0{,}9772 = 0{,}0228$. Jos valitaan riskitaso, joka on tätä suurempi, niin H_0 hylätään ja H_1 hyväksytään (tutkija oikeassa). Jos valitaan esim. 2 %:n riskitaso, H_0 hyväksytään. Päätellään, että tukija väärässä.

Päättely taulukkoarvojen perusteella:

- Jos valitaan 5 %:n riskitaso, niin harvinaisten arvojen raja on $-z_{0,05} = -1,65$. Koska $-2 < -1,65$, niin H_0 hylätään ja H_1 hyväksytään.
- Jos valitaan 1 %:n riskitaso, niin harvinaisten arvojen raja on $-z_{0,01} = -2,33$. Koska $-2 > -2,33$, niin H_0 hyväksytään.

MTTTP5, luento 29.11.2018

Kertausta

$$\boxed{H_0 : \mu = \mu_0}$$

Olk. $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 tunnettu. Jos H_0 on tosi, niin

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1). \qquad \text{Kaava 5.1}$$

Esim. Riisipussien pakkauskoneen pitäisi tuottaa keskimäärin kilon pusseja. Painon oletetaan vaihtelevan normaalijakauman mukaisesti keskihajonnan ollessa 2,1 g. Tutkitaan koneen toimivuutta. Punnitaan satunnaisesti valitut 20 pussia, joiden keskipainoksi saadaan 1001 g. Toimiiko kone oikein?

$H_0 : \mu = 1000$

$H_1 : \mu \neq 1000$

Jos H_0 on tosi, niin

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1).$$

Saadaan

$$z_{hav.} = \frac{1001-1000}{2{,}1/\sqrt{20}} = 2{,}13$$

Pienin riskitaso, jolla H_0 voidaan hylätä, on $2P(Z > 2{,}13) = 2(1-\Phi(2{,}13)) = 2(1 - 0{,}9834) = 0{,}0338$. Jos valitaan riskitaso, joka on tätä suurempi, niin H_0 hylätään ja H_1 hyväksytään (päätellään: kone ei toimi oikein). Jos valitaan esim. 1 %:n riskitaso, H_0 hyväksytään (päätellään: kone toimii oikein).

Päättely taulukkoarvojen perusteella:

- Jos valitaan 5 %:n riskitaso, niin harvinaisten arvojen raja on $z_{0,05/2} = 1,96$. H_0 hylätään, H_1 hyväksytään.
- Jos valitaan 1 %:n riskitaso, niin harvinaisten arvojen raja on $z_{0,01/2} = 2,57$. H_0 hyväksytään.

6.1.1 Yhden populaation odotusarvoa koskeva päättely (jatkoa)

$\boxed{H_0 : \mu = \mu_0}$

Olk. $X_1, X_2, \ldots, X_n$ on satunnaisotos N(μ, σ^2):sta, missä σ^2 tuntematon.

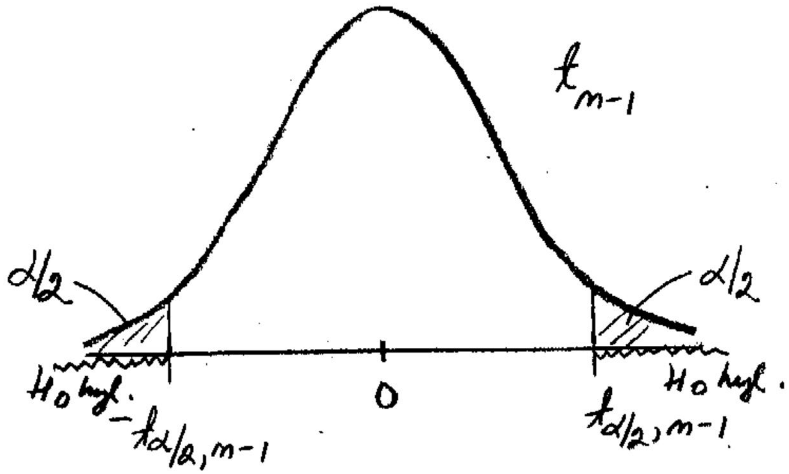
Jos H_0 on tosi, niin

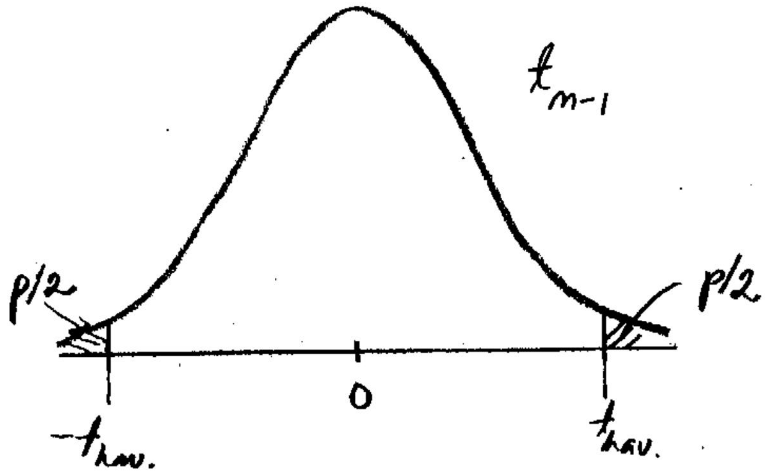
$$t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t_{n-1} \, .$$ Kaava 5.2

- Jos H_1 : $\mu \neq \mu_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $|t_{hav.}| > t_{\alpha/2,\, n-1}$.

  Pienin riskitaso p, jolla H_0 voidaan hylätä, on

  $2P(t_{n-1} > |t_{hav.}|)$.

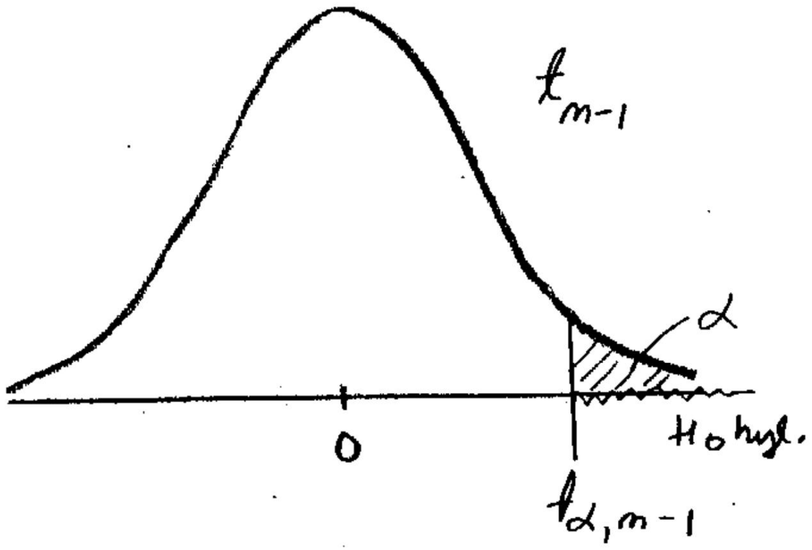


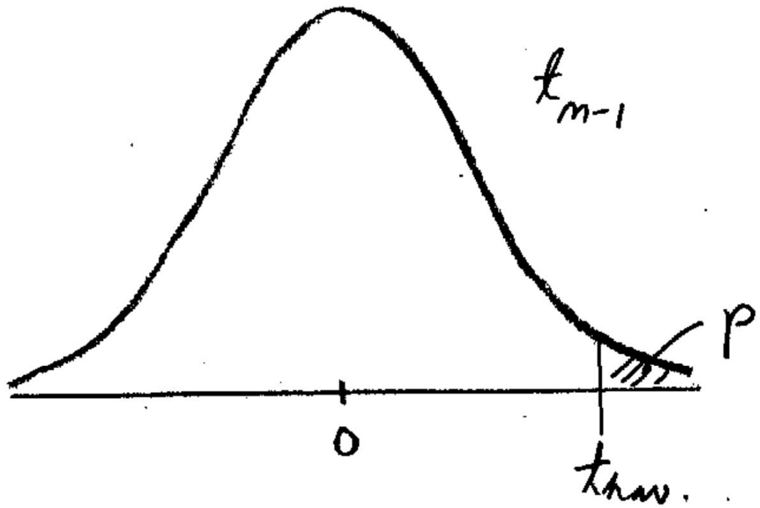


- Jos $H_1 : \mu > \mu_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $t_{hav.} > t_{\alpha,n-1}$

  Pienin riskitaso p, jolla H_0 voidaan hylätä, on

  $P(t_{n-1} > t_{hav})$.

