Tilastollisen päättelyn perusteet, MTTTP5

Luentorunko, lukuvuosi 2018 - 2019

Raija Leppälä

25. syyskuuta 2018

Sisältö

Luku 1

Johdanto

Tilastollinen analyysi voidaan jakaa karkeasti kuvailevaan (descriptive) analyysiin ja tilastolliseen päättelyyn (statistical inference). Kuvaileva tilastotiede pyrkii kuvailemaan tietoaineiston sisältöä erilaisten graafisten esitysten ja tunnuslukujen sekä taulukoiden avulla. Kuvailevaan tilastotieteeseen tutustuttiin tilastotieteen johdantokurssilla.

Opintojaksolla tilastollisen päättelyn perusteet perehdytään tilastolliseen päättelyyn, johon jo alustavasti tutustuttiinkin johdantokurssilla. Empiirisissä tutkimuksissa on käytössä satunnaisotos populaatiosta. Otoksen perusteella pyritään tekemään johtopäätelmiä koko populaatiosta. Voidaan haluta arvioida vaikkapa populaation keskiarvoa (Esim. 1.0.1) tai pyritään selvittämään milloin voidaan sanoa ehdollisten otoskeskiarvojen perusteella, että populaatioissa keskiarvot poikkeavat toisistaan (Esim. 1.0.2).

Esimerkki 1.0.1 Halutaan selvittää potilaiden sairaalassaolopäivien keskimääräistä aikaa. Tutkitaan asiaa tekemällä 100 potilaan satunnaisotos ja saadaan oheiset analyysitulokset.

Statistics	
Mean	4.530
Std. Deviation	3.678
Std. Error of Mean	0.3682
95% Confidence Interval of the Mean upper	5.135
lower	3.925
n	100

Test mean = value	
hypothesized value	5
actual estimate	4.530

t-Test	
test statistic	-1.28
prob $> \|t\|$	0.201

Esimerkki 1.0.2 Ovatko tytöt ja pojat syntyessään keskimäärin samanpainoisia? Eräästä aineistosta (http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.sav, http://mtl.uta.fi/tilasto/tiltp_aineistoja/saidit.xls, $n = 120$) laskettuna poikien painon keskiarvo oli 3640.46 ja tyttöjen 3451.27. Otoskeskiarvojen erotus oli siis 189.19. Voidaanko tämän perusteella yleistää ja sanoa, että pojat ovat syntyessään keskimäärin tyttöjä painavampia?

Analyysin tuloksia:

Means and Standard Deviations

level	number	mean	std dev
pojat	65	3640.46	438.24
tytöt	55	3451.27	523.28

t-Test	df	prob $> \|t\|$
2.156	118	0.033

Tilastollisten päätelmien teko perustuukin satunnaisotoksesta määriteltyjen tunnuslukujen (kuten esim. otoskeskiarvojen) todennäköisyysjakaumiin. Johtopäätelmät tehdään erilaisten tilastollisten testien ja analysointimenetelmien avulla. Tällaiseen päättelyyn sisältyy tiettyä epävarmuutta, jota pyritään hallitsemaan käyttäen hyväksi todennäköisyyslaskentaa ja erilaisia todennäköisyysjakaumia.

Opintojaksolla tilastollisen päättelyn perusteet tutustutaankin aluksi todennäköisyyslaskentaan, todennäköisyysjakaumiin sekä otosjakaumiin sekä niiden käyttöön tilastollisessa päättelyssä. Tämän jälkeen vuorossa on tilastollisen päättelyn peruskäsitteiden esittely. Käydään läpi estimointiin liittyviä käsitteitä, luottamusvälejä sekä tutustutaan joihinkin tilastollisiin testeihin.

Tämä moniste ei sisällä kovin laajaa kokoelmaa esimerkeistä, mutta lisäesimerkkejä löytyy opetuksen toteutuksen yhteydessä julkaistavasta luentomateriaalista. Opiskelun tukena voi myös käyttää opintojakson www-sivuston materiaalin, kirjallisuusluettelossa esitettyä oheiskirjallisuutta sekä tässä monisteessa olevia linkkejä.

Luku 2

Todennäköisyyslaskentaa

2.1 Satunnaisilmiö ja tapahtuma

Esimerkki 2.1.1 Heitettäessä rahaa ei tiedetä saadaanko kruuna vai klaava. Tiedetään, että molemmat vaihtoehdot ovat yhtä todennäköisiä. Heitettäessä noppaa tiedetään, että saadaan silmäluku 1, 2, 3, 4, 5 tai 6, mutta ei tiedetä etukäteen silmälukua. Tiedetään, että jokaisen silmäluvun todennäköisyys on sama. Kortin vetäminen sekoitetusta korttipakasta, lottoaminen, veikkaaminen, bussin saapuminen pysäkille ja päivän sää ovat myös esimerkkejä ilmiöistä, joihin liittyy epävarmuutta.

Satunnaisilmiö on mikä tahansa ilmiö, johon liittyy useita eri tulosmahdollisuuksia sekä epävarmuutta ilmiön tuloksesta. Puhutaan myös satunnaiskokeesta.

Satunnaisilmiöön liittyvien kaikkien mahdollisten tulosten joukkoa kutsutaan *perusjoukoksi* (otosavaruudeksi) E. Käytännössä ollaan kiinnostuneita joistain perusjoukon osajoukoista (sekä niiden esiintymistodennäköisyyksistä). Perusjoukon osajoukko on nimeltään *tapahtuma*. Tapahtumia merkitään A, B, C, ...

Esimerkki 2.1.2 Rahanheitto

$$\begin{aligned} & E = \text{"kaikki mahdolliset tulokset"} = \{\text{kruuna, klaava}\}. \\ \text{Tapahtumia:} \quad & A = \text{"saadaan kruuna"} = \{\text{kruuna}\}, \\ & B = \text{"saadaan klaava"} = \{\text{klaava}\}. \end{aligned}$$

Nopanheitto

$$\begin{aligned} & E = \{1, 2, 3, 4, 5, 6\}. \\ \text{Tapahtumia:} \quad & A = \text{"saadaan parillinen"} = \{2, 4, 6\}, \\ & B = \{1\}, \\ & C = \{1, 2, 3\}, \\ & D = \text{"saadaan suurempi kuin 4"} = \{5, 6\}. \end{aligned}$$

Kortin vetäminen sekoitetusta korttipakasta

$E =$ ”kaikki kortit”.

Tapahtumia: $A =$ ”saadaan pata”,

$B =$ ”saadaan kuningas”,

$C =$ ”saadaan punainen ässä”.

Lottoaminen (40 palloa, joista arvotaan palauttamatta 7)

$E =$ ”kaikki mahdolliset lottorivit”, joita on 18643560 (ks. kombinatoriikka).

Tapahtumia: $A =$ ”saadaan 7 oikein”,

$B =$ ”saadaan 6 oikein”,

$C =$ ”ei saada yhtään oikein”.

Veikkaaminen (13 kohdetta, joissa jokaisessa 3 vaihtoehtoa)

$E =$ ”kaikki mahdolliset rivit”, joita on 1594323 (ks. kombinatoriikka).

Tapahtumia: $A =$ ”saadaan 13 oikein”,

$B =$ ”saadaan 12 oikein”,

$C =$ ”ei saada yhtään oikein”.

2.2 Klassinen todennäköisyys

Olkoon tarkasteltavan satunnaisilmiön perusjoukossa n tulosta, jotka ovat *kaikki yhtä mahdollisia*. Olkoon tapahtumaan A liittyviä tuloksia k kappaletta ($0 \leq k \leq n$). Tällöin tapahtuman A todennäköisyys

$$P(A) = \frac{k}{n}.$$

Esimerkki 2.2.1 Rahanheitto

$$A = \text{”saadaan kruuna”}, \qquad P(A) = \frac{1}{2}.$$

Nopanheitto

$$A = \text{”saadaan parillinen”} = \{2, 4, 6\},$$
$$P(A) = \frac{3}{6},$$
$$B = \{1\},$$
$$P(B) = \frac{1}{6},$$
$$D = \text{”suurempi kuin 4”} = \{5, 6\},$$
$$P(D) = \frac{2}{6}.$$

Lottoaminen

$$A = \text{"saadaan 7 oikein"},$$
$$P(A) = \frac{1}{\text{kaikkien rivien lkm}} = \frac{1}{18643560},$$
$$B = \text{"saadaan 6 oikein"},$$
$$P(B) = \frac{\text{rivien lkm, joissa 6 oik.}}{\text{kaikkien rivien lkm.}}$$

Klassisen todennäköisyyden (voidaan liittää vain äärellisiin perusjoukkoihin) yhteydessä lukujen n ja k määrittäminen ei aina ole yksinkertaista. Joudutaan usein käyttämään hyväksi kombinatoriikkaa.

Tapahtuman A todennäköisyys voidaan myös määritellä arvoksi, jota tapahtuman suhteellinen frekvenssi lähestyy satunnaiskoetoistojen määrää kasvatettaessa.

2.3 Todennäköisyyslaskennan aksioomat ja laskusääntöjä

Matemaattisesti määriteltynä todennäköisyys on joukkofunktio P, joka liittää jokaiseen satunnaisilmiön tapahtumaan A reaaliluvun $P(A)$, jota sanotaan tapahtuman A todennäköisyydeksi ja joka toteuttaa tietyt aksioomat.

Aksiooma 1 Jos A on mikä tahansa satunnaisilmiön tapahtuma, niin

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

Aksiooma 2 $P(E) = 1$. Tällöin kyseessä varma tapahtuma.

Jos A ja B ovat kaksi saman satunnaisilmiön tapahtumaa, niin määritellään niiden yhdiste

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

ja leikkaus

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

Sanotaan, että tapahtumat A ja B ovat *erillisiä*, jos ne molemmat eivät voi tapahtua samanaikaisesti eli $A \cap B = \emptyset$ (mahdoton tapahtuma).

Aksiooma 3 Jos tapahtumat A ja B ovat erillisiä, eli $A \cap B = \emptyset$, niin $P(A \cup B) = P(A) + P(B)$.

Esimerkki 2.3.1 Nopanheitto

$$
\begin{aligned}
&A = \text{”saadaan parillinen”} = \{2, 4, 6\},\\
&P(A) = \frac{3}{6},\\
&B = \text{”saadaan ykkönen”} = \{1\},\\
&P(B) = \frac{1}{6},\\
&A \cup B = \text{”saadaan parillinen tai ykkönen”},\\
&A \cap B = \emptyset, \quad \text{joten} \quad P(A \cup B) = P(A) + P(B).
\end{aligned}
$$

Laskusääntö 1 Mahdottoman tapahtuman todennäköisyys on nolla:

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

Määritellään A:n komplementtitapahtuma

$$A^C = \text{”}A \text{ ei tapahdu”}$$

Laskusääntö 2

$$P(A^C) = 1 - P(A).$$

Esimerkki 2.3.2 Nopanheitto

$$
\begin{aligned}
&A = \text{”silmäluku pienempi kuin 6”},\\
&A^C = \text{”silmäluku 6”},\\
&P(A) = 1 - P(A^C) = 1 - \frac{1}{6}.
\end{aligned}
$$

Esimerkki 2.3.3 Tarkastellaan vakioveikkausrivin täyttämistä täysin satunnaisesti. Olkoon $A =$ ”saadaan korkeintaan 11 oikein”.

$$P(A) = 1 - P(A^C) = 1 - P(\text{saadaan 12 tai 13 oikein}).$$

Laskusääntö 3 Jos tapahtumat A_1, A_2, $\ldots$, A_k ovat pareittain erillisiä eli mitkään kaksi tapahtumaa eivät voi esiintyä samanaikaisesi, niin

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

Esimerkki 2.3.4 Vedetään kortti sekoitetusta pakasta. Laske todennäköisyys, että kortti on ruutu-, hertta- tai ristikortti. (Vast. $\frac{39}{52}$)

Laskusääntö 4 (yleinen yhteenlaskusääntö) Jos A ja B ovat saman satunnaisilmiön tapahtumia, niin

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

Esimerkki 2.3.5 Vedetään kortti sekoitetusta pakasta. Laske todennäköisyys, että kortti on patakortti tai ässä.

$$P(\text{kortti pata tai ässä}) = P(\text{kortti pata}) + P(\text{kortti ässä}) - P(\text{kortti pataässä})$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}.$$

Määritellään A:n ehdollinen todennäköisyys ehdolla B: Olkoot A ja B saman satunnaisilmiön tapahtumia siten, että $P(B) > 0$. Tällöin tapahtuman A ehdollinen todennäköisyys ehdolla, että tiedetään tapahtuman B esiintyneen on

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Esimerkki 2.3.6 Nopanheitossa on saatu pariton silmäluku. Mikä on silmäluvun 5 todennäköisyys? A = {5}, B = {1, 3, 5},

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}.$$

Laskusääntö 5 (yleinen kertolaskusääntö) Jos $P(B) > 0$, niin

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

Tapahtumat A ja B ovat (tilastollisesti, stokastisesti) *riippumattomia*, jos $P(A \mid B) = P(A)$. Tällöin siis B:n tapahtuminen tai tapahtumatta jääminen ei vaikuta A:n tapahtumisen todennäköisyyteen ja A:n tapahtuminen tai tapahtumatta jääminen ei vaikuta B:n tapahtumisen todennäköisyyteen.

Jos tapahtumat A ja B ovat riippumattomia, niin

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

Tapahtumien riippumattomuus voidaan yleistää: Tapahtumat A_1, A_2, ..., A_k ovat riippumattomia, jos minkään niistä tapahtuminen tai tapahtumatta jääminen ei vaikuta muiden tapahtumien todennäköisyyksiin. Tällöin

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1)\, P(A_2) \cdots P(A_k).$$

Riippumattomuuskäsite ja esitetty todennäköisyyden laskukaava voidaan yleistää myös eri satunnaisilmiöiden välille, jolloin tapahtumat voivat olla eri satunnaisilmiöistä. Puhutaan yhdistetystä satunnaisilmiöstä.

