

TILASTOLLINEN PÄÄTTELY

S-P. Kantola
Janne Engblom
Syksy 2007

SISÄLLYSLUETTELO

1	TODENNÄKÖISYYSLASKENTAA.....	3
1.1	Peruskäsitteitä.....	3
1.2	Todennäköisyyden tulkinnoista	5
1.3	Todennäköisyyden laskusääntöjä.....	7
1.3.1	Kombinatoriikkaa.....	7
1.3.2	Klassinen todennäköisyys.....	10
1.3.3	Aksiomaattinen todennäköisyys.....	11
1.4	Todennäköisyysjakaumat.....	19
1.4.1	Satunnaismuuttuja ja todennäköisyysjakauma.....	19
1.4.2	Diskreettejä todennäköisyysjakaumia	27
1.4.3	Jatkuvia todennäköisyysjakaumia	32
2	TILASTOLLINEN PÄÄTTELY.....	36
2.1	Johdanto	36
2.2	Estimointi	37
2.2.1	Piste-estimointi.....	44
2.2.2	Väliestimointi.....	47
2.3	Tilastollinen merkitsevyystestaus	50
2.3.1	Hypoteesien testaaminen	51
2.3.2	Testisuureiden jakaumista.....	64
2.3.3	Keskiarvotestit	67
2.3.4	Varianssin testaus.....	78
2.3.5	Suhteellisen osuuden testaus.....	82
2.3.6	Korrelaatiokertoimen testaus	87
2.3.7	Eräitä ei-parametrisia testejä.....	88
3	KIRJALLISUUTTA	92

1 TODENNÄKÖISYYSLASKENTAA

1.1 Peruskäsitteitä

Todennäköisyysslaskenta tarkastelee *satunnaisilmiöitä*. Satunnaisilmiönä voidaan pitää mitä hyvänsä sellaista ilmiötä, johon liittyy useita eri tulostmahdollisuuksia ja tarkastelijan kannalta epävarmuutta siitä, mikä ilmiön tulos tulee olemaan. Voidaan siis ajatella, että satunnaisilmiö on ilmiö, jonka tulokseen vaikuttaa satuma.

Esimerkkejä satunnaisilmiöistä ovat

rahanheitto, tulostmahdollisuuksina kruuna ja klaava,
 arpakuution heittäminen, tulostmahdollisuuksina arpakuution silmäluvut,
 huomisen päivän sää, tulostmahdollisuuksina erilaiset säätilat,
 haastateltavan umpimähkäinen valinta, tulostmahdollisuuksina kaikki mahdolliset vastaukset tehtävään kysymykseen.

Tarkasteltavan satunnaisilmiön mahdollisia tuloksia sanotaan *alkeistapahtumiksi*. Satunnaisilmiön kaikkien mahdollisten tulosten eli kaikkien alkeistapahtumien joukkoa sanotaan *perusjoukoksi* ja merkitään E :llä. Esimerkiksi rahanheiton alkeistapahtumat ovat kruuna (kr) ja klaava (kl) sekä perusjoukko $E = \{kr, kl\}$ ja arpakuution heiton alkeistapahtumat silmäluvut $1, \dots, 6$ ja perusjoukko $E = \{1, 2, 3, 4, 5, 6\}$.

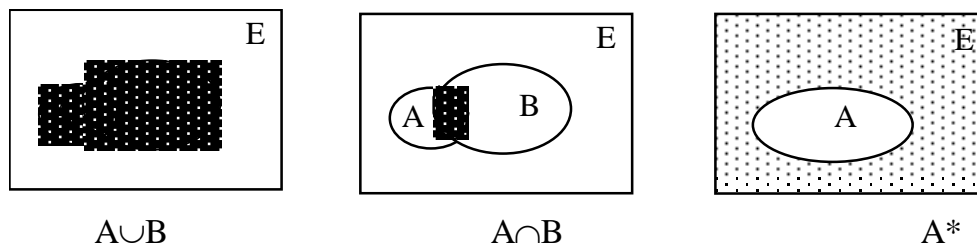
Tapahtumiksi sanotaan perusjoukon osajoukkoja. Näin ollen yksittäisten alkeistapahtumien muodostamat joukot, kuten rahanheiton $A = \{kl\}$, eli “saadaan klaava” ja nopanheiton $A = \{1\}$, eli “saadaan silmäluvuksi 1”, ovat tapahtumia. Tapahtuma voi olla myös eri alkeistapahtumien yhdistelmä, esimerkiksi nopanheiton tapahtuma “saadaan parillinen silmäluku” voidaan ilmoittaa alkeistapahtumien 2, 4 ja 6 muodostamana joukkona, ts. tapahtumana $A = \{2, 4, 6\}$. Joukkoon A kuuluvia alkeistapahtumia sanotaan tapahtumalle A *suotuisiksi alkeistapahtumiksi*.