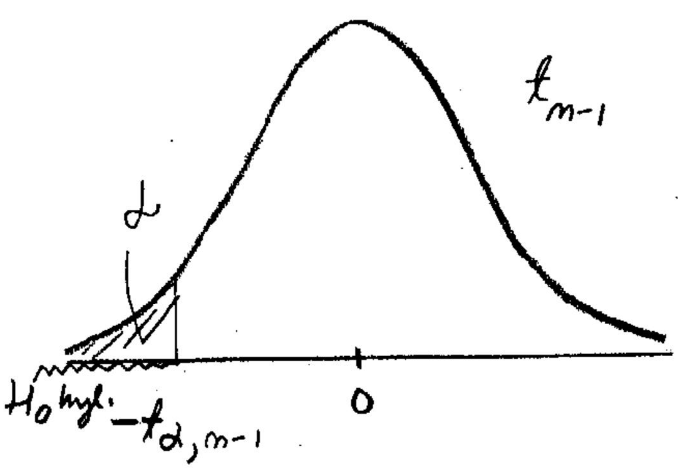


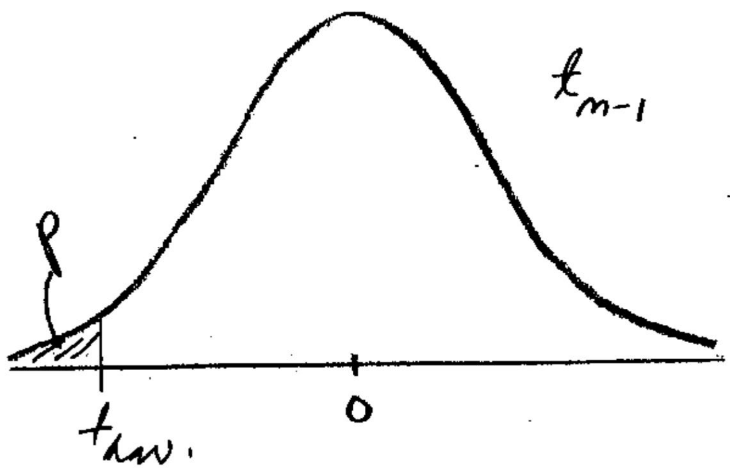


- Jos $H_1 : \mu < \mu_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $t_{hav.} < -t_{\alpha,n-1}$

  Pienin riskitaso p, jolla H_0 voidaan hylätä, on

  $P(t_{n-1} < t_{hav.})$.





Esim. 6.1.3 Lepakot paikallistavat hyönteisiä lähettämällä korkeataajuista ääntä. Kaiun kuulemiseen kuluvan ajan perusteella ne pystyvät paikallistamaan hyönteiset. Tutkijat arvelivat, että keskimääräinen tunnistusmatka voisi olla yli 35 cm. He keräsivät aineiston mitaten etäisyydet (cm), joista lepakot löysivät hyönteisiä. Mitatut etäisyydet olivat 62, 52, 68, 23, 34, 45, 27, 42, 83, 56, 40. Voidaanko saatujen tulosten perusteella pitää tutkijoiden arvioita oikeana?

$H_0 : \mu = 35$

$H_1 : \mu > 35$

Jos H_0 on tosi, niin

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}\ .$$

Saadaan

$\bar{x}$ = 48,36, s = 18,085, $t_{hav.} = \frac{48{,}36 - 35}{18{,}085/\sqrt{11}} =$ 2,450

Koska $t_{hav.} = 2,450 < t_{0,01;10} = 2,764$, H_0 hyväksytään 1 % riskitasolla.

Koska $t_{hav.} = 2,450 > t_{0,025;10} = 2,228$, H_0 hylätään 2,5 % riskitasolla

Siis $0,01 < p < 0,025$ (SPSS: 0,034/2 = 0,017)

SPSS-ohjeet:

Analyze -> Compare Means -> One-Sample T Test -> Test Variable Matka, Test Value 35

SPSS-tulostus

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Matka	11	48,36	18,085	5,453

One-Sample Test

	Test Value = 35			
	t	df	Sig. (2-tailed)	Mean Difference
Matka	2,451	10	,034	13,364

Esim. Valmistaja ilmoittaa kynttilöittensä keskimääräiseksi palamisajaksi 9,5 tuntia. Tutkit asiaa ja teet valmistajan kynttilöistä 6 kynttilän satunnaisotoksen. Mittaat näiden palamisajat ja saat keskiarvoksi 8,8 tuntia ja keskihajonnaksi 0,48 tuntia. Voitko uskoa valmistajan väitteen?

$$H_0 : \mu = 9,5$$

$$H_1 : \mu < 9,5$$

Jos H_0 on tosi, niin $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$. Nyt $\mu_0 = 9{,}5$, $n = 6$,

$\bar{x} = 8{,}8$, $s = 0{,}48$, $t_{hav.} = \frac{8{,}8-9{,}5}{0{,}48/\sqrt{6}} = -3{,}57$.

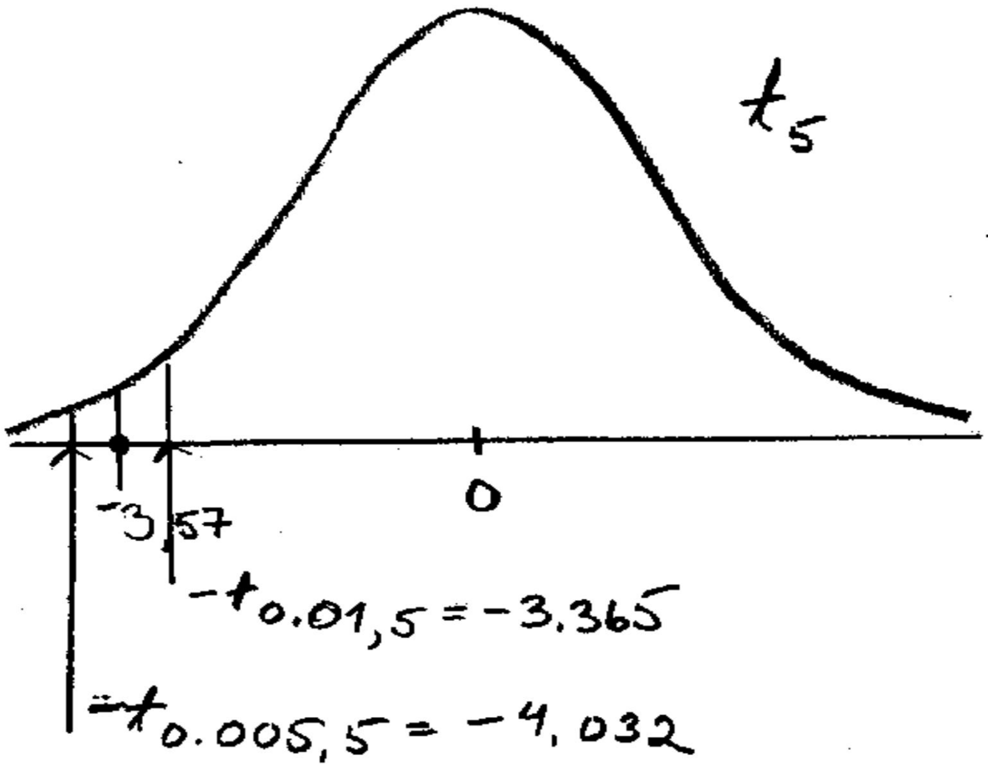
H_0 hyväksytään 0,5 % riskitasolla, koska

$t_{hav.} = -3{,}57 > -t_{0{,}005{,}5} = -4{,}032$

$t_{0{,}01{,}5} = 3{,}365$ H_0 hylätään 1 %:n riskitasolla, koska

$t_{hav.} = -3{,}57 < -t_{0{,}01{,}5} = -3{,}365$.