Esimerkki 2.3.7 Heitetään noppaa kaksi kertaa.

$$A = \text{”1. heiton silmäluku 5”}$$
$$B = \text{”2. heiton silmäluku 5”},$$

A ja B ovat riippumattomat, joten

$$P(\text{”saadaan 5 molemmilla heitoilla”})$$
$$= P(\text{”1. heiton silmäluku 5”})\, P(\text{”2. heiton silmäluku 5”}) = \left(\tfrac{1}{6}\right)\left(\tfrac{1}{6}\right)$$

Esimerkki 2.3.8 Heitetään noppaa kolme kertaa (toistetaan samaa satunnaisilmiötä).

$$A_1 = \text{"1. heiton silmäluku pariton"}$$
$$A_2 = \text{"2. heiton silmäluku pariton"}$$
$$A_3 = \text{"3. heiton silmäluku pariton"}$$

$$P(\text{"saadaan kaikilla heitoilla pariton"})$$
$$= P(\text{"1. heitolla pariton"})\, P(\text{"2. heitolla pariton"})\, P(\text{"3. heitolla pariton"}) = \frac{1}{8}$$

2.4 Kombinatoriikkaa

Tarkastellaan satunnaisilmiötä, jonka voidaan ajatella syntyvän K:ssa eri vaiheessa (yhdistetty satunnaisilmiö). Oletetaan, että i:nnessä vaiheessa on n_i eri tulosmahdollisuutta. Tällöin yhdistetyllä satunnaisilmiöllä on $n_1 n_2 \cdots n_K$ eri tulosta.

Esimerkki 2.4.1 Kuinka monta vakioveikkausriviä voidaan muodostaa? Montako sellaista, joissa ei yhtään oikeaa? (Vast. $3^{13} = 1594323$, $2^{13} = 8192$)

Esimerkki 2.4.2 Kuinka moneen erilaiseen jonoon henkilöt A, B ja C voidaan järjestää? (Vast. $3 \cdot 2 \cdot 1$)

Edellä muodostettiin kirjainten permutaatiot. Jonon mitä tahansa uutta järjestystä sanotaan *permutaatioksi*.

Kuinka moneen erilaiseen järjestykseen n erilaista alkiota voidaan asettaa? Erilaisia järjestyksiä (permutaatioita) on

$$n(n-1)(n-2)\cdots 2 \cdot 1 = n! \qquad (n\text{-kertoma}).$$

Määritellään $0! = 1$.

Kuinka moneen erilaiseen järjestykseen n:stä erilaisesta alkiosta valitut k alkiota voidaan järjestää?

Erilaisia järjestyksiä (permutaatioita) on

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

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

$$\frac{n!}{k!(n-k)!} = \binom{n}{k} \qquad (\text{lue: } n \text{ yli } k\text{:n})$$

kappaletta. Tämä luku on ns. binomikerroin. Kombinaatio on siis alkioiden joukko, jossa järjestyksellä ei ole väliä.

Esimerkki 2.4.3 Kuinka monta erilaista lottoriviä?

$$\binom{40}{7} = \frac{40!}{(40-7)!7!} = 18643560$$

Kuinka monta sellaista, jossa kaikki väärin?

$$\binom{33}{7} = \frac{33!}{(33-7)!7!} = 4272048$$

Kuinka monta sellaista, jossa k oikein?

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

Montako sellaista vakioveikkausriviä, jossa k oikein?

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

Esimerkki 2.4.4 Kuinka monta erilaista jonoa 5 henkilöä voi muodostaa? Entä 10 henkilöä? (Vast. $5! = 120$, $10! = 3628800$)

Esimerkki 2.4.5 Kuinka moneen eri järjestykseen korttipakan 52 korttia voi asettaa? (Vast. $52!$)

Esimerkki 2.4.6 Valitaan luvuista 1, 2, 3, 4, 5, 6 kaksi lukua satunnaisesti palauttamatta lukua valinnan jälkeen. Kyse siis yksinkertaisesta satunnaisotonnasta (YSO) palauttamatta. Muodosta kaikki mahdolliset otokset (populaation osajoukkoja, jossa järjestyksellä ei merkitystä) ja määritä otoksen suurin alkio sekä sen eri arvojen todennäköisyydet.

$$\binom{6}{2} = \frac{6!}{4!2!} = 15 \text{ otosta}$$

otokset		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

$$P(\text{Max} = 2) = \frac{1}{15}$$

$$P(\text{Max} = 3) = \frac{2}{15}$$

$$P(\text{Max} = 4) = \frac{3}{15}$$

$$P(\text{Max} = 5) = \frac{4}{15}$$

$$P(\text{Max} = 6) = \frac{5}{15}$$

Esimerkki 2.4.7 Kuten edellä, mutta otanta systemaattisella otannalla.

otokset		Max
1	4	4
2	5	5
3	6	6

$$P(\text{Max} = 4) = P(\text{Max} = 5) = P(\text{Max} = 6) = \frac{1}{3}$$

Luku 3

Todennäköisyysjakaumia

3.1 Satunnaismuuttuja ja todennäköisyysjakauma

Funktiota, joka liittää yksikäsitteisen reaaliluvun jokaiseen tarkasteltavan satunnaisilmiön perusjoukon tulokseen, sanotaan *satunnaismuuttujaksi*. Eri tuloksiin liittyviä reaalilukuja sanotaan *satunnaismuuttujan arvoksi*. Jatkossa merkitään (useimmiten) satunnaismuuttujia isoin kirjaimin (X, Y, Z, ...) ja satunnaismuuttujan arvoja pienin kirjaimin (x, y, z, ...).

Esimerkki 3.1.1 Satunnaisilmiö nopanheitto. Satunnaismuuttuja X = saatu silmäluku.

Esimerkki 3.1.2 Heitetään kolikkoa neljä kertaa. Määritellään satunnaismuuttuja X = klaavojen lukumäärä heittosarjassa. Etukäteen ei tiedetä montako klaavaa saadaan, mutta voidaan laskea eri arvojen todennäköisyydet. Tässä satunnaismuuttujan X mahdolliset arvot ovat 0, 1, 2, 3 ja 4. Erilaisia heittosarjoja on kaikkiaan 16.

heittosarja	klaavojen lkm	heittosarja	klaavojen lkm
Kl, Kl, Kl, Kl	4	Kr, Kl, Kl, Kr	2
Kr, Kl, Kl, Kl	3	Kl, Kr, Kl, Kr	2
Kl, Kr, Kl, Kl	3	Kr, Kl, Kr, Kl	2
Kl, Kl, Kr, Kl	3	Kl, Kr, Kr, Kr	1
Kl, Kl, Kl, Kr	3	Kr, Kl, Kr, Kr	1
Kl, Kl, Kr, Kr	2	Kr, Kr, Kl, Kr	1
Kr, Kr, Kl, Kl	2	Kr, Kr, Kr, Kl	1
Kl, Kr, Kr, Kl	2	Kr, Kr, Kr, Kr	0

$$P(X=0)=\frac{1}{16},\quad P(X=1)=\frac{4}{16},\quad P(X=2)=\frac{6}{16},$$
$$P(X=3)=\frac{4}{16},\quad P(X=4)=\frac{1}{16}.$$

Esimerkki 3.1.3 Satunnaisilmiönä veikkaaminen (13 kohdetta, joissa jokaisessa 3 vaihtoehtoa). Tällöin voidaan määritellä satunnaismuuttuja X = oikein veikat-

tujen kohteiden lukumäärä. X voi saada arvoja 0, 1, 2, ..., 13. Näiden arvojen todennäköisyydet voidaan laskea (ks. binomijakauma).

Esimerkissä 3.1.2 ilmoitettiin satunnaismuuttujan mahdolliset arvot ja eri arvojen todennäköisyydet. Tällöin muodostettiin satunnaismuuttujan *todennäköisyysjakauma*.

Satunnaismuuttuja voi olla joko *jatkuva* tai *diskreetti*. Edellisissä esimerkeissä satunnaismuuttujat olivat diskreettejä. Satunnaismuuttujaa sanotaan diskreetiksi, jos se voi saada arvokseen äärellisen määrän erisuuria arvoja tai äärettömän määrän siten, että arvot ovat numeroitavissa positiivisia kokonaislukuja käyttäen. Muulloin satunnaismuuttuja on jatkuva.

Diskreetin satunnaismuuttujan todennäköisyysjakauma voidaan usein (ainakin periaatteessa) muodostaa kuten esimerkissä 3.1.2 Jatkuvien muuttujien yhteydessä todennäköisyysjakauma määritellään jatkuvan funktion avulla. Funktiota, joka määrittää satunnaismuuttujan todennäköisyysjakauman kutsutaan *tiheysfunktioksi*, merk. $f(x)$. Diskreetin muuttujan yhteydessä puhutaan *pistetodennäköisyyksistä*.

Satunnaismuuttujan X *kertymäfunktio* F määritellään

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

Kertymäfunktion arvo pisteessä x kertoo siis todennäköisyyden sille, että satunnaismuuttujan X arvo on $\leq x$.

Kertymäfunktion ominaisuuksia:

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

Esimerkki 3.1.4 Heitetään kolikkoa neljä kertaa. Olkoon X = klaavojen lukumäärä heittosarjassa. Määritä ja piirrä X:n kertymäfunktio. Laske $P(X < 0)$, $P(X \leq 0)$, $P(X < 2.5)$ ja $P(X \leq 4)$.

$$\begin{aligned}
&F(x) = P(X \leq x),\\
&P(X < 0) = 0, \qquad P(X \leq 0) = \frac{1}{16},\\
&P(X < 2.5) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{11}{16},\\
&P(X \leq 4) = 1.
\end{aligned}$$

Kertymäfunktio on nyt porrasfunktio, joka voidaan piirtää todennäköisyyksien

$$P(X \leq 0) = \frac{1}{16}, \quad P(X \leq 1) = \frac{5}{16}, \quad P(X \leq 2) = \frac{11}{16},$$
$$P(X \leq 3) = \frac{15}{16}, \quad P(X \leq 4) = 1$$

avulla.

3.2 Diskreetti satunnaismuuttuja

Olkoon diskreetin satunnaismuuttujan X mahdolliset arvot x_1, x_2, ..., ja näiden arvojen todennäköisyydet p_1, p_2, ... Tällöin satunnaismuuttujan X todennäköisyysjakauma määritellään pistetodennäköisyyksien

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

perusteella.

Esimerkki 3.2.1 Heitetään noppaa. Määritellään X = saatu silmäluku.

Todennäköisyysjakauma:

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

Kertymäfunktio:

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < 1 \\ \frac{1}{6}, & 1 \leq x < 2 \\ \frac{2}{6}, & 2 \leq x < 3 \\ \vdots & \\ \frac{6}{6}, & x \geq 6 \end{cases}$$

Samalla tavalla kuin empiiristen jakaumien yhteydessä jakaumaa voitiin kuvailla tunnuslukujen avulla, voidaan myös teoreettisia todennäköisyysjakaumin kuvata samantyyppisillä tunnusluvuilla, jotka määritellään todennäköisyysjakauman avulla.

Empiirisen jakauman keskiarvoa vastaavaksi tunnusluvuksi todennäköisyysjakauman (populaation) yhteydessä määritellään jakauman odotusarvo (populaation keskiarvo) sekä otosvarianssia ja keskihajontaa vastaaviksi (populaation) varianssi ja keskihajonta.

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

Tällöin satunnaismuuttujan X *odotusarvo* $E(X)$ määritellään

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

sekä *varianssi* $\mathrm{Var}(X)$

$$\mathrm{Var}(X) = E[(X-\mu)^2] = \sum_{i=1}^{k} p_i(x_i - E(X))^2 = \sum_{i=1}^{k} p_i(x_i-\mu)^2 = \sigma^2$$

ja keskihajonta

$$\mathrm{Sd}(X) = \sqrt{\mathrm{Var}(X)} = \sigma$$

Huom. Edellä k voi siis olla myös ääretön.

Esimerkki 3.2.2 Heitetään noppaa. Määritellään $X =$ saatu silmäluku. Määritä $E(X)$ ja $\mathrm{Var}(X)$.

$$P(X=1) = \cdots = P(X=6) = \frac{1}{6}$$
$$E(X) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + \cdots + 6\cdot\frac{1}{6} = 3.5$$
$$\mathrm{Var}(X) = (1-3.5)^2\cdot\frac{1}{6} + (2-3.5)^2\cdot\frac{1}{6} + \cdots + (6-3.5)^2\cdot\frac{1}{6} = \frac{35}{12}$$

Esimerkki 3.2.3 Määritellään rahanheitossa $X = 1$, jos saadaan kruuna, 0 muulloin. Laske $E(X)$ ja $\mathrm{Var}(X)$.

$$P(X=1) = P(X=0) = \frac{1}{2}$$
$$E(X) = 1\cdot\frac{1}{2} + 0\cdot\frac{1}{2} = \frac{1}{2}$$
$$\mathrm{Var}(X) = \left(1-\frac{1}{2}\right)^2\cdot\frac{1}{2} + \left(0-\frac{1}{2}\right)^2\cdot\frac{1}{2} = \frac{1}{4}$$

3.3 Jatkuva satunnaismuuttuja

Olkoon jatkuvan satunnaismuuttujan X tiheysfunktio f. Jotta f olisi tiheysfunktio on $f(x) \geq 0$, jokaisella x:n arvolla sekä $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ eli $f(x)$:n ja x-akselin väliin jäävä pinta-ala = 1. Tiheysfunktio kuvaa siis ykkösen suuruisen todennäköisyysmassan jakaumaa. Tällöin X:n odotusarvo $E(X)$ määritellään

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

sekä varianssi $\mathrm{Var}(X)$

$$\mathrm{Var}(X) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} \big(x - E(X)\big)^2 f(x)\,\mathrm{d}x = \sigma^2$$

ja keskihajonta

$$\mathrm{Sd} = \sqrt{\mathrm{Var}(X)} = \sigma.$$

Odotusarvo kuvaa jakauman keskikohtaa ja varianssi mittaa miten tiiviisti todennäköisyysmassa on keskittynyt odotusarvon ympärille (vrt. empiiriset jakaumat).