Jos saatu tulos kuuluu joukkoon A , sanotaan, että *tapahtuma* A *esiintyy*, tai että A *tapahtuu*. Esimerkiksi jos nopanheiton tulos on silmäluku 2, $A = \{2, 4, 6\}$ tapahtuu. Myös koko perusjoukkoa E voidaan pitää tapahtumana. Koska ilmiön tulos kuuluu aina perusjoukkoon, voidaan edellisin käsittein ilmaistuna sanoa, että E tapahtuu aina. Tästä syystä perusjoukkoa E sanotaan *varmaksi tapahtumaksi*. Vastakkainen tilanne esiintyy, jos tapahtuma on tyhjä joukko \emptyset . Tyhjä joukko ei sisällä yhtään alkeistapahtumaa, ja koska satunnaisilmiön tuloksena on aina jokin alkeistapahtumista, voidaan sanoa, että \emptyset ei tapahdu koskaan; se on ns. *mahdoton tapahtuma*.

Koska tapahtumat esitetään joukkoina, eri tapahtumien yhdistelmiä voidaan kuvata joukko-opin operaatioiden avulla. Jos esimerkiksi halutaan viitata tilanteeseen, jossa tapahtumat A tai B tai molemmat esiintyvät yhtäaikaan, voidaan käyttää unionimerkintää $A \cup B$. Seuraavassa esitetään eräiden tapahtumien välisten operaatioiden sanalliset ilmaisut (merkintä A^* tarkoittaa A :n komplementtia):

$A \cup B$ tapahtuu	A tai B tai molemmat tapahtuvat
$A \cap B$ tapahtuu	sekä A että B tapahtuvat
A^* tapahtuu	A ei tapahdu
$A \subset B$	B tapahtuu aina, kun A tapahtuu

Tapahtumia voidaan kuten joukkoja yleensäkin havainnollistaa Venn- diagrammeilla, esimerkiksi seuraavasti



Kun puhutaan tapahtuman A *todennäköisyydestä*, on tarkoitus kuvata sitä, kuinka suuret mahdollisuudet tapahtumalla A on todella tapahtua. Todennäköisyys siis pyrkii kvantifioimaan ilmiöön liittyvää epävarmuutta. Tapahtuman A todennäköisyys $P(A)$ on luku, joka on skaalattu nollan ja ykkösen välille siten, että todennäköisyyden arvo yksi vastaa varmaa tapahtumaa ja nolla mahdotonta tapahtumaa, ts.

$$0 \leq P(A) \leq 1 \text{ ja}$$

jos $P(A) = 1$, A tapahtuu varmasti ja
 jos $P(A) = 0$, A :n tapahtuminen on mahdotonta.

Mitä suurempi todennäköisyys $P(A)$ tapahtumalla A on, sitä suuremmalla varmuudella A tapahtuu.

Todennäköisyyden käsitteeseen ja sen muodolliseen määrittelyyn liittyy syvällisiä matemaattisia ja filosofisia kysymyksiä. Niinpä todennäköisyyttä voidaan lähestyä eri näkökulmista ja tämän mukaan puhutaan todennäköisyyden erilaisista tulkinnoista.

1.2 Todennäköisyyden tulkinnoista

Todennäköisyyslaskennan katsotaan saaneen alkunsa 1600-luvulla uhkapeleihin liittyvien ongelmien tutkimisesta. Todennäköisyyden käsitteen eri puolet olivat kuitenkin tunnettuja jo antiikin filosofeille, jotka käyttivät todennäköisyyttä vastaavia termejä kolmessa merkityksessä: ilmaisemaan objektiivista “toden kaltaisuutta” tai “totuutta lähellä olevaa”, subjektiivista “uskottavuutta” tai “vakuuttavuutta” sekä tapahtuman toistumisfrekvensseihin viitaten “sitä, mikä tavallisesti tapahtuu”. Nykyään todennäköisyyskäsite liitetään lähinnä kahteen viimeksi mainittuun. Nämä eri tulkinnat jakavat nykyiset todennäköisyysteoreetikot omiin koulukuntiinsa, ja vastaavasti puhutaan todennäköisyyden *subjektiivisesta* ja *frekvenssitulkinnasta*.