Siis $0{,}005 < p < 0{,}01$



Esim. Tutkittiin pH-mittarin toimivuutta. Mitattiin 14 neutraalin (pH = 7) liuoksen pH-arvot, joiksi saatiin 7,01, 7,04, 6,97, 7,00, 6,99, 6,97, 7,04, 7,04, 7,01, 7,00, 6,99, 7,04, 7,07, 6,97. Toimiiko mittari oikein?

$H_0 : \mu = 7$

$H_1 : \mu \neq 7$

Jos H_0 on tosi, niin $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$. Nyt $\mu_0 = 7$, $n = 14$, $\bar{x} = 7{,}01$, $s = 0{,}03162$, $t_{hav.} = \frac{7{,}01-7}{0{,}03162/\sqrt{14}} = 1{,}183$.

Jos valitaan α = 0,05, niin $t_{0,05/2;13} = t_{0,025;13} = 2{,}160$.

H_0 hyväksytään 5 %:n riskitasolla, koska

$-2{,}160 < t_{hav.} < 2{,}160$.

Pienin riskitaso, jolla H_0 voidaan hylätä, on > 0,05.

Laskurin http://vassarstats.net/ tulos
http://www.sis.uta.fi/tilasto/mtttp5/syksy2015/pH_mittari.pdf

SPSS-tulostus

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
pH	14	7,0100	,03162	,00845

One-Sample Test

	Test Value = 7					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
pH	1,183	13	,258	,01000	-,0083	,0283

Esim. Eräs yritys tarjoaa valmennuskurssia yliopistoon pyrkijöille. Yritys haluaa tutkia kurssinsa tehokkuutta. Tutkitaan pareja, joilla on samanlaiset lähtötiedot. Toinen osallistuu valmennuskurssille, toinen ei. Saadaan aineisto, jossa pyrkijöiden valintakoepisteet. Onko kurssi tehokas?

	Osallistui	Ei osallistunut	Erotus d
Pari 1	82	75	7
Pari 2	73	71	2
Pari 3	59	52	7
Pari 4	48	46	2
Pari 5	69	70	-1
Pari 6	93	83	10

Tutkitaan, onko erotus peräisin jakaumasta, jonka odotusarvo nolla. Siis

$H_0 : \mu_D = 0$

$H_1 : \mu_D > 0$

Jos H_0 on tosi, niin $t = \frac{\bar{D}}{s/\sqrt{n}} \sim t_{n-1}$. Nyt $\sum d_i = 27$,

$\sum d_i^2 = 207$, $\bar{d} = 4{,}5$ $SS_d = 207 - 27^2/6 = 85{,}5$, s = 4,135, $t_{hav.} = \frac{4{,}5}{4{,}135/\sqrt{6}} = 2{,}666$.

$t_{0,025;5} = 2{,}571$, $t_{0,01;5} = 3{,}365$, joten $0{,}01 < p < 0{,}025$

Ks. SPSS-tuloste

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
D	6	4,50 $= \bar{d}$	4,135 $= s$	1,688 $= s/\sqrt{n}$

Taulukosta

$t_{0,025,5} = 2,571 < 2,666$ H_0 hyl.

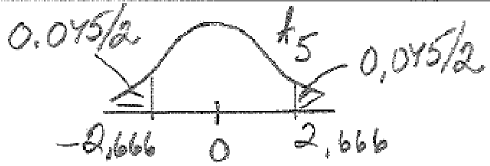
$t_{0,01,5} = 3,365 > 2,666$ H_0 hyv.

One-Sample Test

	Test Value = 0 $H_0: \mu = 0$					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
D	2,666	5	,045	4,500	,16	8,84

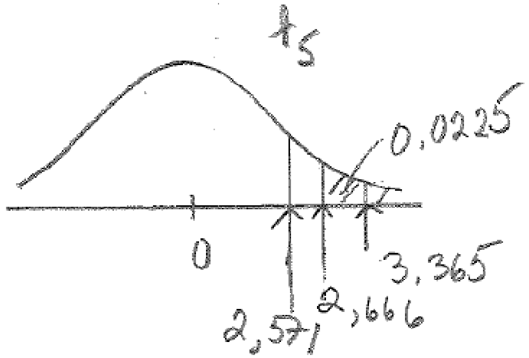
$0,01 < p\text{-arvo} < 0,025$

$$t = \frac{4,5 - 0}{4,135/\sqrt{6}}$$



1-suuntaisessa testissä

p-arvo = 0.045/2 = 0.0225



MTTTP5, luento 3.12.2018

6.1.2 Yhdessä populaatiossa tietyn tyyppisten alkioiden prosentuaalista osuutta koskeva päättely

$$\boxed{H_0 : \pi = \pi_0}$$

Oletetaan, että populaatiossa viallisia π %. Olkoon X_1, X_2, . . . , X_n satunnaisotos tästä populaatiosta.

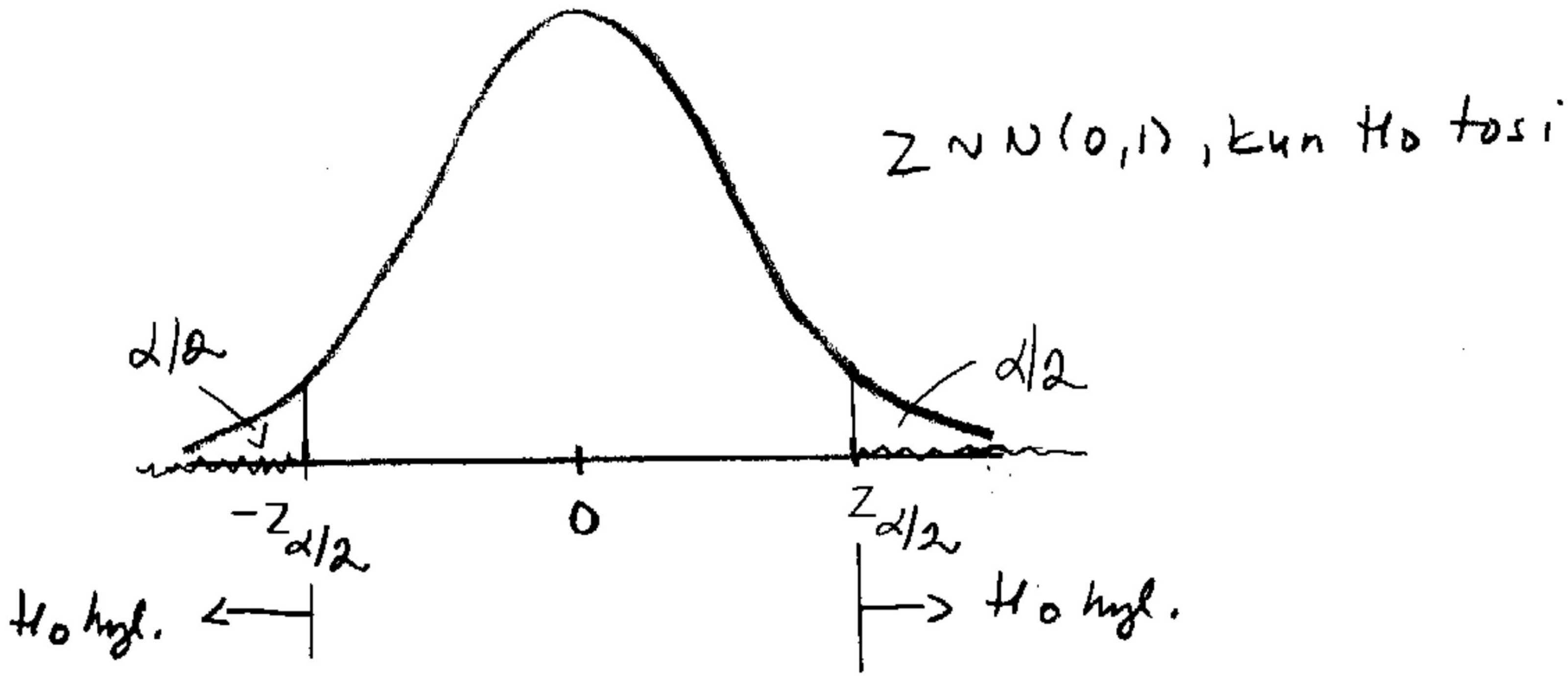
Jos H_0:n on tosi, niin viallisten prosenttiosuus otoksessa