Olkoon X jatkuva satunnaismuuttuja sekä a ja b reaalilukuja ($a \leq b$), tällöin

$$\begin{aligned}
&P(X \leq a) = P(X < a) = F(a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x,\\
&P(X \geq a) = P(X > a) = 1 - P(X \leq a) = 1 - F(a),\\
&P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)\\
&\qquad\qquad = F(b) - F(a).
\end{aligned}$$

Esimerkki 3.3.1 Olkoon X satunnaisesti väliltä $[0, 1]$ valittu reaaliluku. Tällöin $f(x)$=1. Määritä X:n tiheysfunktio sekä kertymäfunktio. Laske lisäksi $P(X > 0.25)$, $P(0.5 \leq X \leq 0.75)$, $P(X \leq a)$. Laske vielä $E(X)$ ja $\mathrm{Var}(X)$.

$$F(x) = P(X \leq x) = \begin{cases} x \cdot 1 = x, & 0 \leq x \leq 1\\ 0, & x < 0\\ 1, & x > 1 \end{cases}$$

$$\begin{aligned}
&P(X > 0.25) = 1 - P(X \leq 0.25) = 1 - F(0.25) = 1 - 0.25 = 0.75,\\
&P(0.5 \leq X \leq 0.75) = F(0.75) - F(0.5) = 0.75 - 0.5 = 0.25,\\
&E(X) = \int_0^1 1 \cdot x\,\mathrm{d}x = \frac{1}{2} - 0 = \frac{1}{2},\\
&\mathrm{Var}(X) = \int_0^1 f(x)(x - \frac{1}{2})^2\,\mathrm{d}x\\
&\qquad = \int_0^1 1 \cdot (x - \frac{1}{2})^2\,\mathrm{d}x = \int_0^1 (x^2 - x + \frac{1}{4})\,\mathrm{d}x = \frac{1}{12}.
\end{aligned}$$

Olkoon $E(X) = \mu$ ja $\mathrm{Var}(X) = \sigma^2$. Tällöin muuttuja X standardoidaan tekemällä muunnos

$$Z = \frac{(X - \mu)}{\sigma}.$$

3.4 Odotusarvon ja varianssin ominaisuuksia

Odotusarvon ominaisuuksia:

1. $E(a) = a$, a vakio
2. $E(aX + b) = aE(X) + b$, X satunnaismuuttuja ja a, b vakioita ($aX + b$ myös satunnaismuuttuja)

3. Olkoot $X_1, X_2, \ldots, X_n$ satunnaismuuttujia, jolloin myös $X_1+X_2+\cdots+X_n$ on satunnaismuuttuja ja

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

4. Jos satunnaismuuttujat X ja Y ovat riippumattomia, niin

$$E(XY) = E(X)\,E(Y).$$

Satunnaismuuttujien riippumattomuus määritellään vastaavalla tavalla kuin tapahtumien riippumattomuuskin. Diskreetin satunnaismuuttujan yhteydessä: Satunnaismuuttujat ovat riippumattomia, joss

$$P(X = x_i,\, Y = y_j) = P(X = x_i)(Y = y_j), \quad \forall i, j$$

Varianssin ominaisuuksia:

1. $\operatorname{Var}(a) = 0$, a vakio
2. $\operatorname{Var}(X) = E(X^2) - (E(X))^2$
3. $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$, a, b vakioita
4. $\operatorname{Sd}(aX + b) = |a| \operatorname{Sd}(X)$, a, b vakioita
5. Jos satunnaismuuttujat X_1, X_2, …, X_n ovat riippumattomia, niin

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_n) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2) + \cdots + \operatorname{Var}(X_n)$$

6. Olkoot X ja Y satunnaismuuttujia. Tällöin

$$\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) \pm 2 \operatorname{Cov}(X, Y),$$

missä $\operatorname{Cov}(X, Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big] = \sigma_{XY}$ on satunnaismuuttujien X ja Y välinen *kovarianssi*, joka on nolla, jos X ja Y ovat riippumattomia. Kovarianssi liittyy muuttujien X ja Y yhteisjakaumaan. Satunnaismuuttujien X ja Y välinen korrelaatiokerroin

$$\rho_{XY} = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Sd}(X) \operatorname{Sd}(Y)}.$$

Esimerkki 3.4.1 Olkoon $E(X) = \mu$ ja $\operatorname{Var}(X) = \sigma^2$. Määritellään $Z = (X - \mu)/\sigma$. Laske $E(Z)$ ja $\operatorname{Var}(Z)$.

$$E(Z) = E\left(\frac{(X - \mu)}{\sigma}\right) = \frac{1}{\sigma}\big(E(X) - \mu\big) = \frac{1}{\sigma}(\mu - \mu) = 0,$$

$$\operatorname{Var}(Z) = \operatorname{Var}\left(\frac{(X - \mu)}{\sigma}\right) = \frac{1}{\sigma^2} \operatorname{Var}(X - \mu) = \frac{1}{\sigma^2} \operatorname{Var}(X) = 1.$$

Esimerkki 3.4.2 Olkoot X ja Y riippumattomia satunnaismuuttujia sekä määritellään $Z = X - Y$. Olkoon $\operatorname{Sd}(X) = \sigma_X$ ja $\operatorname{Sd}(Y) = \sigma_Y$ sekä $E(X) = \mu_X$ ja $E(Y) = \mu_Y$. Laske Z:n odotusarvo ja keskihajonta.

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

$$\begin{aligned}\operatorname{Var}(Z) &= \operatorname{Var}(X - Y) = \operatorname{Var}(X) + \operatorname{Var}(-Y) \\ &= \operatorname{Var}(X) + (-1)^2 \operatorname{Var}(Y) = \sigma_X^2 + \sigma_Y^2,\end{aligned}$$

$$\operatorname{Sd}(Z) = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

Esimerkki 3.4.3 Olkoot X_1, X_2, …, X_n riippumattomia satunnaismuuttujia siten, että $E(X_i) = \mu$ ja $\mathrm{Var}(X_i) = \sigma^2$. Määritellään $Y = (X_1 + X_2 + \cdots + X_n)/n$. Laske $E(Y)$ ja $\mathrm{Var}(Y)$.

$$\begin{aligned} E(Y) &= E\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n}\big(E(X_1) + E(X_2) + \cdots + E(X_n)\big) = \frac{1}{n} \cdot n \cdot \mu = \mu, \\ \mathrm{Var}(Y) &= \mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \left(\frac{1}{n}\right)^2 \big(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\big) \\ &= \left(\frac{1}{n}\right)^2 \cdot\ n \cdot \sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Esimerkki 3.4.4 Sijoitat 1000 euroa. Mahdollisia sijoituskohteita A ja B, joissa molemmissa pienin sijoitusmäärä 500 euroa. Olkoon X = tuotto 100 euron sijoituksesta A:han, Y = tuotto 100 euron sijoituksesta B:hen. Olkoon lisäksi $P(X = -5) = 0.4$, $P(X = 20) = 0.6$, $P(Y = 0) = 0.6$, $P(Y = 25) = 0.4$ sekä sijoitukset toisistaan riippumattomia. Miten sijoittaisit?

Mahdolliset vaihtoehdot:

1. 1000 euroa A:han,
2. 1000 euroa B:hen,
3. 500 euroa kumpaankin.

$$\begin{aligned} &E(X) = -5 \cdot 0.4 + 20 \cdot 0.6 = 10, \\ &E(Y) = 0 \cdot 0.6 + 25 \cdot 0.4 = 10, \\ &\mathrm{Var}(X) = 0.4(-5 - 10)^2 + 0.6(20 - 10)^2 = 150, \\ &\mathrm{Var}(Y) = 0.6(0 - 10)^2 + 0.4(25 - 10)^2 = 150. \end{aligned}$$

Olkoon W tuotto sijoituksesta.

1. $W = 10X$: $E(10X) = 10 \cdot E(X) = 100$, $\mathrm{Var}(10X) = 10^2\,\mathrm{Var}(X) = 15000$.
2. $W = 10Y$: $E(10Y) = 10 \cdot E(Y) = 100$, $\mathrm{Var}(10Y) = 10^2\,\mathrm{Var}(Y) = 15000$.
3. $W = 5X + 5Y$: $E(W) = E(5X) + E(5Y) = 5E(X) + 5E(Y) = 100$, $\mathrm{Var}(W) = 5^2\,\mathrm{Var}(X) + 5^2\,\mathrm{Var}(Y) = 7500$.

Vaihtoehto 3 on paras.

Esimerkki 3.4.5 Sijoitetaan 1000 euroa. Mahdollisia kohteita A ja B. Olkoon $X = 1$ euron tuotto kohteesta A, $Y = 1$ euron tuotto kohteesta B. Olkoon X ja Y riippumattomia sekä $E(X) = E(Y) = \mu$ ja $\mathrm{Var}(X) = \mathrm{Var}(Y) = \sigma^2$. Miten sijoitat?

Sijoitetaan kohteeseen A α euroa ja kohteeseen B $(1000-\alpha)$ euroa. Tuotto $W = \alpha \cdot X + (1000-\alpha)Y$.

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

siis ei riipu α:sta!

$$\begin{aligned}\operatorname{Var}(W) &= \operatorname{Var}(\alpha X) + \operatorname{Var}\big((1000-\alpha)Y\big) = \alpha^2 \operatorname{Var}(X) + (1000-\alpha)^2 \operatorname{Var}(Y) \\ &= \sigma^2(2\alpha^2 - 2000\alpha + 1000000).\end{aligned}$$

Jos $\alpha = 0$, niin $\operatorname{Var}(W) = 1000000\sigma^2$. Jos $\alpha = 1000$, niin $\operatorname{Var}(W) = 1000000\sigma^2$.

Minimoidaan $f(\alpha) = 2\alpha^2 - 2000\alpha + 1000000$.

$$f'(\alpha) = 4\alpha - 2000 = 0 \iff \alpha = 500$$

Tällöin $\operatorname{Var}(W) = 500000\sigma^2$.

Kannattaa sijoittaa 500 euroa molempiin, koska tällöin tuotolla on pienin varianssi.

3.5 Joitain todennäköisyysjakaumia

3.5.1 Bernoulli-jakauma

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

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

Olkoon lisäksi

$$\begin{aligned} P(A) &= P(X=1) = p, \\ P(A^C) &= P(X=0) = q = 1-p. \end{aligned}$$

Tällöin sanotaan, että X noudattaa Bernoulli-jakaumaa parametrilla p. Merkitään $X \sim \operatorname{Ber}(p)$.

Jos $X \sim \operatorname{Ber}(p)$, niin

$$E(X) = p \qquad \text{ja} \qquad \operatorname{Var}(X) = p(1-p) = pq.$$

Esimerkki 3.5.1

- Rahanheitto
- Veikkauksessa yhden kohteen arvaaminen
- Nopanheitto onnistumisena silmäluvun 6 saaminen

3.5.2 Binomijakauma

Tarkastellaan vakioveikkausta. Määritellään satunnaismuuttuja X = oikein arvattujen kohteiden kokonaislukumäärä. Tehtävänä on määrittää X:n todennäköisyysjakauma. Tällöin päädytään nk. binomijakaumaan.

Olkoon satunnaisilmiössä onnistumisen todennäköisyys p. Toistetaan tätä satunnaisilmiötä n kertaa. Määritellään X = onnistumisten kokonaislukumäärä. Tällöin sanotaan, että X noudattaa binomijakaumaa parametrein n ja p. Merkitään $X \sim \text{Bin}(n, p)$. Jos $X \sim \text{Bin}(n, p)$, niin

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

ja

$$E(X) = np \qquad \text{sekä} \qquad \text{Var}(X) = np(1-p) = npq.$$

Binomijakaumaa noudattava satunnaismuuttuja määritellään siis itse asiassa Bernoulli-jakaumaa noudattavien satunnaismuuttujien summana. Olkoon $X_i \sim \text{Ber}(p)$, jolloin toistettaessa Bernoulli-koetta n kertaa, onnistumisten kokonaislukumäärä voidaan määritellä

$$X = X_1 + X_2 + \cdots + X_n$$

ja tällöin siis $X \sim \text{Bin}(n, p)$.

Tämän summamuuttujan avulla saadaan laskettua binomijakauman odotusarvo ja varianssi.

Esimerkki 3.5.2 Veikataan satunnaisesti yksi rivi. Määritellään X = oikein arvattujen kohteiden kokonaislukumäärä. Määritä X:n jakauma sekä sen odotusarvo. Laske $P(X = 0)$, $P(X = 13)$, $P(X > 11)$, $P(X > 3)$.

$$\begin{aligned}
&X \sim \text{Bin}\left(13, \frac{1}{3}\right), \\
&P(X = k) = \binom{13}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{13-k}, \\
&P(X = 0) = \binom{13}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{13-0} = \left(\frac{2}{3}\right)^{13}, \\
&P(X = 13) = \binom{13}{13} \left(\frac{1}{3}\right)^{13} \left(\frac{2}{3}\right)^{13-13} = \left(\frac{1}{3}\right)^{13}, \\
&P(X = 12) = \binom{13}{12} \left(\frac{1}{3}\right)^{12} \left(\frac{2}{3}\right)^{13-12} = \cdots \approx 0.000016, \\
&P(X = 11) = \binom{13}{11} \left(\frac{1}{3}\right)^{11} \left(\frac{2}{3}\right)^{13-11} = \cdots \approx 0.000196, \\
&P(X > 11) = P(X = 12) + P(X = 13).
\end{aligned}$$

Esimerkki 3.5.3 Pelaat ystäväsi kanssa peliä, jossa heitetään rahaa. Jos tulee klaava, saat ystävältäsi euron, jos tulee kruuna, annat ystävällesi euron. On heitetty rahaa 20 kertaa ja olet tappiolla 14 euroa eli on tullut 17 kruunaa ja 3 klaavaa. Onko syytä tutkia rahaa tarkemmin? Jos raha harhaton, niin $X =$ klaavojen lukumäärä 20 heitossa $\sim \mathrm{Bin}\big(20, \frac{1}{2}\big)$. Millä todennäköisyydellä olet vähintään 14 euroa tappiolla?

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

$$\begin{aligned} P(X \le 3) &= P(X = 0 \text{ tai } X = 1 \text{ tai } X = 2 \text{ tai } X = 3) \\ &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= \left[\binom{20}{0} + \binom{20}{1} + \binom{20}{2} + \binom{20}{3}\right]\left(\frac{1}{2}\right)^{20} \\ &= \left[\frac{20!}{0!\,20!} + \frac{20!}{1!\,19!} + \frac{20!}{2!\,18!} + \frac{20!}{3!\,17!}\right]\left(\frac{1}{2}\right)^{20} \\ &= (1 + 20 + 190 + 1140)\left(\frac{1}{2}\right)^{20} = 1351 \cdot \left(\frac{1}{2}\right)^{20}. \end{aligned}$$

On siis sattunut tapahtuma, jonka todennäköisyys on hiukan yli 1/1000 tai pelissä oleva raha on harhainen ja antaa kruunan useammin kuin klaavan.

3.5.3 Diskreetti tasajakauma

Noppaa heitettäessä voidaan määritellä satunnaismuuttuja $X =$ silmäluku. X:n mahdolliset arvot ovat 1, 2, 3, 4, 5, 6 ja jokaisen esiintymistodennäköisyys 1/6. Tätä jakaumaa kutsutaan diskreetiksi tasajakaumaksi välillä $(1, 6)$.

Jos satunnaismuuttujan X arvot ovat kokonaislukuja

$$a,\ a+1,\ a+2,\ a+3, \ldots,\ a+(n-1) = b$$

ja kukin n:stä arvo yhtä todennäköinen, niin sanotaan, että X noudattaa diskreettiä tasajakaumaa välillä (a, b). Merkitään $X \sim \mathrm{Tasd}(a, b)$. Tällöin

$$E(X) = \frac{a+b}{2} \qquad \text{ja} \qquad \mathrm{Var}(X) = \frac{n^2 - 1}{12}.$$

Esimerkki 3.5.4 Nopanheitto.

$$X \sim \mathrm{Tasd}(1, 6), \qquad E(X) = \frac{1+6}{2} = 3.5, \qquad \mathrm{Var}(X) = \frac{6^2 - 1}{12} = \frac{35}{12}.$$

Esimerkki 3.5.5 Olkoon X yksinumeroinen satunnaisluku. Mahdolliset arvot ovat siis 0, 1, 2, ..., 9 ja jokaisen arvon todennäköisyys 1/10. Tällöin $X \sim \mathrm{Tasd}(0, 9)$, $E(X) = (0 + 9)/2$ ja $\mathrm{Var}(X) = (10^2 - 1)/12$.

3.5.4 Jatkuva tasajakauma

Satunnaismuuttuja noudattaa jatkuvaa tasajakaumaa välillä $[a, b]$, jos sen tiheysfunktio f on

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

Merkitään $X \sim \text{Tas}(a, b)$. Tällöin

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

Esimerkki 3.5.6 Aiemmat esimerkit

$$X \sim \text{Tas}(0, 1), \qquad E(X) = \frac{1+0}{2} = 0.5, \qquad \text{Var}(X) = \frac{(1-0)^2}{12} = \frac{1}{12}.$$

3.5.5 Normaalijakauma

Seuraava todennäköisyysjakauma on tilastotieteessä hyvin keskeinen. Tarkastellaan jatkuvaa satunnaismuuttujaa X, joka voi saada arvokseen kaikki reaaliluvut. Satunnaismuuttuja X noudattaa normaalijakaumaa parametrein μ ja σ^2 ($\sigma > 0$), jos sen tiheysfunktio on

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}[(x-\mu)/\sigma]^2}, \qquad -\infty < x < \infty.$$

Tällöin $E(X) = \mu$ ja $\text{Var}(X) = \sigma^2$. Merkitään $X \sim N(\mu, \sigma^2)$.

Jos $X \sim N(\mu, \sigma^2)$, niin sen tiheysfunktio on yksihuippuinen jakauma, symmetrinen odotusarvon suhteen varianssin kertoessa jakauman levittäytymisestä odotusarvon ympärille.

Jos $X \sim N(0, 1)$, niin sen tiheysfunktio on

$$f(x) = \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

Kyseessä nk. *standardoitu normaalijakauma*. Usein merk. $Z \sim N(0, 1)$, $f(z) = \phi(z)$ ja $F(z) = P(Z \leq z) = \Phi(z)$.

Normaalijakauman tiheysfunktion integraaliafunktiota (kertymäfunktiota) ei tunneta. Standardoidun normaalijakauman kertymäfunktion $\Phi(z) = P(Z \leq z)$ arvoja on taulukoitu. Taulukoiden avulla voidaan laske erilaisia todennäköisyyksiä. Normaalijakauman symmetrisyydestä seuraa, että $\Phi(z) = 1 - \Phi(-z)$.

Esimerkki 3.5.7 Olkoon $Z \sim N(0,1)$. Laske $P(Z \leq 1)$, $P(Z \leq 1.1)$, $P(Z \leq 1.14)$, $P(Z \leq$ -1), $P(Z \leq 0)$, $P(-1 \leq Z \leq 1)$, $P(-2 \leq Z \leq 2)$, $P(-3 \leq Z \leq 3)$.

$$
\begin{aligned}
&P(Z \leq 1) = \Phi(1) = 0.8413,\\
&P(Z \leq 1.1) = \Phi(1.1) = 0.8643,\\
&P(Z \leq 1.14) = \Phi(1.14) = 0.8729,\\
&P(Z \leq -1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587,\\
&P(Z \leq 0) = 0.5,\\
&P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) = \Phi(1) - \big(1 - \Phi(1)\big)\\
&\qquad\qquad = \Phi(1) - 1 + \Phi(1) = 2\Phi(1) - 1 = 0.6826,\\
&P(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) = \Phi(2) - \big(1 - \Phi(2)\big)\\
&\qquad\qquad = 2\Phi(2) - 1 = 0.9544,\\
&P(-3 \leq Z \leq 3) = \Phi(3) - \Phi(-3) = \Phi(3) - \big(1 - \Phi(3)\big)\\
&\qquad\qquad = 2\Phi(3) - 1 = 0.9974.
\end{aligned}
$$

Esimerkki 3.5.8 Olkoon $Z \sim N(0,1)$. Määritä z, kun a) $\Phi(z) = 0.75$ b) $\Phi(z) = 0.26$.

a) $P(Z \leq z) = 0.75 \Longrightarrow z \approx 0.67,$

b) $P(Z \leq z) = 0.26 \Longleftrightarrow P(Z \leq -z) = 1 - 0.26 = 0.74$
$\Longrightarrow -z = 0.64 \Longrightarrow z = -0.64.$

Jos $X \sim N(\mu, \sigma^2)$, niin $P(X \leq a)$ voidaan laskea käyttäen standardoitua normaalijakaumaa, sillä on osoitettavissa, että jos $X \sim N(\mu, \sigma^2)$, niin

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$

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

$$P(X \leq a) = P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right),$$

$$P(X \geq a) = 1 - P(X \leq a) = 1 - P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

ja

$$
\begin{aligned}
P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a)\\
&= P\left(\frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) - P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right)\\
&= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).
\end{aligned}
$$

Esimerkki 3.5.9 Sinulla on sijoitusvaihtoehdot A ja B. Oletat, että sijoitusten tuottoprosentit noudattavat normaalijakaumia odotusarvoina 10.4 ja 11.0 sekä hajontoina 1.2 ja 4.0. Haluat tehdä sijoituksen, jolla on todennäköisempää saada vähintään 10 prosentin tuotto. Kumman sijoitusvaihtoehdon valitset?

X = tuotto sijoituksesta A, $X \sim N(10.4, 1.2^2)$

Y = tuotto sijoituksesta B, $Y \sim N(11.0, 4.0^2)$

$P(X \geq 10) = 1 - P(X \leq 10) = 1 - \Phi(\frac{10-10.4}{1.2}) = 1 - \Phi(-0.33) = 1 - (1 - \Phi(0.33)) = 0.6293.$

$P(Y \geq 10) = 1 - P(Y \leq 10) = 1 - \Phi(\frac{10-11}{4}) = 1 - \Phi(-0.25) = 1 - (1 - \Phi(0.25)) = 0.5987.$

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

Esimerkki 3.5.10 Laske todennäköisyydet, että normaalijakaumassa satunnaismuuttujan arvo on korkeitaan

a) hajonnan päässä odotusarvosta,
b) kahden hajonnan päässä odotusarvosta,
c) kolmen hajonnan päässä odotusarvosta.

$X \sim N(\mu, \sigma^2)$.

a)
$$P(-\sigma \leq X - \mu \leq \sigma) = P\left(-\frac{\sigma}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{\sigma}{\sigma}\right) = \Phi(1) - \Phi(-1) = \cdots = 0.6826$$

b)
$$P(-2\sigma \leq X - \mu \leq 2\sigma) = P\left(-\frac{2\sigma}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{2\sigma}{\sigma}\right) = \Phi(2) - \Phi(-2) = \cdots = 0.9544$$

c)
$$P(-3\sigma \leq X - \mu \leq 3\sigma) = P\left(-\frac{3\sigma}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{3\sigma}{\sigma}\right) = \Phi(3) - \Phi(-3) = \cdots = 0.9974$$

Normaalijakaumaan liittyviä keskeisiä teoreettisia tuloksia:

1. Jos $X \sim N(\mu, \sigma^2)$, niin $aX + b \sim N(a\mu + b, a^2\sigma^2)$, ($a$, b vakioita).
2. Jos X_1, X_2, …, X_n ovat riippumattomia ja $X_i \sim N(\mu_i, \sigma_i^2)$, niin

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

3. *Keskeinen raja-arvolause*: Olkoon X_1, X_2, ..., X_n riippumattomia satunnaismuuttujia, joista kukin noudattaa omaa jakaumaansa. Olkoon $E(X_i) = \mu_i$ ja $\mathrm{Var}(X_i) = \sigma_i^2$, $i = 1, 2, \ldots, n$. Tällöin (hyvin yleisten ehtojen vallitessa) satunnaismuuttuja $X_1 + X_2 + \cdots + X_n$ noudattaa likimain normaalijakaumaa (kun n riittävän iso) parametrein $\mu_1 + \mu_2 + \cdots + \mu_n$ ja $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$.

Esimerkki 3.5.11 Olkoot X_1, X_2, X_3 ja X_4 riippumattomia ja kukin $X_i \sim N(0,1)$. Määritellään $\overline{X} = (X_1 + X_2 + X_3 + X_4)/4$. Laske $P(\overline{X} \geq 1)$.

$$\begin{aligned}
E(\overline{X}) &= \frac{1}{4}E(X_1 + X_2 + X_3 + X_4) \\
&= \frac{1}{4}\big(E(X_1) + E(X_2) + E(X_3) + E(X_4)\big) = 0, \\
\mathrm{Var}(\overline{X}) &= \left(\frac{1}{4}\right)^2 \mathrm{Var}(X_1 + X_2 + X_3 + X_4) \\
&= \left(\frac{1}{4}\right)^2 \big(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \mathrm{Var}(X_3) + \mathrm{Var}(X_4)\big) \\
&= \left(\frac{1}{4}\right)^2 \cdot 4 \cdot 1 = \frac{1}{4}, \\
\overline{X} &\sim N\left(0, \frac{1}{4}\right), \\
P(\overline{X} \geq 1) &= 1 - P(\overline{X} \leq 1) \\
&= 1 - P\left(\frac{\overline{X} - 0}{1/2} \leq \frac{1-0}{1/2}\right) = 1 - \Phi(2) = 0.0228.
\end{aligned}$$

Olkoot X_1, X_2, ..., X_n riippumattomia ja kukin $X_i \sim N(\mu, \sigma^2)$, niin tällöin

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Otoskeskiarvon jakauma on siis normaalijakauma (ks. otosjakaumat)! Vaikka X_i:t eivät olisikaan normaalisti jakautuneita, niin $\overline{X}$ olisi likimain normaalisti jakautunut keskeisen raja-arvolauseen perusteella.

Binomijakaumaa voidaan approksimoida normaalijakaumalla. Jos $X \sim \mathrm{Bin}(n,p)$, niin silloinhan $X = X_1 + X_2 + \cdots + X_n$, missä $X_i \sim \mathrm{Ber}(p)$. Keskeisen raja-arvolauseen mukaan (jos n on riittävän suuri) X noudattaa likimain normaalijakaumaa parametrein np ja npq. Approksimaatio on hyvä, jos n on suuri ja p ei ole kovin pieni eikä suuri.

Esimerkki 3.5.12 Henkilö osallistuu tenttiin, jossa sataan väitteeseen vastataan väitteen olevan tosi tai epätosi ja vain toinen vaihtoehto on oikea. Jos henkilö vastaa kaikkiin kohtiin valitsemalla vaihtoehdon aina täysin satunnaisesti, niin millä todennäköisyydellä hän saa korkeintaan 60 oikeaa vastausta?

X = oikeiden vastausten lkm.

$$X \sim \text{Bin}\left(100, \frac{1}{2}\right), \qquad E(X) = 100 \cdot \frac{1}{2} = 50, \qquad \text{Var}(X) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25,$$

$$P(X \leq 60) = \sum_{k=0}^{60} \binom{100}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{100-k} = \sum_{k=0}^{60} \binom{100}{k} \left(\frac{1}{2}\right)^{100} \approx 0.9824$$

(laskettu Excel:illä),

$$X \overset{\text{likimain}}{\sim} N(50, 25), \quad \text{jolloin} \quad P(X \leq 60) \approx \Phi\left(\frac{60 - 50}{\sqrt{25}}\right) = \Phi(2) = 0.9772.$$

Kun binomijakaumaa approksimoidaan normaalijakaumalla, niin diskreettiä jakaumaa arvioidaan jatkuvalla. Paremman arvion saamiseksi voidaan tehdä nk. jatkuvuuskorjaus. Arvioitaessa $P(X \leq a)$, missä a on kokonaisluku, lasketaankin $P(X \leq a + 0.5)$. Tässä esimerkissä

$$P(X \leq 60) \approx \Phi\left(\frac{60.5 - 50}{\sqrt{25}}\right) = \Phi(2.1) = 0.9821.$$

Esimerkki 3.5.13 Levykaupan omistaja arvioi, että 20 % asiakkaista suorittaa ostoksen. Laske todennäköisyys, että 180 asiakkaan joukosta ainakin 45 suorittaa ostoksen (binomijakaumaa voidaan approksimoida normaalijakaumalla).

X = ostosten suorittajien lkm.

$$X \sim \text{Bin}(180, 0.2), \quad E(X) = 180 \cdot 0.2 = 36, \quad \text{Var}(X) = 180 \cdot 0.2 \cdot 0.8 = 28.8,$$

$$P(X \geq 45) = 1 - P(X \leq 44) = 1 - \sum_{k=0}^{44} \binom{180}{k} 0.2^k \cdot 0.8^{180-k}$$

$$= 1 - 0.94054 = 0.059458 \qquad \text{(laskettu Excel:illä)}$$

Nyt $X \overset{\text{likimain}}{\sim} N(36, 28.8)$, jolloin

$$P(X \geq 45) = 1 - P(X \leq 44) \approx 1 - \Phi\left(\frac{44.5 - 36}{\sqrt{28.8}}\right) = 1 - \Phi(1.58) = 0.0571.$$

Ilman jatkuvuuskorjausta:

$$P(X \geq 45) = 1 - P(X \leq 44) \approx 1 - \Phi\left(\frac{44 - 36}{\sqrt{28.8}}\right) = 0.0681.$$

Esimerkki 3.5.14 GMAT-testiä käytetään useiden yliopistojen pääsykokeena. Kokeen tuloksen on todettu noudattavan normaalijakaumaa odotusarvona 525 ja keskihajontana 100. Sadan yliopistoon pyrkijän ryhmä osallistui ennen pääsykoetta valmennuskurssille. Pääsykokeessa heidän GMAT-testin keskiarvo oli 541.4. Menestyivätkö he pääsykokeessa muita paremmin?

X = testipistemäärä, $X \sim N(525, 100^2)$

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

$P(\overline{X} \geq 541.4) = 1 - P(\overline{X} \leq 541.4) = 1 - \Phi(\frac{541.5-525}{10}) = 1 - \Phi(1,64) = 0.0505.$

Eivät menestyneet paremmin kuin muut, koska ei ole harvinaista saada otoskeskiarvoa, joka suurempi kuin 541.4 silloin, kun menestyminen tavanomaista.