$$p \sim N\left(\pi_0, \frac{\pi_0(100-\pi_0)}{n}\right), likimain,$$

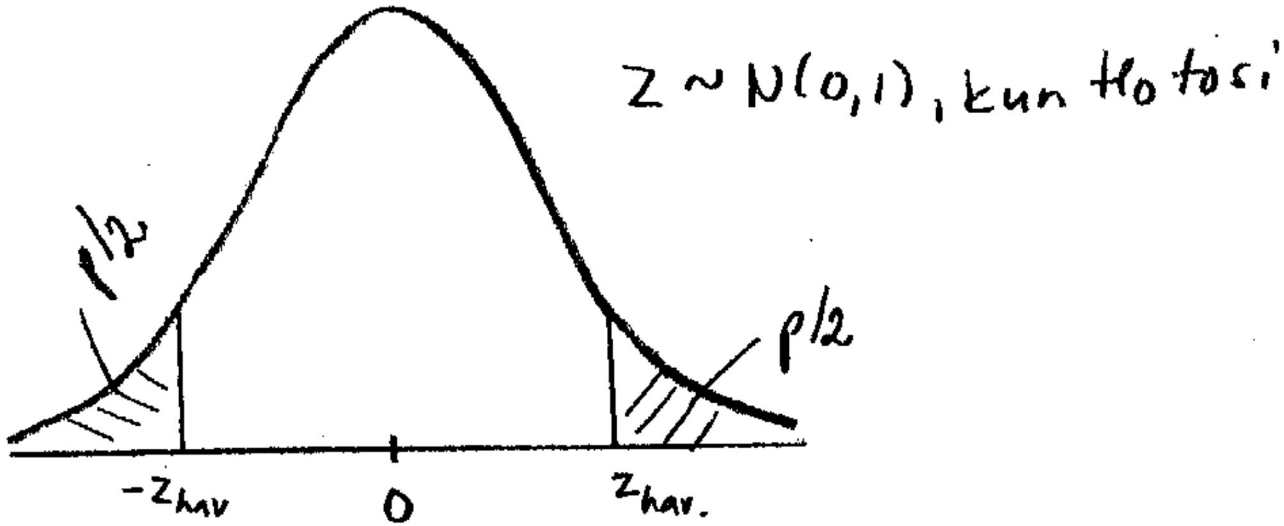
joten

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}} \sim N(0,1), likimain.$$

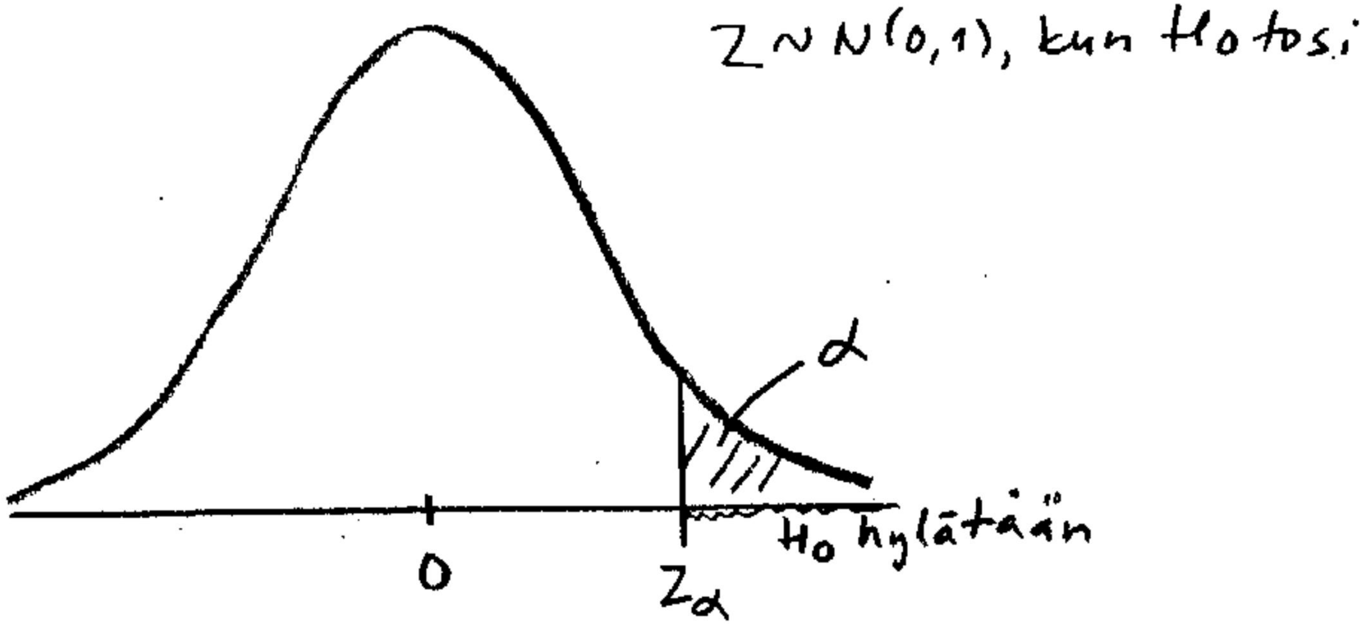
- Jos $H_1 : \pi \neq \pi_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $|z_{hav}| > z_{\alpha/2}$.



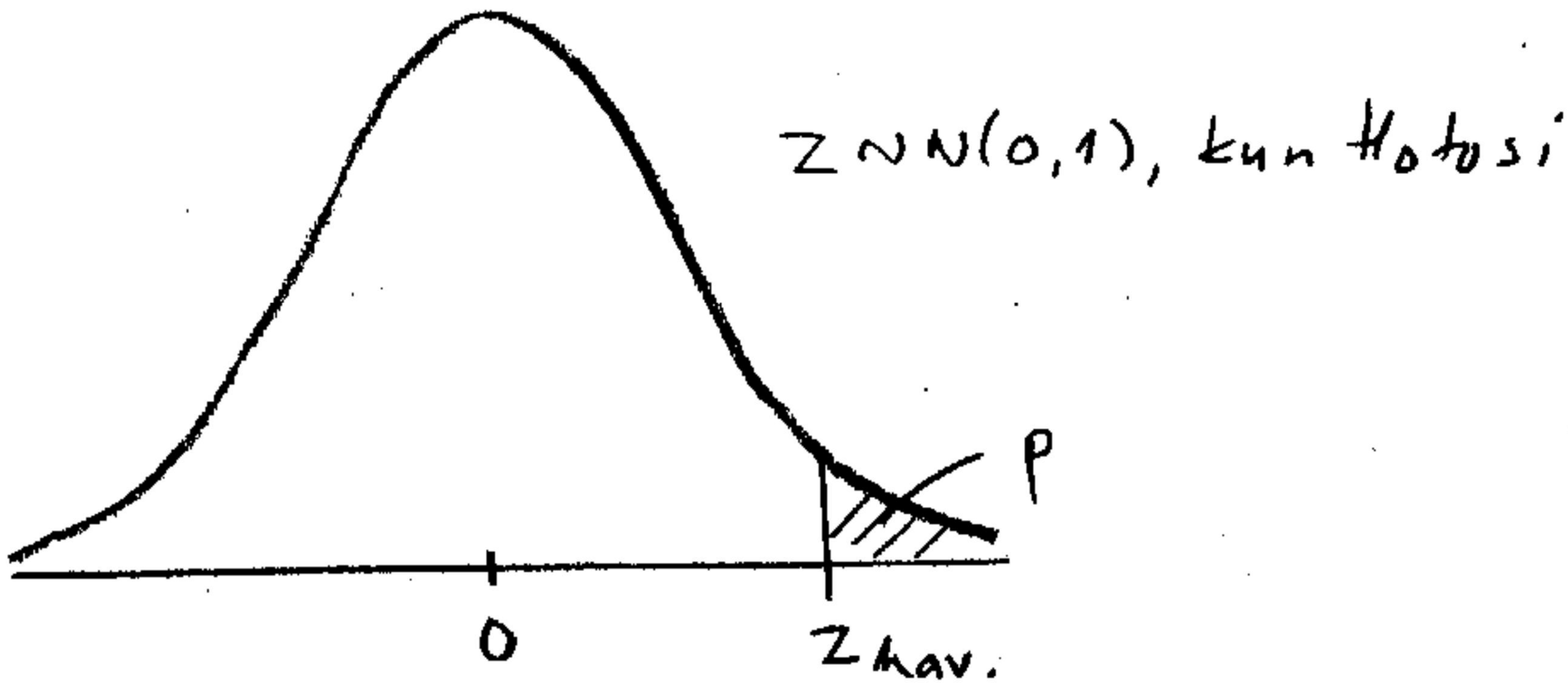
Pienin riskitaso p, jolla H_0 voidaan hylätä, on $2P(Z > |z_{hav}|)$.