Esimerkki 3.5.15 Oletetaan, että opiskelijoiden älykkyysosamäärä $\sim N(\mu, 225)$. Kuinka suuri otos tarvitaan, kun halutaan, että otoskeskiarvo poikkeaa μ:stä korkeintaan ± 2 pistettä todennäköisyydellä 0.99?

$$X \sim N(\mu, 225), \qquad \overline{X} \sim N\left(\mu, \frac{225}{n}\right),$$

$$\begin{aligned} P(-2 \leq \overline{X} - \mu \leq 2) &= P\left(-\frac{2}{15/\sqrt{n}} \leq \frac{\overline{X} - \mu}{15/\sqrt{n}} \leq \frac{2}{15/\sqrt{n}}\right) \\ &= \Phi\left(\frac{2\sqrt{n}}{15}\right) - \Phi\left(\frac{-2\sqrt{n}}{15}\right) \\ &= \Phi\left(\frac{2\sqrt{n}}{15}\right) - \left[1 - \Phi\left(\frac{2\sqrt{n}}{15}\right)\right] \\ &= 2\Phi\left(\frac{2\sqrt{n}}{15}\right) - 1 = 0.99 \end{aligned}$$

$$\iff \Phi\left(\frac{2\sqrt{n}}{15}\right) = \frac{1.99}{2} = 0.995 \iff \frac{2\sqrt{n}}{15} = 2.58 \iff n = \frac{2.58^2 \cdot 15^2}{2^2} \approx 374.$$

Luku 4

Satunnaisotos, otossuure ja otosjakauma

Kun populaatio on hyvin suuri tai ääretön ei tietenkään voida tutkia koko populaatiota. Tällöin tilastolliset johtopäätelmät, jotka koskevat populaation eli perusjoukon (äärellinen tai ääretön) ominaisuuksia tehdään otoksen avulla. Jotta erilaisten otoksesta laskettujen tunnuslukujen luotettavuutta voidaan arvioida otos valitaan poimimalla se todennäköisyysotannalla. Todennäköisyysotannassa kaikki mahdolliset n alkion otokset voidaan luetella, tunnetaan jokaisen mahdollisen otoksen poimintatodennäköisyys ja otokset poimitaan näiden todennäköisyyksien mukaan sekä tiedetään, miten otoksen perusteella yleistetään tulokset koko populaatioon.

Jatkossa tarkastellaan yksinkertaisella satunnaisotannalla tehtyyn otokseen liittyviä tuloksia. Lisäksi ollaan kiinnostuneita vain yhdestä populaation alkioihin liittyvästä ominaisuudesta, muuttujasta.

Yksinkertainen satunnaisotos (YSO) poimitaan siten, että jokaisella n alkion suuruisella otoksella on yhtä suuri todennäköisyys tulla poimituksi. Käytännössä ei muodosteta kaikkia n alkion osajoukkoja, joista sitten satunnaisesti valitaan yksi, vaan alkiot poimitaan yksi kerrallaan kunnes otoskoko on n. YSO voidaan tehdä joko palauttamatta tai palauttaen.

4.1 Satunnaisotos

Olkoon $X_1, X_2, \ldots, X_n$ n:n satunnaismuuttujan jono. Tätä jonoa sanotaan *satunnaisotokseksi*, 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 sitä, että jokainen $X_i \sim N(\mu, \sigma^2)$ ja X_i:t ovat riippumattomia.

Kun äärettömästä populaatiosta tehdään otanta yksinkertaisella satunnaisotannalla (palauttaen tai palauttamatta) ja tarkastellaan yhtä tiettyä muuttujaa (tilastoyksikön ominaisuutta), on kyse satunnaisotoksesta. Jos populaatio on äärellinen YSO palauttaen johtaa satunnaisotokseen, mutta palauttamatta ei, koska riippumattomuusoletus ei ole voimassa. Kuitenkin, jos populaatio on suuri YSO palauttamattakin johtaa lähes riippumattomiin satunnaismuuttujiin.

Satunnaisotos määritellään siis satunnaismuuttujien perusteella. Nämä satunnaismuuttujat saavat arvot, kun otos on tehty. Siis otoksen tekemisen jälkeen satunnaisotokselle saadaan arvot, jotka vaihtelevat otoksesta toiseen.

Satunnaismuuttujista muodostetut funktiot ovat satunnaismuuttujia, joten myös satunnaisotoksesta muodostetut funktiot ovat satunnaismuuttujia.

Esimerkki 4.1.1 Otoskeskiarvo $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ on satunnaismuuttuja, joka saa arvon kun otos on tehty. Arvo vaihtelee otoksesta toiseen.

Esimerkki 4.1.2 Otosvarianssi $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ on on satunnaismuuttuja, joka saa arvon tehdyn otoksen perusteella.

4.2 Otossuureet ja otosjakaumat

Satunnaisotoksen avulla määriteltyä funktiota, joka siis on satunnaismuuttuja, kutsutaan *otossuureeksi*. Koska otossuure on satunnaismuuttuja, liittyy siihen todennäköisyysjakauma. Otossuureen todennäköisyysjakaumasta käytetään nimitystä *otanta-* tai *otosjakauma*.

Tarkasteltavan otossuureen todennäköisyysjakauma pyritään määrittämään, jolloin saadaan selville miten otossuure voi vaihdella otoksesta toiseen. Tämä auttaa, kun olemme kiinnostuneita populaatioon liittyvistä arvioista perustaen arviot otokseen.

Joidenkin otossuureiden otosjakaumia:

1. Otoskeskiarvon jakauma tunnetaan silloin kun, otos on normaalijakaumasta. Jos $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, niin tällöin

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

   Lisäksi voidaan keskeisen raja-arvolauseen perusteella sanoa, että (otoskoon ollessa riittävän suuri) otoskeskiarvo on likimain normaalisti jakautunut, vaikka satunnaisotos olisi peräisin jostain muusta kuin normaalijakaumasta.
   Otoskeskiarvon hajontaa $\sigma/\sqrt{n}$ sanotaan *otoskeskiarvon keskivirheeksi*.
2. Olkoon p viallisten prosenttiosuus otoksessa. Jos populaatiossa on π % viallisia, niin

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

   Otossuureen p hajontaa $\sqrt{\frac{\pi(100-\pi)}{n}}$ sanotaan p:n keskivirheeksi.

3. Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja Y_1, Y_2, …, Y_m satunnaisotos $N(\mu_2, \sigma_2^2)$:sta. Tällöin

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

Otoskeskiarvojen erotuksen keskivirhe on $\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$.

Esimerkki 4.2.1 Olkoon $X_1, X_2, \ldots, X_{10}$ satunnaisotos normaalijakaumasta parametrein 0 ja 1. Laske $P(-1 \leq X_1 \leq 1)$ ja $P(-1 \leq \overline{X} \leq 1)$.

$$X_i \sim N(0,1), \qquad i = 1, \ldots, 10,$$
$$P(-1 \leq X_1 \leq 1) = \Phi(1) - \Phi(-1) = \ldots = 0.6826,$$
$$\overline{X} \sim N\left(0, \frac{1}{10}\right),$$
$$P(-1 \leq \overline{X} \leq 1) = P\left(\frac{-1-0}{\sqrt{1/10}} \leq \frac{\overline{X}-0}{\sqrt{1/10}} \leq \frac{1-0}{\sqrt{1/10}}\right)$$
$$= \Phi(\sqrt{10}) - \Phi(-\sqrt{10}) = 2\Phi(\sqrt{10}) - 1 \approx 1.$$

Esimerkki 4.2.2 Erään tilastotoimiston mukaan väestössä keski-ikäisten miesten verenpaineen keskiarvo on 128 ja keskihajonta 15. Halutaan selvittää, poikkeaako keski-ikäisten yritysjohtajien verenpaineen keskiarvo koko väestön vastaavasta. Mitataan 72 yritysjohtajan verenpaineet ja saadaan keskiarvoksi 130.5.

Olkoon X = verenpaine, $X \sim N(128, 15^2)$,

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

$P(\overline{X} \geq 130.5) = 1 - P(\overline{X} \leq 130.5) = 1 - \Phi(\frac{130.5-128}{15/\sqrt{72}}) = 1 - \Phi(1.41) = 0.0793.$

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.

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

Olkoon X = nollien lukumäärä pyöritettäessä 3700 kertaa.

Jos ruletti toimii oikein, niin $X \sim$ Bin(3700, 1/37), E(X) = 100, Var(X) = 3700(1/37)(36/37). Tällöin $X \sim$ 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. Todistat, että pelipaikan ruletti toimii väärin.

Esimerkki 4.2.4 Koneiden A ja B pitäisi valmistaa keskimäärin samanmittaisia tankoja. Molempien koneiden tuotannossa tankojen pituuksissa X ja Y on jonkin verran vaihtelua. Vaihtelua voidaan luonnehtia normaalijakaumalla, jonka varianssi on 0.2. Epäillään kuitenkin koneen A tuottavan keskimäärin pidempiä tankoja. Tutkitaan asiaa valitsemalla satunnaisesti koneen A tuotannosta 20 ja koneen B tuotannosta 10 tankoa. Koneen A tuotannosta valittujen tankojen keskipituudeksi saatiin 40.0 ja koneelta B valittujen 39.5. Onko epäily aiheellinen?

$$
\begin{aligned}
&\text{Kone A: } X_i \sim N(\mu, 0.2), \qquad i = 1, \ldots, 20, \\
&\text{Kone B: } Y_i \sim N(\mu, 0.2), \qquad i = 1, \ldots, 10, \\
&\overline{X} \sim N\left(\mu, \frac{0.2}{20}\right), \qquad \overline{Y} \sim N\left(\mu, \frac{0.2}{10}\right), \\
&E(\overline{X} - \overline{Y}) = \mu - \mu = 0, \\
&\operatorname{Var}(\overline{X} - \overline{Y}) = \operatorname{Var}(\overline{X}) + (-1)^2 \operatorname{Var}(\overline{Y}) = 0.03.
\end{aligned}
$$

Siis $\overline{X} - \overline{Y} \sim N(0, 0.03)$, joten

$$P(\overline{X} - \overline{Y} \geq 0.5) = 1 - P(\overline{X} - \overline{Y} \leq 0.5) = 1 - \Phi\left(\frac{0.5 - 0}{\sqrt{0.03}}\right) = 1 - \Phi(2.89) = 0.0019.$$

Epäily on aiheellinen, koska jos koneet tuottaisivat keskimäärin samanmittaisia tankoja, niin olisi harvinaista saada otokset, joiden keskiarvojen erotus olisi suurempi kuin 0.5.

Luku 5

Parametrien estimointi

5.1 Piste-estimointi

Estimointi on populaation tuntemattoman parametrin arviointia sopivan otossuureen avulla. Näin tehtäessä puhutaan *piste-estimoinnista*. Esimerkiksi voidaan estimoida populaation odostusarvoa otoskeskiarvolla, populaation varianssia otosvarianssilla.

Esimerkki 5.1.1 Olkoon populaatiossa π % viallisia. Pyritään arvioimaan π:tä otoksen perusteella. Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos ko. populaatiosta. Määritellään

$$X_i = \begin{cases} 1, & \text{jos alkio viallinen;} \\ 0, & \text{jos alkio viaton.} \end{cases}$$

Näin siis $X_i \sim \text{Ber}\left(\frac{\pi}{100}\right)$, jolloin $E(X_i) = \frac{\pi}{100}$ ja $\text{Var}(X_i) = \frac{\pi}{100}\left(1 - \frac{\pi}{100}\right)$. Viallisten kokonaislukumäärä otoksessa on $X = X_1 + X_2 + \cdots + X_n$.

Luonnollinen arvio π:lle olisi vastaava luku otoksessa eli viallisten prosenttiosuus otoksessa

$$p = \frac{100X}{n} = \frac{100(X_1 + X_2 + \cdots + X_n)}{n}.$$

Kun p on otossuure, jolla estimoidaan π:tä, sanotaan, että p on π:n *estimaattori*. Kun otos on tehty, voidaan p:lle laskea arvo eli *estimaatti*. Otossuureen p odotusarvo ja varianssi:

$$\begin{aligned} E(p) &= E\left(\frac{100X}{n}\right) = \frac{100}{n} E(X_1 + \cdots + X_n) = \frac{100}{n} \cdot n \cdot \frac{\pi}{100} = \pi, \\ \text{Var}(p) &= \text{Var}\left(\frac{100X}{n}\right) = \left(\frac{100}{n}\right)^2 \text{Var}(X_1 + \cdots + X_n) \\ &= \left(\frac{100}{n}\right)^2 \cdot n \cdot \frac{\pi}{100}\left(1 - \frac{\pi}{100}\right) = \frac{\pi(100 - \pi)}{n}. \end{aligned}$$

Koska $E(p) = \pi$, niin sanotaan, että p on π:n *harhaton* estimaattori. Harhattomuus tarkoittaa siis sitä, että estimaattori antaa keskimäärin oikeita arvoja. Otossuureen p hajontaa sanotaan otoksen *prosenttiosuuden keskivirheeksi*.

Keskeisen raja-arvolauseen perusteella voidaan sanoa, että

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

Saatiin siis selville otossuureen p otosjakauma.

Yksi tapa estimoida populaation parametri on tehdä se otoksesta lasketun vastaavan tunnusluvun avulla (analogiaperiaate):

estimoitava parametri	estimaattori
odotusarvo	otoskeskiarvo
populaation varianssi	otosvarianssi
populaation mediaani	otosmediaani
”viallisten”%-osuus populaatiossa	”viallisten”%-osuus otoksessa

On tietysti monenlaisia otossuureita, joita voitaisiin käyttää parametrien estimoinnissa. Estimaattorille asettaa kuitenkin erilaisia vaatimuksia. Harhattomuus on yksi estimaattorin toivottu ominaisuus.

Olkoon θ populaation tuntematon, estimoitava parametri ja $\hat{\theta}$ sen estimaattori. Tällöin sanotaan, että $\hat{\theta}$ on θ:n harhaton estimaattori, jos $E(\hat{\theta}) = \theta$.

Harhattomuuden lisäksi estimaattorilla toivotaan olevan pienin mahdollinen varianssi. Jos estimaattori on harhaton ja sillä on pienin varianssi parametrin kaikkien harhattomien estimaattoreiden joukossa, sanotaa estimaattoria *harhattomaksi minimivarianssiseksi estimaattoriksi* eli *tehokkaimmaksi estimaattoriksi*.