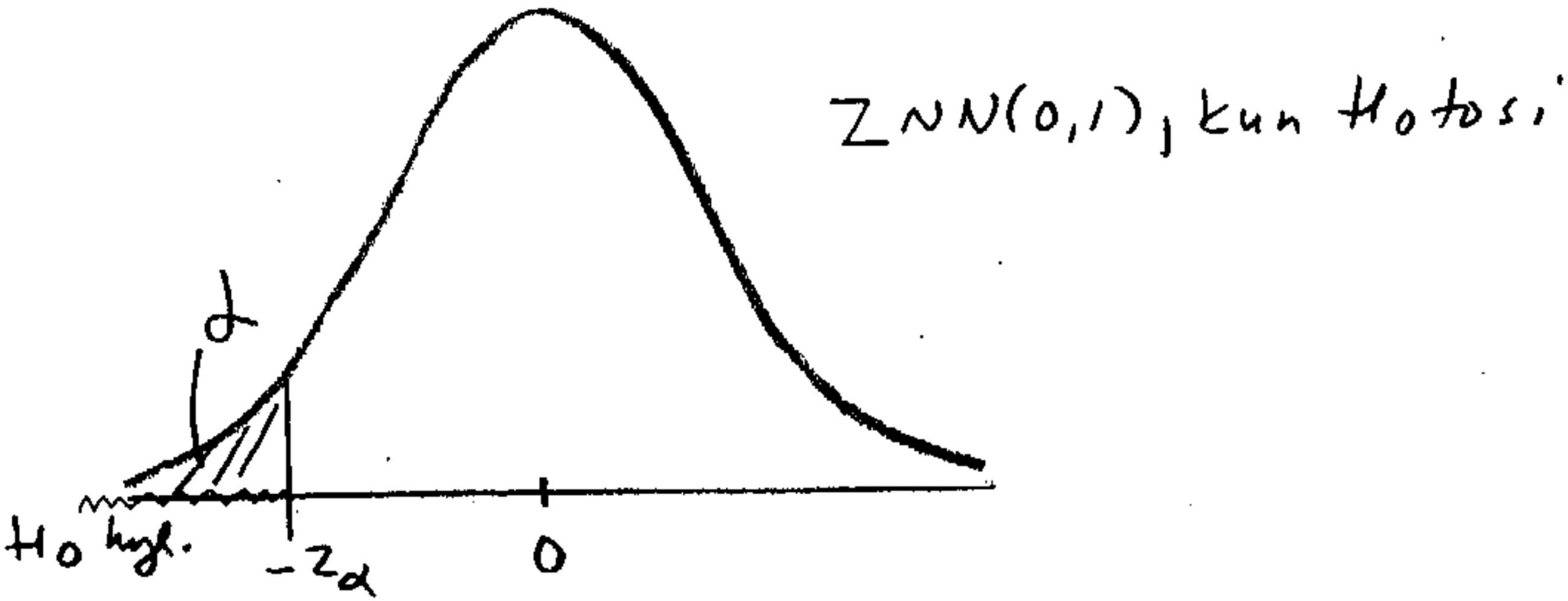
- Jos $H_1 : \pi > \pi_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $z_{hav} > z_\alpha$.



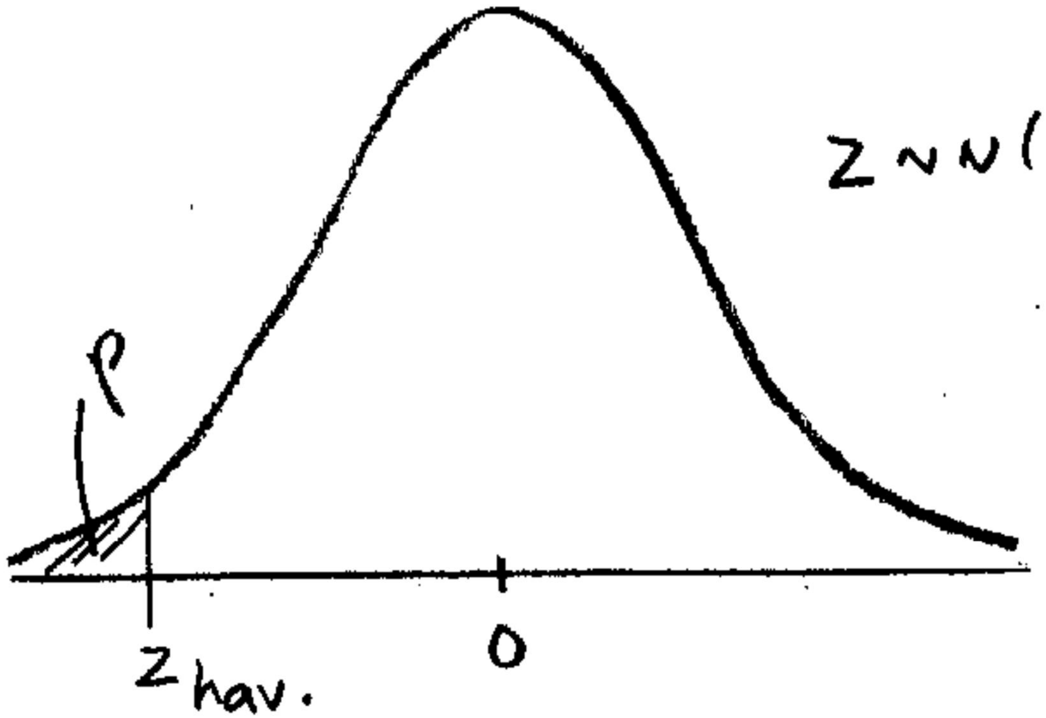
Pienin riskitaso p, jolla H_0 voidaan hylätä, on $P(Z > z_{hav})$.



- Jos H_1 : $\pi < \pi_0$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $z_{hav} < -z_\alpha$.



Pienin riskitaso p, jolla H_0 voidaan hylätä, on $P(Z < z_{hav})$.



$Z \sim N(0,1)$, kun H_0 tosi

Esim. 6.1.6. Ystäväsi väittää, että suomalaisista on 10% vasenkätisiä. Tutkit asiaa ja valitset satunnaisesti 400 suomalaista, joista vasenkätisiä on 47. Uskotko ystäväsi väitteen?

$$H_0 : \pi = 10$$

$$H_1 : \pi > 10$$

Jos Ho tosi,

$$Z = \frac{p - 10}{\sqrt{10(100 - 10)/n}} \sim N(0,1), likimain.$$

Otoksesta laskettu z:n arvo on

$$z = \frac{11{,}75 - 10}{\sqrt{10(100-10)/400}} = 1{,}17$$

Pienin riskitaso, jolla H_0 voidaan hylätä yksisuuntaisessa testissä, on $P(Z > 1{,}17) = 1 - \Phi(1{,}17) = 1 - 0{,}8790 = 0{,}121$. Uskotaan siis ystävän väite.

Jos valitaan 5 %:n riskitaso, niin yksisuuntaisessa testissä kriittinen arvo on $z_{0,05} = 1{,}64$ (koska $\Phi(1{,}64) = 0{,}9495$) ja kaksisuuntaisessa testissä $z_{0,05/2} = 1{,}96$ (koska $\Phi(1{,}96) = 0{,}975$).