Kahdesta parametrin harhattomasta estimaattorista on tehokkaampi se, jolla on pienempi varianssi. Otoskoon kasvaessa toivotaan estimoinnin tarkentuvan eli estimaattorin jakauman keskittyvän yhä tiiviimmin estimoitavan parametrin ympärille. Jos estimaattorin varianssi lähenee nollaa otoskoon kasvaessa rajatta, sanotaan, että estimaattori on tarkentuva. Luonnollinen vaatimus tietenkin estimaattorille on myös se, että käytetään kaikki otoksessa oleva informaation hyväksi.

Esimerkki 5.1.2 Otoskeskiarvo $\overline{X}$ on jakauman odotusarvon μ harhaton estimaattori, koska $E(\overline{X}) = \mu$. Aiemmin on myös todettu, että $\mathrm{Var}(\overline{X}) = \sigma^2/n$. Lisäksi voidaan osoittaa, että normaalijakauman tapauksessa μ:n harhattomien estimaattoreiden joukossa otoskeskiarvolla on pienin varianssi.

Esimerkki 5.1.3 Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos populaatiosta, jonka varianssi on σ^2. Voidaan osoittaa, että otosvarianssi $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ on σ^2:n harhaton estimaattori eli $E(S^2) = \sigma^2$.

Vaikka otosvarianssi onkin populaation varianssin harhaton estimaattori, niin otoshajonta ei yleensä ole populaation hajonnan harhaton estimaattori.

On olemassa monenlaisia estimointimenetelmiä edellä esitellyn lisäksi, mm. pienimmän neliösumman menetelmä, maximum likelihood -menetelmä.

5.2 Luottamusvälejä

Piste-estimointi tuottaa siis (otoksen teon jälkeen) yhden luvun, jolla arvioidaan estimoitavaa parametria. Estimointiin liittyy tietysti aina epävarmuutta. Usein halutaankin määrätä yksittäisen arvon sijaan väli, jolla arvellaan tuntemattoman parametrin olevan. Tällöin puhutaan *väliestimoinnista*. Väliestimoinnissa muodostetaan nk. *luottamusväli* vastaavan piste-estimaattorin ja piste-estimaattorin otantajakauman keskihajonnan eli *estimaattorin keskivirheen* avulla.

Olkoon A ja B satunnaisotoksen perusteella määriteltyjä satunnaismuuttujia. Väli (A, B) on parametrin θ $100(1-\alpha)$ %:n luottamusväli, jos $P(A \leq \theta \leq B) = 1-\alpha$.

Kyseessä on siis satunnaisväli, joka sisältää populaation tuntemattoman estimoitavan parametrin todennäköisyydellä $1-\alpha$. Kun otos on tehty, voidaan A:lle ja B:lle laskea arvot. Näin saadaan väli (a, b), joka joko sisältää parametrin θ tai ei sisällä. Välistä (a, b) käytetään myös nimitystä luottamusväli. Koska päättely halutaan tehdä melko suurella varmuudella, valitaan α esim. 0.05 tai 0.01. Tällöin siis määritetään 95 %:n tai 99 %:n luottamusväli. Luottamustaso on 0.95 tai 0.99.

Määritellään standardoituun normaalijakaumaan liittyvä merkintä, jota tarvitaan mm. luottamusvälien määrittämisessä. Olkoon $Z \sim N(0,1)$. Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$. Samoin $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$. Esimerkiksi $z_{0.05} = 1.64$ ja $z_{0.05/2} = z_{0.025} = 1.96$.

Graafisesti, ks.http://mtl.uta.fi/tilasto/tiltp2/syksy2004/zalfa.pdf.

5.2.1 Populaation odotusarvon luottamusväli

Esimerkki 5.2.1 Halutaan arvioida poikien keskimääräistä syntymäpituutta. Otoksessa 65 pojan syntymäpituuden keskiarvo oli 50.95 cm ja keskihajonta 1.97 cm (SAIDIT -aineisto). Miten voisi arvioida poikapopulaation keskiarvoa?

Seuraavaksi arvioidaan normaalijakauman odotusarvoa, kun tunnetaan populaation varianssi. Näinhän ei tietysti voida poikien keskipituuden arvioinissa edellä olettaa.

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

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

jolloin

$$P\left(-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

Kirjoittamalla lausuttu tapahtuma toiseen muotoon saadaan

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Voidaan sanoa, että epäyhtälöt toteutuvat todennäköisyydellä 0.95.

Väliä $(\overline{X} - 1.96\,\sigma/\sqrt{n},\ \overline{X} + 1.96\,\sigma/\sqrt{n}\,)$ sanotaan *μ:n 95 %:n luottamusväliksi*. Luottamusvälin määritelmässä $\overline{X}$ on siis satunnaismuuttuja, jonka arvot vaihtelevat otoksesta toiseen. Havaitun otoksen perusteella saadaan kiinteä väli, jota myös kutsutaan luottamusväliksi.

Tämän säännön mukaan laskettu väli pitää sisällään 95 %:n todennäköisyydellä tuntemattoman populaatiokeskiarvon μ. Poimittaessa monta otosta ja laskettaessa joka kerta edellä esitetty luottamusväli, niin luottamusväleistä n. 95 % on sellaisia, jotka sisältävät μ:n.

Vastaavalla tavalla kuin 95 %:n luottamusväli, voidaan muodostaa myös 99 %:n luottamusväli.

Yleisesti, jos $0 < \alpha < 1$ (tavallisesti 0.05 tai 0.01), niin $100(1-\alpha)$ %:n luottamusväli populaation odotusarvolle μ, kun varianssi tunnettu, on

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Esimerkki 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 pussituskone tuottaa säätöjen jälkeen keskimäärin kilon pusseja?

Odotusarvon μ luottamusväli, kun σ tunnettu

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Nyt $\overline{x} = 1002$, $\sigma = 2.5$, $n = 20$, $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, joten 95 %:n luottamusväli μ:lle on

$$1002 \pm 1.96 \cdot \frac{2.5}{\sqrt{20}},$$

Saatu luottamusväli (1000.9, 1003.1) ei sisällä kiloa. Päätellään, että kone ei tuota keskimäärin kilon pusseja. Sama päättely tehdään 99 %:n luottamusvälin (1000.6, 1003.4) perusteella

Edellä esitetyssä oletettiin, että meillä on satunnaisotos normaalijakaumasta, jolloin otoskeskiarvon jakauma on myös normaalijakauma. Esitettyä luottamusvälin laskukaavaa voidaan kuitenkin käyttää otoskoon ollessa suuri siinäkin tapauksessa, että satunnaisotos on peräisin jostain muusta kuin normaalijakaumasta. Tällöinhän keskeisen raja-arvolauseen perusteella otoskeskiarvon jakauma on likimain normaalijakauma.

Edellä esitetyssä oletettiin myös, että jakauman varianssi on tunnettu. Käytännössä harvemmin tietysti populaation varianssia tunnetaan. Tällöin se onkin estimoitava otoksen perusteella käyttäen otosvarianssia.

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

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

noudattaa ns. Studentin t-jakaumaa vapausastein $n - 1$.

Studentin t-jakauma, joka määritellään nk. vapausastein (df), on jatkuva, origon suhteen symmetrinen jakauma, merkitään t_{df} (tai $t(df)$). Suurilla vapausasteilla t-jakauma lähestyy standardoitua normaalijakaumaa.

Olkoon t_{df} Studentin t-jakaumaa noudattava satunnaismuuttuja. Määritellään $t_{\alpha;df}$ siten, että $P(t_{df} \geq t_{\alpha;df}) = \alpha$ ja $P(t_{df} \geq t_{\alpha/2;df}) = \alpha/2$.

Näitä $t_{\alpha;df}$ arvoja on taulukoitu.

Esimerkki 5.2.3

$$\begin{aligned} &P(t_{10} > 1.812) = 0.05, \\ &P(t_{20} \leq 2.086) = 1 - 0.025 = 0.975, \\ &P(t_{120} \leq -1.98) = 0.025. \end{aligned}$$

Graafisesti, ks. http://mtl.uta.fi/tilasto/tiltp2/syksy2004/talfa.pdf

Esimerkki 5.2.4

$$\begin{aligned} t_{0.05;10} &= 1.812, \\ t_{0.05;30} &= 1.697, \\ t_{0.01;10} &= 2.764, \\ t_{0.01;30} &= 2.457. \end{aligned}$$

Nyt $100(1 - \alpha)$ *%:n luottamusväli populaation odotusarvolle* μ*, kun varianssi tuntematon*, on

$$\overline{X} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}.$$

Vaikka otos ei olisikaan peräisin normaalijakaumasta, voidaan taas riittävän suurilla n:n arvoilla luottamusväli laskea edellä esitetyllä tavalla.

Esimerkki 5.2.5 Halutaan arvioida poikien ja tyttöjen keskimääräisiä syntymäpituuksia, otoksena SAIDIT -aineisto.

$100(1 - \alpha)$ %:n luottamusväli μ:lle, kun σ tuntematon

$$\overline{X} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}.$$

Pojat:

$$n = 65, \quad \overline{x} = 50.95, \quad s = 1.972, \quad t_{0.025;64} \approx 2,$$

joten 95 %:n luottamusväli poikien keskipituudelle on

$$50.95 \pm 2 \cdot \frac{1.972}{\sqrt{65}} \qquad (50.46, 51.44).$$

Tytöt:

$$n = 55, \quad \overline{x} = 50.24, \quad s = 2.027, \quad t_{0.025;54} \approx 2,$$

joten 95 %:n luottamusväli tyttöjen keskipituudelle on

$$50.24 \pm 2 \cdot \frac{2.027}{\sqrt{55}} \qquad (49.69, 50.79).$$

Esimerkki 5.2.6 Keskimääräinen neliövuokra Tampereen Hervannassa 2011.

$$n = 26, \quad \overline{x} = 12.32, \quad s = 2.25, \quad t_{0.025;25} = 2.060,$$

joten 95 %:n luottamusväli keskineliövuokralle on

$$12.32 \pm 2.060 \cdot \frac{2.25}{\sqrt{26}} \qquad (11.41, 13.23).$$

5.2.2 Prosentuaalisen osuuden luottamusväli

Esimerkki 5.2.7 Puolue haluaa arvioida kannatusprosenttinsa ja kysyy sadalta kansalaiselta mielipidettä. Sadan vastaajan joukossa on kannattajia 18 %. Todellista kannatusprosenttia π ei siis tiedetä, mutta sitä voidaan arvioida muodostamalla luottamusväli kyselyssä saatujen lukujen perusteella.

Olkoon populaatiossa tietyn tyyppisiä alkioita π %, kutsutaan näitä jatkossa viallisiksi. Halutaan arvioida tätä lukua π satunnaisotoksen (otoskoko n) perusteella. Olkoon p = viallisten prosenttiosuus otoksessa. Nyt $p \sim N(\pi, \pi(100 - \pi)/n)$ (likimain), joten

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

Tämän perusteella saadaan (menetellen kuten odotusarvon luottamusvälin yhteydessä ja korvaamalla p:n hajonnassa π estimaattorillaan p) $100(1 - \alpha)$ *%:n luottamusväli π:lle*:

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

Esimerkki 5.2.8 Yritys tekee tiettyä komponenttia, jota käytetään auton moottorissa. Yritys valvoo tuotantoaan; virheellisten komponenttien osuus ei saisi olla suurempi kuin 4 %. Laaduntarkkailussa tehtiin 500 komponentin otos, jossa 28 komponenttia osoittautui virheellisiksi. Voidaanko päätellä, että prosessi tuottaa virheellisiä komponentteja yli sallitun rajan?

Muodostetaan 95 %:n luottamusväli suhteelliselle osuudelle. Luottamusväli on $p \pm 1.96\sqrt{p(100-p)/n}$. Nyt $p = 5.6$ ja $n = 500$, joten luottamusvälin alaraja on 3.6 ja ylärajaksi 7.6. Virheellisten osuuden arvellaan olevan välillä 3.6 % – 7.6 %, joten vaihtelu on sallituissa rajoissa, koska 4 % kuuluu arvioidulle välille.

5.2.3 Kahden populaation odotusarvojen erotuksen luottamusväli

Esimerkki 5.2.9 Jos halutaan selvittää, ovatko pojat ja tytöt syntyessään keskimäärin saman painoisia, niin tehdään tyttö- ja poikapopulaatioista satunnaisotokset ja arvioidaan otoskeskiarvojen avulla kahden populaation odotusarvojen yhtäsuuruutta.

Käytännössä populaatioiden varianssit ovat tuntemattomia, mutta lähdetään liikkeelle olettaen ne tunnetuiksi.

Olkoon $(X_1, X_2, \ldots, X_n)$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja olkoon $(Y_1, Y_2, \ldots, Y_m)$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta, missä σ_1 ja σ_2 *tunnettuja* sekä satunnaisotokset toisistaan riippumattomia. Tällöin $\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m)$, johon perustuen odotusarvojen erotuksen $\mu_1 - \mu_2$ $100(1-\alpha)$ *%:n luottamusväli* on

$$\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}},\ \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right).$$

Käytännössä tietysti tilanne on sellainen, että populaatioiden variansseja ei tunneta. Olettaen *varianssit tuntemattomiksi, mutta yhtä suuriksi* voidaan otoskeskiarvojen erotuksen varianssia estimoida otosvarianssien avulla ja saadaan odotusarvojen erotuksen $\mu_1 - \mu_2$ $100(1-\alpha)$ *%:n luottamusväli*

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

missä

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

Suurten otosten tapauksessa tuloksia voidaan käyttää myös muidenkin kuin normaalijakaumien yhteydessä. Jos populaatioiden varianssit ovat tuntemattomia eikä ole perusteltua olettaa yhtä suuruutta, niin silloin suurten otosten tapauksessa on mahdollista muodostaa odotusarvojen erotukselle luottamusväli, jonka määritys riippuu populaatio-oletuksista.

Esimerkki 5.2.10 Ovatko tytöt ja pojat syntyessään keskimäärin saman painoisia? Ks. Esim. 1.0.2.

Luottamusväli odotusarvojen erotukselle, kun populaation varianssit tuntemattomia, mutta yhtä suuria

$$\overline{X} - \overline{Y} \pm t_{\alpha/2;n+m-2}\, s\sqrt{\frac{1}{n} + \frac{1}{m}}, \quad \text{missä} \quad s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

$$\overline{x} = 3640.46, \quad s_X = 438.24, \quad n = 65,$$
$$\overline{y} = 3451.27, \quad s_Y = 523.28, \quad m = 55,$$
$$t_{0.025;65+55-2} \approx 1.98,$$
$$s^2 = \frac{(65-1)438.24^2 + (55-1)523.28^2}{65+55-2},$$
$$s \approx 479.0,$$
$$s\sqrt{\frac{1}{n}+\frac{1}{m}} \approx 87.76,$$
lv: $189.19 \pm 1.98 \cdot 87.76$.