Esim. Ruletissa on 37 numeroa, joista pyöritettäessä jokaisella pitäisi olla sama todennäköisyys tulla tulokseksi. Pelipaikka voittaa numerolla nolla. Rulettia pyöritetään 3700 kertaa. Saadaan nollia 140 eli 3,78 %. Toimiiko ruletti oikein?

$$H_0 : \pi = 100 \cdot 1/37$$

$$H_1 : \pi > 100 \cdot 1/37$$

Jos H_0 tosi, niin

$$Z = \frac{p - \frac{100}{37}}{\sqrt{\frac{100}{37}\left(100 - \frac{100}{37}\right)/n}} \sim N(0,1), likimain.$$

Otoksesta laskettu z:n arvo on

$$z = \frac{\frac{140}{37} - \frac{100}{37}}{\sqrt{\frac{100}{37}(100 - \frac{100}{37})/3700}} \approx 4{,}06$$

Pienin riskitaso, jolla H_0 voidaan hylätä yksisuuntaisessa testissä, on $P(Z > 4{,}06) = 1 - \Phi(4{,}06) \approx 0$. Ruletti ei toimi oikein.

Esim. 6.1.7. Öljy-yhtiö väittää, että erään kaupungin asunnoista 20 % lämmitetään öljyllä. Onko kuitenkin syytä olettaa, että vähemmän kuin viidesosa asunnoista lämmitetään öljyllä, jos 1000 satunnaisesti valitusta kaupungin asunnosta vain 160 lämmitettiin öljyllä?

Nyt

$$H_0\colon \pi = 20$$
$$H_1\colon \pi < 20.$$

Jos H_0 tosi, niin

$$Z = \frac{p - 20}{\sqrt{20(100 - 20)/n}} \sim N(0,1), likimain.$$

Aineiston perusteella testisuureen arvoksi saadaan

$$z = \frac{16 - 20}{\sqrt{20(100 - 20)/1000}} = -3{,}16 < -z_{0,001} = -3{,}08$$

H_0 hylätään 0,1 %:n riskitasolla. Päätellään, että alle viidesosa lämmitetään öljyllä.

Pienin riskitaso, jolla H_0 voidaan hylätä, on $P(Z < -3{,}16) = 1 - \Phi(3{,}16) = 1 - 0{,}9992 = 0{,}0008$.

Esim. Aiempien tutkimusten perusteella 20 % kahvin juojista valitsi kahvin hinnan perusteella. Haluttiin selvittää, onko ostokäyttäytymisessä tapahtunut muutosta. Kysyttiin valinnan perusteita 100 kahvin juojalta. Päätellään muutosta tapahtuneen, jos p > 28 % (p = otoksessa valintansa hinnan perusteella tekevien %-osuus). Mikä on tällöin testissä käytetty α?

H_0: $\pi = 20$

Jos H_0 tosi, niin $p \sim N\left(20, \frac{20(100-20)}{100}\right)$, $likimain$.

Käytetty riskitaso

α = P(H_0 hylätään, kun se on tosi)

= P(p > 28, kun π = 20)

= 1 – P(p ≤ 28, kun π = 20)

$$= 1 - \Phi\left(\frac{28-20}{\sqrt{20(100-20)/100}}\right) = 1 - \Phi(2) = 0{,}0228$$

Esim. Verkkokauppa pyrkii toimimaan siten, että tuotteet lähetetään kolmen työpäivän kuluttua tilauksesta. Tämä ei kuitenkaan aina ole mahdollista. Verkkokauppa haluaa toimia siten, että satunnaisia viivästymisiä voi olla 10 %. Viimeisen kuukauden ajalta satunnaisesti valituista 150 tilauksesta 21 lähetettiin myöhässä. Verkkokauppa toteaa, että lähetyksissä ei ole 10 % suurempaa viivästymistä. Millä riskitasolla päättely on voitu tehdä?

H_0: $\pi = 10$
H_1: $\pi > 10$

Jos Ho tosi,

$$Z = \frac{p - 10}{\sqrt{10(100 - 10)/n}} \sim N(0,1), likimain.$$

Aineiston perusteella testisuureen arvoksi saadaan

$$z = \frac{14-10}{\sqrt{10(100-10)/150}} = 1{,}63$$

Pienin riskitaso, jolla H_0 voidaan hylätä, on P(Z > 1,63) = 1 - Φ(1, 63) = 1 – 0,9484 = 0,0516.

Päättely on voitu tehdä esim. 5 %:n riskitasolla.

6.1.3 Kahden jakauman sijainnin vertailu

$$\boxed{H_0 : \mu_1 = \mu_2} \text{ tai } \boxed{H_0\ \mu_1 - \mu_2 = 0}$$

$X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu_1, \sigma_1^2)$:sta,

$Y_1, Y_2, \ldots, Y_m$ on satunnaisotos $N(\mu_2, \sigma_2^2)$:sta.

Oletetaan, että varianssit tunnettuja ja satunnaisotokset riippumattomia.

Kun H_0 on tosi, niin

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

Tätä käytetään testisuureena, päättely kuten aiemminkin normaalijakaumaa noudattavien testisuureiden tapauksissa.

Esim. Hedelmien viljelijä tietään pitkäaikaisen seurannan perusteella, että samanlaisissa olosuhteissa viljeltyinä lajike A ja B tuottavat keskimäärin yhtä suuren sadon. Lisäksi molempien lajikkeiden sadon vaihtelu on normaalisti jakautunut varianssina 66 g^2. Viljelijä ottaa käyttöön uudet viljelyalueet. Hän epäilee, että uudet kasvumaat eivät enää tuota keskimäärin samansuuruisia satoja. Hän valitsee satunnaisesti molempien lajikkeista sadoista 25 saaden lajikkeen A sadon keskiarvoiksi 116 g ja lajikkeen B sadon keskiarvoksi 111 g. Onko viljelijän epäily aiheellista?

Aseta tilanteeseen sopiva H_0 ja H_1. Suorita testaus 1 %:n riskitasolla. Laske pienin riskitaso, jolla H_0 voidaan hylätä.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Kun H_0 tosi, niin $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$,

$$z_{hav.} = \frac{116 - 111}{\sqrt{\frac{66}{25} + \frac{66}{25}}} = 2{,}18.$$

Koska $z_{0,01/2} = 2{,}58$, niin H_0 hyväksytään 1 %:n riskitasolla. Viljelijän epäily ei aiheellinen. Pienin riskitaso, jolla H_0 voidaan hylätä, on

$$\begin{aligned} 2(1 - P(Z < 2{,}18)) &= 2(1 - \Phi(2{,}18)) \\ &= 2(1 - 0{,}9854) \\ &= 0{,}0292. \end{aligned}$$

MTTTP5, luento 4.12.2018

6.1.3 Kahden jakauman sijainnin vertailu (jatkoa)

Tutkimustilanteita

y = neliöhinta
x = sijainti (2 aluetta)

y = lepopulssi
x = sukupuoli

y = musikaalisuus
x = sukupuoli

y = kaksion koko
x = sijainti (keskusta/ei-keskusta)

Y = tehopisteet
x = pelipaikka (puolustaja/hyökkääjä)

y = lumilaudan hinta
x = kenelle tarkoitettu (miehille/naisille)

$\boxed{H_0 : \mu_1 = \mu_2}$ tai $\boxed{H_0\ \mu_1 - \mu_2 = 0}$

$X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu_1, \sigma_1^2)$:sta,

$Y_1, Y_2, \ldots, Y_m$ on satunnaisotos $N(\mu_2, \sigma_2^2)$:sta.

Oletetaan, että varianssit tunnettuja ja satunnaisotokset riippumattomia. Kun H_0 on tosi, niin

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1).$$

Tätä käytetään testisuureena, päättely kuten aiemminkin normaalijakaumaa noudattavien testisuureiden tapauksessa.

Variansseja ei useinkaan tunneta. Jos oletetaan ne tuntemattomiksi mutta yhtä suuriksi, niin H_0:n ollessa tosi

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n + m - 2),$$

$$s^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$$

Tätä käytetään testisuureena, päättely tehdään kuten aiemminkin t-jakaumaa noudattavien testisuureiden tapauksessa.

Oletusta varianssien yhtäsuuruudesta voidaan testata. SPSS tulostaa tähän liittyvän p-arvon.

Esim. 6.1.9 Lepopulssi

Group Statistics

	Sukupuoli	N	Mean	Std. Deviation	Std. Error Mean
Lepopulssi	Mies	$n = 44$	70,6364	11,27684 $= s_1$	1,70005
	Nainen	$m = 36$	77,5556	9,80800 $= s_2$	1,63467

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		$H_0: \sigma_1^2 = \sigma_2^2$		$H_0: \mu_1 = \mu_2$	
		F	Sig.	t	df
Lepopulssi	Equal variances assumed	,125	,725	-2,893	78
	Equal variances not assumed			-2,934	77,685

Independent Samples Test

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
Lepopulssi	Equal variances assumed	,005	-6,91919	2,39180
	Equal variances not assumed	,004	-6,91919	2,35845

$$s\sqrt{\frac{1}{n}+\frac{1}{m}}$$

$$s^2=\frac{(n-1)s_1^2+(m-1)s_2^2}{n+m-2}$$

Analyze -> Compare Means -> Independent-Samples T Test -> Test Variable(s): Lepopulssi, Grouping Variable: Sukupuoli: Group 1: 1 (Mies), Group 2: 2 (Nainen)

Esim. 6.1.10 Testi lasten kehityshäiriön tunnistamiseen

Suoritusajat testissä ryhmittäin:

Normaali	204, 218, 197, 183, 227, 233, 191
Kehityshäiriö	243, 228, 261, 202, 343, 242, 220, 239

$$H_0 : \mu_N = \mu_K$$

$$H_1 : \mu_N < \mu_K$$

$\sum x_i$ = 204 +... + 191 = 1453

$\sum x_i^2$ = 204^2 + ... + 191^2 = 303737

SS_N = 303737 – $7 \cdot (1453/7)^2$ = 2135,714

$\sum y_i$ = 243 +... + 239 = 1978

$\sum y_i^2$ = 243^2 + ... + 239^2 = 501692

SS_K = 501692 – $8 \cdot (1978/8)^2$ = 12631,5

$$s^2 = \frac{2135{,}714 + 12631{,}5}{7 + 8 - 2} = 1135{,}94, \; s = 33{,}7$$

$$t_{hav.} = \frac{\frac{1453}{7} - \frac{1978}{8}}{33{,}7\sqrt{\frac{1}{7} + \frac{1}{8}}} = -2{,}28$$

$-t_{0,01;13} = -2{,}65 < t_{hav.} < -2{,}16 = -t_{0,025;13}$

H_0 voidaan hylätään 2,5 %:n riskitasolla, mutta ei 1 %:n riskitasolla.

SPSS-tulos

Group Statistics

	ryhmä	N	Mean	Std. Deviation	Std. Error Mean
testi	normaali	$m_N =$ 7	207,5714 $\bar{x}_N$	18,86670 s_N	7,13094
	kehityshäiriö	$m_K =$ 8	247,2500 $\bar{x}_K$	42,47941 s_K	15,01874

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		$H_0: \sigma_N^2 = \sigma_K^2$		$H_0: \mu_N = \mu_K$	
		F	Sig.	t	df
testi	Equal variances assumed	,926	,353	-2,275	13
	Equal variances not assumed	H_0 hyv.		-2,387	9,923

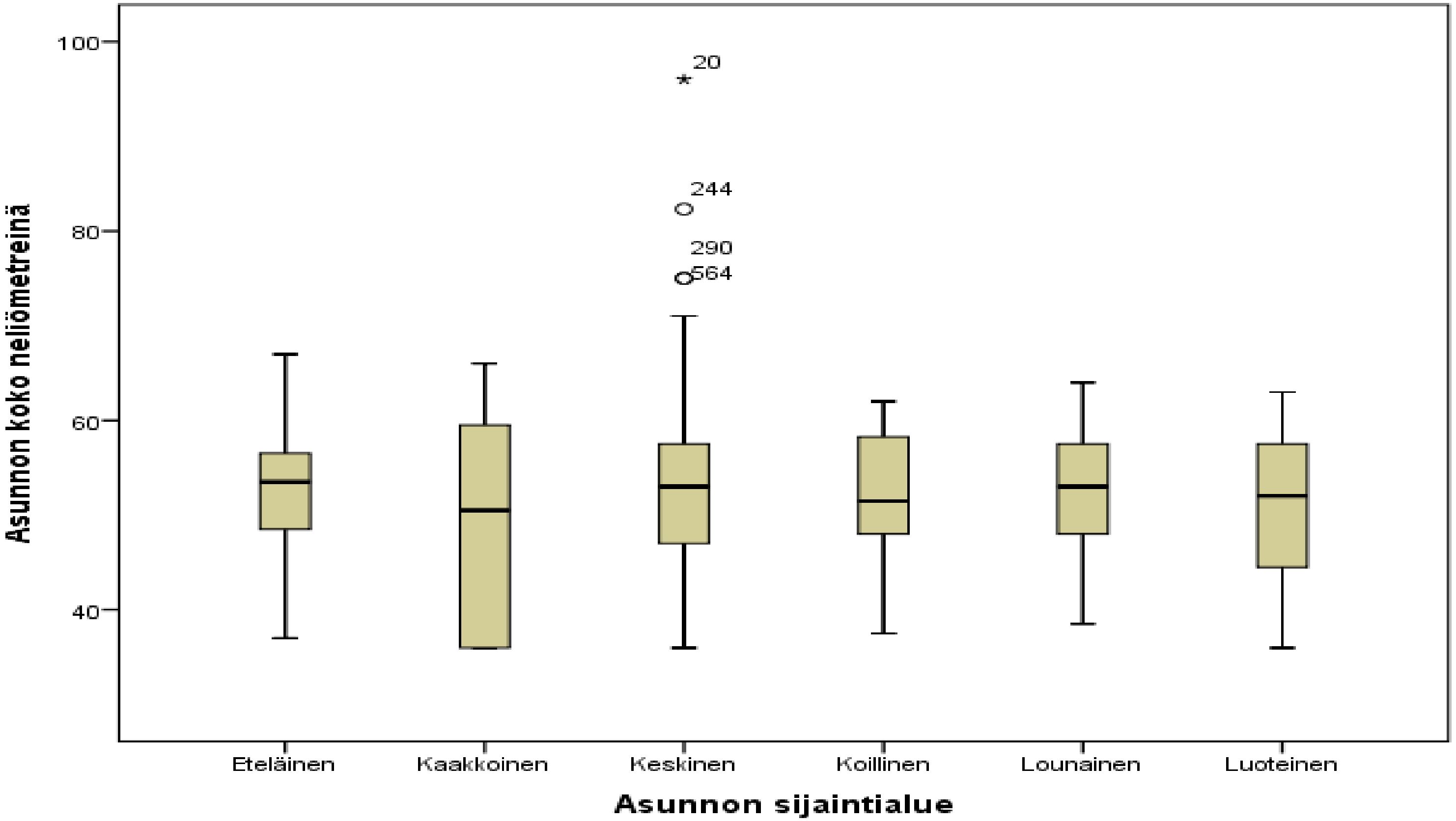
Independent Samples Test

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
testi	Equal variances assumed	,041	-39,67857	17,44332
	Equal variances not assumed	,038	-39,67857	16,62567

$H_1: \mu_N < \mu_K$

p-arvo $0,041/2 = 0,0205$

Esim. Ovatko kaksiot keskimäärin erikokoisia Tampereen eri alueilla? Aineisto: Tre_myydyt_kaksiot_2016.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/



Ei eroja:

Group Statistics

	Asunnon sijaintialue	N	Mean	Std. Deviation
Asunnon koko neliömetreinä	Eteläinen	53	52,46	5,965
	Koillinen	44	52,09	6,379

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Asunnon koko neliömetreinä	Equal variances assumed	,248	,620	,296	95	,768
	Equal variances not assumed			,294	89,211	,769

On eroja:

Group Statistics

	Asunnon sijaintialue	N	Mean	Std. Deviation
Asunnon koko neliömetreinä	Kaakkoinen	112	49,75	10,217
	Keskinen	277	52,56	7,888