Esimerkki 5.2.11 Tarkasteltaessa miesten ja naisten eroja musikaalisuuden suhteen saatiin suoritetussa kokeessa 20 miehelle ja 25 naiselle seuraavat pisteluvut:

Miehet: 50 45 47 56 37 40 52 50 45 33 31 48 49 42 42 57 46 28 43 37
Naiset: 38 43 46 38 49 42 28 50 47 32 41 48 49 42 42 57 46 52 33 37
48 37 36 39 35

Näistä lasketut tunnusluvut:

	Miehet	Naiset
Otoskeskiarvo	43.9	42.2
Otoskeskihajonta	7.86	6.98

Onko musikaalisuuden suhteen eroja miesten ja naisten välillä?

Kuten esim. 5.2.10 luottamusväli odotusarvojen erotukselle, kun populaation varianssit tuntemattomia, mutta yhtä suuria.

$$\overline{X} - \overline{Y} \pm t_{\alpha/2;n+m-2} s\sqrt{\frac{1}{n}+\frac{1}{m}}, \quad \text{missä} \quad s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

$$\alpha = 0.05, \quad t_{0.025;20+25-2} \approx 2.021, \quad n_M = 20, \quad n_N = 25,$$
$$\overline{Y}_M = 43.9, \quad \overline{Y}_N = 42.2,$$
$$s^2 = \frac{(n_M-1)s_M^2 + (n_N-1)s_N^2}{n+m-2} = \frac{(20-1)7.86^2 + (25-1)6.98^2}{20+25-2},$$
$$s \approx 7.38.$$

95 %:n luottamusväli odotusarvojen erotukselle:

$$43.9 - 42.2 \pm 2.021 \cdot 7.38\sqrt{\frac{1}{20}+\frac{1}{25}} \quad \text{eli} \quad (-2.77, 6.17).$$

Koska nolla kuuluu luottamusvälille, ei ole syytä väittää, että naisten ja miesten pisteluvuissa olisi tasoeroja (populaatiossa).

Luottamusväliä kahden populaation odotusarvojen erotukselle voidaan käyttää, kun selitettävä muuttuja on kvantitatiivinen ja selittäjä on kaksiluokkainen (tai luokiteltu siten). Jos luottamusväli sisältää nollan, niin voidaan tehdä johtopäätelmä, että odotusarvot ovat samoja.

5.2.4 SPSS -ohjeita

1) Luottamusväli populaation odotusarvolle

Analyze
Compare Means ▶ One-Sample T Test...

Muuttujan oltava vähintään intervalliasteikollinen.

2) Luottamusväli populaation odotusarvojen erotukselle riippumattomien otosten tilanteessa.

Analyze
Compare Means ▶ Independent-Samples T Test...

Riippuvan muuttujan oltava vähintään intervalliasteikollinen, selittävää muuttujaa tarkastellaan kahdessa luokassa.

3) Luottamusvälit prosentuaalisille osuuksille. Ohjelmistolla lasketaan prosentuaaliset osuudet aineistossa esim. frekvenssijakauman avulla

Analyze
Descriptive Statistics ▶

ja tämän jälkeen itse kyseinen luottamusväli.

Luku 6

Hypoteesien testaus

Tilastollinen *hypoteesi* on väite populaatiosta, sen jakaumasta tai jakauman parametrista. Väittämä voi liittyä myös useampaan populaatioon.

Hypoteesin testaus tarkoittaa väittämän tutkimista otoksen perusteella. Väitteen paikkansa pitävyyttä tutkitaan otoksen (käytettävissä olevan aineiston) perusteella käyttämällä tilanteeseen sopivaa nk. testisuuretta. Otoksesta lasketun testisuureen arvon perusteella joko uskotaan väite tai ei uskota (jolloin vaihtoehtoinen väite hyväksytään). Johtopäätelmän tekeminen perustuu siihen, että selvitetään voidaanko otoksesta laskettua testisuureen arvoa väitteen ollessa tosi pitää ”tavanomaiseten” arvojen joukkoon kuuluvana vai katsotaanko se harvinaisten arvojen joukkoon kuuluvaksi. Jos testisuureeen arvo kuuluu harvinaisten arvojen joukkoon, niin ei uskota väitettä. Mikä sitten on harvinaista? Testauksessa harvinaisiksi arvoiksi katsotaan sellaisten arvojen joukko, jonka todennäköisyys on melko pieni, esim. pienempi kuin 0.05. Testauksessa onkin tapana ilmoittaa nk. p-arvo, joka kertoo todennäköisyyden saada väitteen ollessa tosi otoksesta saatua arvoa harvinaisempi arvo. Tämä on siis pienin riskitaso, jolla asetettu väite voidaan hylätä. Jos siis testaukseen liittyvä p-arvo on pieni, sanotaan vaikkapa 0.01, niin asetettua väitettä ei uskota; se hylätään ja hyväksytään vaihtoehtoinen väittämä. Siis kun p-arvo on pieni (esim. pienempi kuin 0.05, 0.01, 0.001), niin ei ole syytä uskoa väittämää. Se milloin p-arvon katsotaan olevan tarpeeksi pieni, riippuu siitä millainen todennäköisyys sallitaan sille, että tehdään väärä johtopäätelmä; väärä siten, että väittämä hylätään vaikka sen on tosi.

Hypoteesin testauksessa asetetaankin siis kaksi väittämää, joista jompi kumpi on välttämättä voimassa: Nollahypoteesi H_0, jonka ollessa tosi testisuuren jakauma tunnetaan sekä vaihtoehtoinen hypoteesi H_1. Nollahypoteesi H_0 tulee aina asettaa käytetyn testin sanelemalla tavalla.

Esimerkki 6.0.1

$$H_0\colon \mu = 25 \qquad H_0\colon \mu = 25$$
$$H_1\colon \mu \neq 25 \qquad H_1\colon \mu < 25$$
$$H_0\colon \mu = 10 \qquad H_0\colon \pi = 50$$
$$H_1\colon \mu = 20 \qquad H_1\colon \pi \neq 50$$

Käytännön tilanteissa hypoteesit on hyvä kirjoittaa tutkimustilanteeseen liittyvin termein.

Hypoteesi voi olla joko yksinkertainen tai yhdistetty. Hypoteesi on yksinkertainen, jos parametriväittämään on kiinnitetty vain yksi arvo, esim. $H\colon \theta = 10$. Yhdistetyssä hypoteesissa väitetään parametrin olevan jollain välillä esim. $\theta \neq 10$, $\theta < 10$.

Jatkossa tarkastellaan tilanteita, jossa nollahypoteesi on yksinkertainen ja vaihtoehtoinen hypoteesi yhdistetty.

Vaihtoehtoinen hypoteesi voi olla *kaksisuuntainen* (-puolinen)

$$H_1\colon \theta \neq \theta_0$$

tai *yksisuuntainen* (-puolinen)

$$H_1\colon \theta > \theta_0,$$
$$H_1\colon \theta < \theta_0.$$

Yksisuuntaista hypoteesia käytettäessä on oltava ennakkoinformaatiota θ:sta.

Tilastollinen *testi* on sääntö, joka johtaa asetetun hypoteesin hyväksymiseen tai hylkäämiseen. Testauksessa määritellään *otossuure, jonka jakauma tunnetaan nollahypoteesin ollessa tosi*. Havaitun otoksen perusteella saadaan otossuurelle arvo, jonka avulla päätellään, voidaanko väittämä hyväksyä. *Väittämän hyväksymisen tai hylkääminen perustetaan siihen, kuinka todennäköistä on saada saatua arvo pienempi tai suurempi arvo nollahypoteesin ollessa tosi.*

Testauksessa käytetystä otossuureesta käytetään nimitystä *testisuure*.

Esimerkki 6.0.2

$$H_0\colon \mu = 25,$$
$$H_1\colon \mu \neq 25.$$

Tutkittaessa onko jakauman odotusarvo 25, tehdään satunnaisotos kyseisestä jakaumasta. Otoskeskiarvohan on odotusarvon harhaton estimaattori, joten käytetään sitä hyväksi asetetun nollahypoteesin testauksessa. Lasketaan otoksesta keskiarvo ja verrataan sitä H_0:n mukaiseen tilanteeseen. Pienet poikkeamat 25:stä eivät johda H_0:n hylkäämiseen, mutta kun poikkeama on ”tarpeeksi suuri” väittämä hylätään ja tehdään johtopäätelmä, että H_0 ei ole tosi eli otos ei ole peräisin jakaumasta jonka odotusarvo on 25.

Kuinka suuren poikkeaman sitten pitää olla, jotta H_0 voidaan hylätä? Tämän kysymyksen ratkaisussa voidaan käyttää otossuureen jakaumaa H_0:n ollessa tosi.

Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos $N(25, \sigma^2)$:sta, missä σ^2 on tunnettu. Tällöin otoskeskiarvo

$$\overline{X} \sim N\left(25, \frac{\sigma^2}{n}\right)$$

ja

$$Z = \frac{\overline{X} - 25}{\sigma/\sqrt{n}} \sim N(0,1).$$

Nyt voidaankin käyttää Z-testisuuretta asetetun hypoteesin testaukseen. Lasketaan otoksen perusteella Z:lle arvo z ja katsotaan kuinka hyvin arvo sopii standardoidun normaalijakauman arvoksi.

Nythän

$$P\left(-1.96 \leq \frac{\overline{X} - 25}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

Jos otoksesta saatu Z:n arvo on vaikkapa 2.5, niin silloin H_0 on tosi, mutta on saatu harvinainen otos tai H_0 on epätosi.

Tässä testaus voidaan suoritetaan siten, että H_0 hyväksytään, jos otoksesta laskettu $|z| \leq 1.96$ ja hylätään, jos $|z| > 1.96$. Testi hylkää oikean H_0:n todennäköisyydellä 0.05. Tätä todennäköisyyttä kutsutaan testin *merkitsevyys- eli riskitasoksi*, merk. α. Tämän todennäköisyyden testaaja voi itse kiinnittää. Kuitenkin sen halutaan olevan melko pieni, kuten 0.05, 0.025, 0.01, 0.001.

Testauksen vaiheet:

1. Valitaan riskitaso α.
2. Muodostetaan testisuureen otosjakauma, kun H_0 *on tosi*. Määrätään testisuureen harvinaisten arvojen joukko eli testin *kriittinen alue*.
3. Lasketaan otoksesta testisuureelle arvo.
4. Hylätään H_0, jos saatu arvo kuuluu kriittiselle alueelle, muulloin hyväksytään.

Testin kriittinen alue riippuu tietysti valitusta α:sta, mutta myös vaihtoehtoisesta hypoteesista (yksi-/kaksisuuntainen).

Jos kaksisuuntaisessa testissä

$$H_0 \colon \mu = 25,$$
$$H_1 \colon \mu \neq 25,$$

niin H_0 hyväksytään, jos otoksesta laskettu $|z| \leq z_{\alpha/2}$, ja hylätään, jos $|z| > z_{\alpha/2}$.

Jos yksisuuntaisessa testissä

$$H_0 \colon \mu = 25,$$
$$H_1 \colon \mu < 25,$$

niin H_0 hyväksytään, jos otoksesta laskettu $z \geq -z_\alpha$ ja hylätään, jos $z < -z_\alpha$.

Jos yksisuuntaisessa testissä

$$H_0 \colon \mu = 25,$$
$$H_1 \colon \mu > 25,$$

niin H_0 hyväksytään, jos otoksesta laskettu $z \leq z_\alpha$ ja hylätään, jos $z > z_\alpha$.

Testauksessa ei siis tiedetä kumpi hypoteesi on tosi. Jos todellinen tilanne on se, että H_0 on tosi, niin todennäköisyys sille, että H_0 hylätään testin perusteella on α. Voi olla tietysti myös sellainen tilanne, että todellisuudessa H_1 on tosi. Tällöinkin voidaan tehdä virhe testauksessa hyväksymällä H_0. Tätä virhettä kutsutaan 2. lajin virheeksi.

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

Pyritään siihen, että molemmat virheet olisivat mahdollisimman pieniä. Kuitenkin nämä virhetodennäköisyydet riippuvat toisistaan siten, että kun toinen pienenee toinen suurenee. Vaihtoehtoisen hypoteesin tilanteessa testisuureiden jakauman käsittely hankalampaa (esim. H_1: $\mu < 25$).

Testin voimakkuus määritellään todennäköisyydeksi hylätä epätosi nollahypoteesi.

Ks. lisää testauksesta http://www.fsd.uta.fi/menetelmaopetus/hypoteesi/testaus.html.

6.1 Erilaisia testejä

6.1.1 Yhden populaation odotusarvoa koskeva päättely

Esimerkki 6.1.1 Valmistaja väittää, että kynttilöiden keskimääräinen palamisaika on 7 h. Voidaanko tämä väite uskoa?

Asetetaan populaation odotusarvoa koskeva väite

$$H_0\colon \mu = \mu_0.$$

Oletetaan aluksi, että $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 on tunnettu. Tällöin H_0:n ollessa tosi

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

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

Oletetaan seuraavaksi, että $X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta, missä σ^2 on *tuntematon*. Tällöin H_0:n ollessa tosi

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1).$$

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

Edellä esitettyjä testejä voidaan käyttää suurten otosten tapauksessa vaikka populaatio ei olisikaan normaalinen.

Esimerkki 6.1.2 Kauppias väitti, että kananmunien keskipaino on 50 g. Tehdään 25 alkion satunnaisotos ja saadaan $\overline{x} = 47$, $s = 5$. Onko kauppiaan väittämään uskomista?

Asetetaan hypoteesit

$$H_0 \colon \mu = 50$$
$$H_1 \colon \mu < 50.$$

Jos H_0 tosi, niin $t = \frac{\overline{X}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$.

$$t_{\text{hav.}} = \frac{47 - 50}{5/\sqrt{25}} = -3 < -t_{0.005;24} = -2.797$$

H_0 hylätään 0.5 %:n riskitasolla ja H_1 hyväksytään. Pienin riskitaso, jolla H_0 voidaan hyvätä on < 0.005.

Esimerkki 6.1.3 Lepakot paikallistavat hyönteisiä lähettämällä korkeataajuista ääntä. Kaiun kuulemiseen kuluvan ajan perusteella ne pystyvät paikallistamaan hyönteiset. Tutkijat arvelevat, että keskimääräinen tunnistusmatka voisi olla yli 35 cm. He keräävät 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 arviota oikeana?