Independent Samples Test

		Levene's Test for Equality of Variances		t		
		F	Sig.	t	df	Sig. (2-tailed)
Asunnon koko neliömetreinä	Equal variances assumed	25,812	,000	-2,905	387	,004
	Equal variances not assumed			-2,608	167,057	,010

Esim. Eräs yritys valmistaa 10 metrin teräskaapeleita tehtaissa X ja Y. Tarkasteltiin kaapeleiden murtolujuuksia (kilonewtoneina). Tehtaan X tuotannosta valittiin satunnaisesti 9 ja tehtaan Y tuotannosta 16 kaapelia, joiden murtolujuudet mitattiin. Saatiin tulokset:

Tehdas X: $\bar{x} = 30{,}11, \sum(x_i - \bar{x})^2 = 0{,}8013$

Tehdas Y: $\bar{y} = 29{,}63, \sum(y_i - \bar{y})^2 = 3{,}0206$

Asetetaan

$$H_0 : \mu_X = \mu_Y$$

$$H_1 : \mu_X \neq \mu_Y$$

Nyt

$$s^2 = \frac{0{,}8013 + 3{,}0206}{9 + 16 - 2} = 0{,}1662, s = 0{,}4076$$

$$t_{hav.} = \frac{30{,}11 - 29{,}63}{0{,}4076\sqrt{\frac{1}{9} + \frac{1}{16}}} = 2{,}83$$

Pienin riskitaso, jolla H_0 voidaan hylätä, on 0,0047x2 = 0,0094, ks. http://stattrek.com/online-calculator/t-distribution.aspx tai http://onlinestatbook.com/2/calculators/t_dist.html

Keskimääräisissä murtolujuuksissa on siis eroja.

Yhteenveto kurssin sisällöstä, joitain esimerkkejä

Todennäköisyyslaskentaa

- satunnaisilmiö ja tapahtuma
- klassinen todennäköisyys
- todennäköisyyslaskennan aksioomat ja laskusääntöjä
- kombinatoriikkaa

Todennäköisyysjakaumia

- satunnaismuuttuja ja todennäköisyysjakauma
- diskreetti satunnaismuuttuja
- jatkuva satunnaismuuttuja
- odotusarvon ja varianssin ominaisuuksia
- joitain todennäköisyysjakaumia

    Bernoulli-jakauma

    binomijakauma

    diskreetti tasajakauma

    jatkuva tasajakauma

    normaalijakauma

    t-jakauma

Satunnaisotos, otossuure ja otosjakauma

Esim. Jos populaatiossa viallisia π %, niin viallisten prosenttiosuus otoksessa

$$p \sim N\left(\pi, \frac{\pi(100-\pi)}{n}\right), likimain.$$

Esim.

Jos $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, niin

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Esim.

Olkoot $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta sekä otokset riippumattomia.

Tällöin

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Parametrien estimointi

- piste-estimointi
- estimaattori, estimaattorin ominaisuudet, estimaattorin keskivirhe
- luottamusvälit (väliestimointi)

μ:lle

Aineisto Lumilaudat.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

- Miesten lumilautojen keskihinta

One-Sample Statistics

Kenelle lauta on tarkoitettu		N	Mean	Std. Deviation	Std. Error Mean
Miehet	Laudan hinta euroina	51	522,84	136,219	19,074
Naiset	Laudan hinta euroina	18	511,06	102,737	24,215
Lapset	Laudan hinta euroina	11	278,18	34,005	10,253

One-Sample Test

		Test Value = 0	
		95% Confidence Interval of the Difference	
Kenelle lauta on tarkoitettu		Lower	Upper
Miehet	Laudan hinta euroina	484,53	561,16
Naiset	Laudan hinta euroina	459,97	562,15
Lapset	Laudan hinta euroina	255,34	301,03

$\bar{X} \pm t_{0,05/2;50} \cdot 136,219/\sqrt{51}$

$522,84 \pm 2 \cdot 19,074$

π:lle

Aineisto Audi_A6.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

- Diesel-autojen osuus

Käyttövoima

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Bensiini	99	74,4	74,4	74,4
	Diesel	34	25,6	25,6	100,0
	Total	133	100,0	100,0	

$$p \pm z_{\alpha/2}\sqrt{p(100-p)/n}$$

$$25{,}6 \pm 1{,}96\sqrt{25{,}6 \cdot 74{,}4/133}$$

$$25{,}6 \pm 7{,}4$$

$$(18{,}2 \,,\ 33{,}0)$$

$(\mu_1 - \mu_2)$:lle

Aineisto Plasma.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

- Onko keskimääräisissä kolesterolimäärissä eroja miehillä ja naisilla?

Group Statistics

	sukupuoli	N	Mean	Std. Deviation	Std. Error Mean
päivässä saatujen kalorien määrä	mies	42	2155,786	916,5653	141,4291
	nainen	273	1741,404	620,2701	37,5405

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of
		$H_0: \sigma_1^2 = \sigma_2^2$		
		F	Sig.	t
päivässä saatujen kalorien määrä	Equal variances assumed	2,007	,158	3,750.
	Equal variances not assumed			2,832

Independent Samples Test

		t-test for Equality of Means		
		df	Sig. (2-tailed)	Mean Difference
päivässä saatujen kalorien määrä	Equal variances assumed	313	,000	414,3821
	Equal variances not assumed	46,946	,007	414,3821

Independent Samples Test

$s\sqrt{\frac{1}{n}+\frac{1}{m}}$

		t-test for Equality of Means		
			95% Confidence Interval of the Difference	
		Std. Error Difference	Lower	Upper
päivässä saatujen kalorien määrä	Equal variances assumed	110,4912	196,9827	631,7815
	Equal variances not assumed	146,3266	120,0019	708,7622

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

$$\bar{x} - \bar{y} \pm t_{\alpha/2, n+m-2} \cdot s\sqrt{\frac{1}{n}+\frac{1}{m}}$$

$$t_{0,05/2;313} = 1,96$$

$$414,3821 \pm 1,96 \cdot 110,4912$$

Hypoteesien testaus

H_0: $\mu = \mu_0$

- Onko kynttilöiden keskimääräinen palamisaika 24 h

  20 kynttilän palamisajat minuutteina:

  1426, 1438, 1425, 1435, 1432, 1434, 1427, 1441, 1439, 1427, 1435, 1435, 1427, 1432, 1445, 1439, 1429, 1435, 1434, 1447

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Palamisaika	20	1434,01	$S=$ 6,149	1,375 $s/\sqrt{n}$

One-Sample Test

	Test Value = 1440					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Palamisaika	-4,354	19	,000 $<0,001$	-5,987	-8,86	-3,11

$H_0: \mu = 1440$

$H_1: \mu < 1440$

$t_{0,005,19} = 2,861$

$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, kun H₀ tosi,

H_0: $\pi = \pi_0$

Aineisto Saidit.sav sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Syntyykö tyttöjä ja poikia yhtä paljon?

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	poika	65	54,2	54,2 = p	54,2
	tyttö	55	45,8	45,8	100,0
	Total	120	100,0	100,0	

$H_0 : \pi = 50$

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}}$$

$$Z_{hav.} = \frac{54{,}2 - 50}{\sqrt{50 \cdot 50/120}} = 0{,}92$$

Yksisuuntaisessa testissä

p-arvo $= P(Z > 0{,}92) = 1 - \Phi(0{,}92) = 0{,}1788$

poikia ja tyttöjä syntyy yhtä paljon

H_0: $\mu_1 = \mu_2$

Aineisto Plasma.sav sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

- Onko keskimääräisissä kolesterolimäärissä eroja miehillä ja naisilla?

  $t = 3{,}750$, p-arvo $< 0{,}001$, on eroja

Lopuksi

- Osaamistavoitteet

  http://www.sis.uta.fi/tilasto/mtttp5/syksy2018/luento_1.pdf

- Tentit

  ks. https://coursepages.uta.fi/mtttp5/syksy-2018/kurssi-info/

- Jatkoksi: Tilastomenetelmien perusteet MTTTA1

  https://www10.uta.fi/opas/opetusohjelma/marjapuuro.htm?id=38653