Nyt

$$\begin{aligned}
&H_0 \colon \mu = 35, \\
&H_1 \colon \mu > 35, \\
&t_{\text{hav.}} = \frac{48.36 - 35}{18.085/\sqrt{11}} = 2.450, \\
&t_{0.01;10} = 2.764 > 2.450, t_{0.025;10} = 2.228 < 2.450.
\end{aligned}$$

Täten H_0 hyväksytään 1 %:n riskitasolla, mutta hylätään 2.5 %:n riskitasolla.

Siis $0.01 < \text{p} < 0.025$.

Esimerkki 6.1.4 Testataan hypoteesia, että populaation odotusarvo on 50. On saatu 30 alkion otoksen perusteella otoskeskiarvoksi 65 ja keskihajonnaksi 11.6. Mikä on pienin riskitaso, jolla nollahypoteesi voidaan hylätä yksisuuntaisessa testissä? Entä kaksisuuntaisessa?

Asetetaan hypoteesit

$$H_0\colon \mu = 50$$
$$H_1\colon \mu > 50.$$

Jos H_0 tosi, niin $t = \frac{\overline{X}-50}{s/\sqrt{n}} \sim t(n-1)$. Nyt

$$t_{\text{hav.}} = \frac{65-50}{11.6/\sqrt{30}} = 2.36,$$
$$t_{0.025;29} = 2.045 < 2.36,$$
$$t_{0.01;29} = 2.462 > 2.36,$$
$$0.01 < p < 0.025.$$

Kaksisuuntaisessa testissä $0.01 < p/2 < 0.025$ eli $0.02 < p < 0.05$.

6.1.2 Yhdessä populaatiossa tietyn tyyppisten alkioiden prosentuaalista osuutta koskeva päättely

Esimerkki 6.1.5 Puolue väittää kannatuksensa olevan 10 %. Voidaanko väite uskoa?

Asetetaan prosenttiosuutta koskeva väite

$$H_0\colon \pi = \pi_0.$$

Oletetaan, että populaatiossa on π % viallisia. Olkoon $X_1, X_2, \ldots, X_n$ satunnaisotos tästä populaatiosta. Jos H_0 on tosi, $p \sim N(\pi_0, \pi_0(100-\pi_0)/n)$, likimain ja

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

missä p on viallisten %-osuus otoksessa.

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

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

Nyt

$$
\begin{aligned}
&H_0\colon \pi = 10,\\
&z = \frac{p-\pi_0}{\sqrt{\pi_0(100-\pi_0)/n}} \overset{\text{likimain}}{\sim} N(0,1), \qquad \text{kun } H_0 \text{ tosi (eli } \pi_0 = 10),\\
&z_{\text{hav.}} = \frac{11.75-10}{\sqrt{10\cdot 90/400}} = 1.17.
\end{aligned}
$$

Yksisuuntaisessa testissä pienin riskitaso, jolla H_0 voidaan hylätä on

$$P(Z > 1.17) = 1 - \Phi(1.17) = 0.121.$$

Kaksisuuntaisessa testissä pienin riskitaso, jolla H_0 voidaan hylätä on

$$2P(Z > 1.17) = 2(1 - \Phi(1.17)) = 0.242.$$

Uskot ystävän väitteen.

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

Nyt

$$
\begin{aligned}
&H_0\colon \pi = 20,\\
&H_1\colon \pi < 20,\\
&z = \frac{p-\pi_0}{\sqrt{\pi_0(100-\pi_0)/n}} \overset{\text{likimain}}{\sim} N(0,1), \qquad \text{kun } H_0 \text{ tosi (eli } \pi_0 = 20),\\
&z_{\text{hav.}} = \frac{16-20}{\sqrt{20\cdot 80/1000}} = -3.16,\\
&-z_{0.001} = -3.08.
\end{aligned}
$$

Koska $z_{\text{hav.}} < -3.08$, niin H_0 hylätään 0,1 %:n riskitasolla. On syytä olettaa, että öljylämmitteisiä asuntoja on vähemmän kuin 20 %. Voidaan myös laskea p-arvo, joka on

$$P(Z < -3.16) = 1 - \Phi(3.16) = 1 - 0.9992 = 0.0008.$$

6.1.3 Kahden jakauman sijainnin vertailu

Esimerkki 6.1.8 Ovatko tytöt ja pojat syntyessään keskimäärin samanmittaisia?

Asetetaan kahden populaation odotusarvoa koskeva väite

$$H_0\colon \mu_1 = \mu_2 \qquad \text{tai} \qquad H_0\colon \mu_1 - \mu_2 = 0.$$

Oletetaan aluksi, että $X_1, X_2, \ldots, X_n$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta ja $Y_1, Y_2, \ldots, Y_m$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta, missä σ_1 ja σ_2 tunnettuja sekä satunnaisotokset toisistaan riippumattomia. Jos H_0 tosi, niin

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

Jos $H_1\colon \mu_1 \neq \mu_2$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $|z| > z_{\alpha/2}$. Jos $H_1\colon \mu_1 < \mu_2$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $z < -z_\alpha$. Jos $H_1\colon \mu_1 > \mu_2$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $z > z_\alpha$.

Jos variansseja σ_1 ja σ_2 ei tunneta, mutta ne voidaan olettaa yhtä suuriksi, niin H_0:n ollessa tosi

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

missä

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

Jos $H_1\colon \mu_1 \neq \mu_2$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $|t| > t_{\alpha/2;n+m-2}$. Jos $H_1\colon \mu_1 < \mu_2$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $t < -t_{\alpha;n+m-2}$. Jos $H_1\colon \mu_1 > \mu_2$, niin H_0 hylätään riskitasolla α, jos otoksesta laskettu $t > t_{\alpha;n+m-2}$.

Esimerkki 6.1.9 Tutkitaan onko lepopulssi naisilla ja miehillä keskimäärin samansuuruinen. On kerätty aineisto, jossa on 36 naista ja 44 miestä. Pulssin keskiarvo otoksessa naisilla on 77.6 ja miehillä 70.6 sekä keskihajonnat 9.8 ja 11.3.

$$\begin{aligned}
&H_0\colon \mu_N = \mu_M,\\
&H_1\colon \mu_N > \mu_M,\\
&\overline{x}_N = 77.6, \quad s_N = 9.8, \quad n_N = 36,\\
&\overline{x}_M = 70.6, \quad s_M = 11.3, \quad n_M = 44,\\
&t = \frac{\overline{X} - \overline{Y}}{s\sqrt{1/n + 1/m}} \sim t(n+m-2), \qquad \text{kun } H_0 \text{ tosi},\\
&s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},\\
&s^2 = \frac{(36-1)\cdot 9.8^2 + (44-1)\cdot 11.3^2}{36+44-2} = 10.65^2,\\
&t_{\text{hav.}} = \frac{77.6 - 70.6}{10.65\sqrt{1/36 + 1/44}} = 2.92.
\end{aligned}$$

Koska $2.92 > t_{0.01;78} \approx 2.39$, niin H_0 hylätään 1 %:n riskitasolla. Päätellään, että naisilla lepopulssi on keskimäärin korkeampi kuin miehillä.

Esimerkki 6.1.10 Psykologi on kehittänyt testin, joka koostuu muutamasta yksinkertaisesta käsin suoritettavista tehtävistä ja jonka tarkoitus on paljastaa mahdollinen lievä kehityshäiriö. Hän on poiminut satunnaisotoksen sekä normaaleista lapsista että kehityshäiriöisistä. Suoritusajat ovat:

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

Kelpaako testi tarkoitukseen?

$$\begin{aligned} &H_0\colon \mu_N = \mu_K, && \overline{x}_N = 207.57, && \overline{x}_K = 247.25, \\ &H_1\colon \mu_N < \mu_K, && s_N^2 = 18.87^2, && s_K^2 = 42.48^2, \\ & && n_N = 7, && n_K = 8. \end{aligned}$$

Kuten esim. 6.1.9. Riippumattomien otosten t-testi odotusarvojen erotukselle.

$$\begin{aligned} t_{\text{hav.}} &= \frac{207.57 - 247.25}{33.7\sqrt{1/7 + 1/8}} = -2.28, \\ -t_{0.01;13} &= -2.65 < -2.28, \\ -t_{0.025;13} &= -2.16 > -2.28. \end{aligned}$$

H_0 voidaan hylätä esim. 2.5 %:n riskitasolla, mutta ei 1 %:n riskitasolla. Jos kiinnitetään 2.5 %:n riski, niin tehdään päätelmä, että testi kelpaa.

6.2 SPSS -ohjeita

1) H_0: $\mu = \mu_0$

Analyze
Compare Means ▶ One Sample T Test...

Muuttujan oltava vähintään intervalliasteikollinen.

2) H_0: $\pi = \pi_0$. Lasketaan vastaava %-osuus otoksesta ja sen avulla z-testisuurelle arvo. Prosenttiosuuden saa selville muodostamalla frekvenssijakauman muuttujasta.

3) H_0: $\mu_1 = \mu_2$ (riippumattomat otokset)

Analyze
Compare Means ▶ Independent-Samples T Test...

Riippuvan muuttujan oltava vähintään intervalliasteikollinen; selittävä muuttuja kahdessa luokassa.

Kirjallisuutta

Kirjallisuusluettelo, jota on käytetty tukena tämän luentorungon kirjoittamisessa.

Agresti, A. & Finlay, B., Statistical Methods for the Social Sciences. Prentice Hall, 1997.

Anderson, T. W. & Sclove, S. L., Introductory Statistical Analysis. Houghton Mifflin Company, 1974.

Clarke, G. M. & Cooke, D., A Basic course in Statistics. Arnold, 1998.

Devore, J. & Peck, R., Statistics, The Exploration and Analysis of Data. West Publishing Company, 1986.

Grönroos, M. Johdatus tilastotieteeseen. Kuvailu, mallit ja päättely, Finn Lectura, 2003.

Helenius, H., Tilastollisten menetelmien perustiedot. Statcon Oy, 1992.

Karjalainen, L. & Ruuskanen, A. Tilastomatematiikka. Pii-kirjat, 1994.

Liski, E. & Puntanen, S., Tilastotieteen peruskurssi I & II. Tampereen yliopisto.

Manninen P., Tilastotiedettä yhteiskuntatieteilijöille. Gaudeamus, 1978.

McClave J& Sincich, T., Statistics, 9th ed., Prentice Hall, 1003.

Mellin, I., Johdatus tilastotieteeseen, 1. kirja, tilastotieteen johdantokurssi, Helsingin yliopisto.

Mellin, I., Johdatus tilastotieteeseen, 2. kirja, tilastotieteen jatkokurssi, Helsingin yliopisto.

Moore, D., The Basic Practice of Statistics, Freeman, 1997.

Moore, D., Introduction to the Practice of Statistics, 3rd ed., Freeman, 1998.

Newbold, P., Statistics for Business and Economics. Prentice Hall, 1995.

Ott, L. & Mendenhall, W., Understanding Statistics. Duxbury Press, 1985.

Siegel, A., Statistics and Data Analysis An Introduction. John Wiley & Sons, 1988.

Tilastollisten päättelyn perusteet, MTTTP5, kaavakokoelma

1 EMPIIRISET JAKAUMAT

(1.1) $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(1.2) $$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) = \frac{1}{n-1}\,SS_x$$

2 TODENNÄKÖISYYSLASKENTAA

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

(2.2) $$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

3 TODENNÄKÖISYYSJAKAUMIA

Diskreetti satunnaismuuttuja X

(3.1) $$E(X) = \mu = \sum_{i=1}^{k} x_i P(X = x_i)$$

(3.2) $$\mathrm{Var}(X) = \sigma^2 = \sum_{i=1}^{k}(x_i - \mu)^2 P(X = x_i)$$

Jatkuva satunnaismuuttuja X

(3.3) $$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

(3.4) $$\mathrm{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x$$

(3.5) $$\mathrm{Cov}(X,Y) = E\big(X - E(X)\big)\big(Y - E(Y)\big)$$

(3.6) $X \sim \text{Ber}(p), \qquad P(X=1)=p, \qquad P(X=0)=1-p,$
$E(X)=p, \qquad \text{Var}(X)=p(1-p)$

(3.7) $X \sim \text{Bin}(n,p), \qquad P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},$
$E(X)=np, \qquad \text{Var}(X)=np(1-p)$

(3.8) $X \sim \text{Tasd}(a,b),$
$P(X=a)=P(X=a+1)=\cdots=P(X=b)=\dfrac{1}{n}, \quad \text{missä } b=a+(n-1),$
$E(X)=\dfrac{a+b}{2}, \qquad \text{Var}(X)=\dfrac{n^2-1}{12}$

(3.9) $X \sim \text{Tas}(a,b), \qquad f(x)=\dfrac{1}{b-a}, \quad a \le x \le b,$
$E(X)=\dfrac{a+b}{2}, \qquad \text{Var}(X)=\dfrac{(b-a)^2}{12}$

(3.10) $X \sim N(\mu,\sigma^2), \qquad f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}(x-\mu)^2/\sigma^2},$
$E(X)=\mu, \qquad \text{Var}(X)=\sigma^2$

4 LUOTTAMUSVÄLEJÄ

<u>μ:lle</u>

(4.1) $\overline{X} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

(4.2) $\overline{X} \pm t_{\alpha/2;n-1}\dfrac{s}{\sqrt{n}}$

<u>π:lle</u>

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

<u>$(\mu_1-\mu_2)$:lle</u>

(4.4) $\overline{X}-\overline{Y} \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n}+\dfrac{\sigma_2^2}{m}}$

(4.5) $\overline{X}-\overline{Y} \pm t_{\alpha/2;n+m-2}\, s\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}\,, \qquad \text{missä } s^2=\dfrac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$

5 TESTISUUREITA

$\underline{H_0 \colon \mu = \mu_0}$

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{5.1}$$

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \tag{5.2}$$

$\underline{H_0 \colon \pi = \pi_0}$

$$Z = \frac{p - \pi_0}{\sqrt{\pi_0(100 - \pi_0)/n}} \overset{\text{likimain}}{\sim} N(0,1) \tag{5.3}$$

$\underline{H_0 \colon \mu_1 = \mu_2}$

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0,1) \tag{5.4}$$

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{1/n + 1/m}} \sim t(n+m-2), \qquad \text{missä } s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \tag{5.5}$$

THE CUMULATIVE DISTRIBUTION FUNCTION OF THE STANDARD NORMAL DISTRIBUTION

$\Phi(z)$

INTEGER AND FIRST DECIMAL OF z	SECOND DECIMAL OF z									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997

	PERCENTAGE POINTS OF STUDENT'S t-DISTRIBUTIONS				
	VALUES OF t FOR VARIOUS α AND DEGREES OF FREEDOM (df)				
	SIGNIFICANCE LEVEL α FOR ONE-TAILED TEST				
df	0,1	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576