Todennäköisyyttä koskevan matemaattisen teorian ensi vaihe oli *klassiseksi* kutsuttu lähestymistapa. Klassinen todennäköisyys kärsii eräistä loogisista heikkouksista, mutta intuitiivisena ja soveltuvissa tapauksissa toimivana sitä on käytetty johdatuksena todennäköisyyslaskennan alkeisiin. Nykyinen matemaattinen todennäköisyysteoria perustuu Kolmogorovin 1930-luvulla esittämiin aksioomiin. Erotukseksi klassisesta määrittelystä tätä kutsutaan *aksiomaattiseksi todennäköisyydeksi*. Matemaattisesti korrekti aksioomajärjestelmä ei kuitenkaan ratkaise kysymystä todennäköisyyden tulkinnasta, päinvastoin: jos tällaista abstraktia järjestelmää halutaan käyttää hyväksi, sen peruskäsitteille, joihin todennäköisyys itsessään kuuluu, on annettava jokin tulkinta.

Frekvenssitulkinta

Todennäköisyyden frekvenssitulkintaa sanotaan myös *tilastolliseksi*, *empiiriseksi* tai *objektiiviseksi* todennäköisyydeksi. Se koskee vain sellaisia satunnaisilmiöitä,

jotka toistuvat tai ovat kokeen luontoisesti toistettavissa periaatteessa äärettömän monta kertaa samanlaisissa olosuhteissa. Satunnaisilmiötä sanotaan tällöin *satunnaiskokeeksi*. Jos tapahtuma A esiintyy f_n kertaa, kun satunnaiskoe toistetaan n kertaa, on tapahtuman A suhteellinen frekvenssi $P_n(A) = f_n/n$.

Jos kokeen toistokertojen määrää lisätään äärettömän suureksi, voidaan ajatella suhteellisen frekvenssin lähestyvän jotakin raja-arvoa. Tätä raja-arvoa frekvenssitulkinnassa sanotaan tapahtuman A todennäköisyydeksi, ts.

$$P(A) = \lim_{n \rightarrow \infty} P_n(A) = \lim_{n \rightarrow \infty} \frac{f_n}{n}$$

Suhteelliselle frekvenssille $P_n(A)$ on aina voimassa

$$0 \leq P_n(A) \leq 1,$$

$P_n(A) = 1$ vain, jos A esiintyy joka kokeessa ja

$P_n(A) = 0$ vain, jos A ei esiinny yhdessäkään kokeessa.

On kuitenkin huomattava, että todennäköisyys $P(A)$ suhteellisten frekvenssien raja-arvona voi olla 0, vaikka tapahtuma A esiintyisi koesarjassa useita kertoja ja $P(A)$ voi olla 1 vaikka A jäisi useita kertoja tapahtumatta.

Frekvenssitulkinta siis liittää todennäköisyyden jonkin tapahtumatyyppin esiintymistiheyteen. Tulkinnan heikkoutena on sen soveltumattomuus ainutkertaisten tapahtumien todennäköisyyksien arviointiin. Lisäksi voidaan kyseenalaistaa suhteellisten frekvenssien raja-arvon olemassaolo: tätä ei voida vahvistaa kokemusperäisesti, sillä koesarjat ovat välttämättä äärellisen pituisia, eivätkä teoreettiset perustelutkaan ole aukottomia.

Subjektiiivinen tulkinta

Todennäköisyyden subjektiiivinen tulkinta ymmärtää todennäköisyyden uskottavuutena ja liittää sen näin inhimilliseen tietoon ja persoonallisiin uskomuksiin. Varhaiset todennäköisyysteorian klassikot määrittivät todennäköisyyden uskomuksen asteena, ja subjektiiivisen koulukunnan edustajat katsovat, että uskomuksen aste ei voi olla puhtaasti objektiivinen, vaan se periaatteessa riippuu uskottavuuden arvioijasta. Tapahtuman A todennäköisyys 1 merkitsee siis täyttä subjektiiivistä varmuutta A :n tapahtumisesta ja todennäköisyys nolla täyttä varmuutta siitä, että A ei tapahdu.

Voidaan tietenkin kysyä, miten pysyviä tai konsistentteja ihmisten subjektiiviset uskomukset ovat ja voidaanko niitä yleensä mitata. Subjektiivisen todennäköisyyden teoria tarkastelee idealisoituja, ns. *rationaalisia uskomuksia*, joiden pohjalta muodostettavat todennäköisyydet toteuttavat todennäköisyyden aksioomat.

Subjektiivisen tulkinnan puitteissa voidaan tarkastella sellaisia todennäköisyyksiä, jotka frekvenssitulkinnalla eivät ole mielekkäitä, esimerkiksi hypoteesien oikeellisuuden tai ainutkertaisten tapahtumien todennäköisyyksiä. Subjektiivista todennäköisyysteoriaa sovelletaan mm. päätöksentekoteoriassa.

1.3 Todennäköisyyden laskusääntöjä

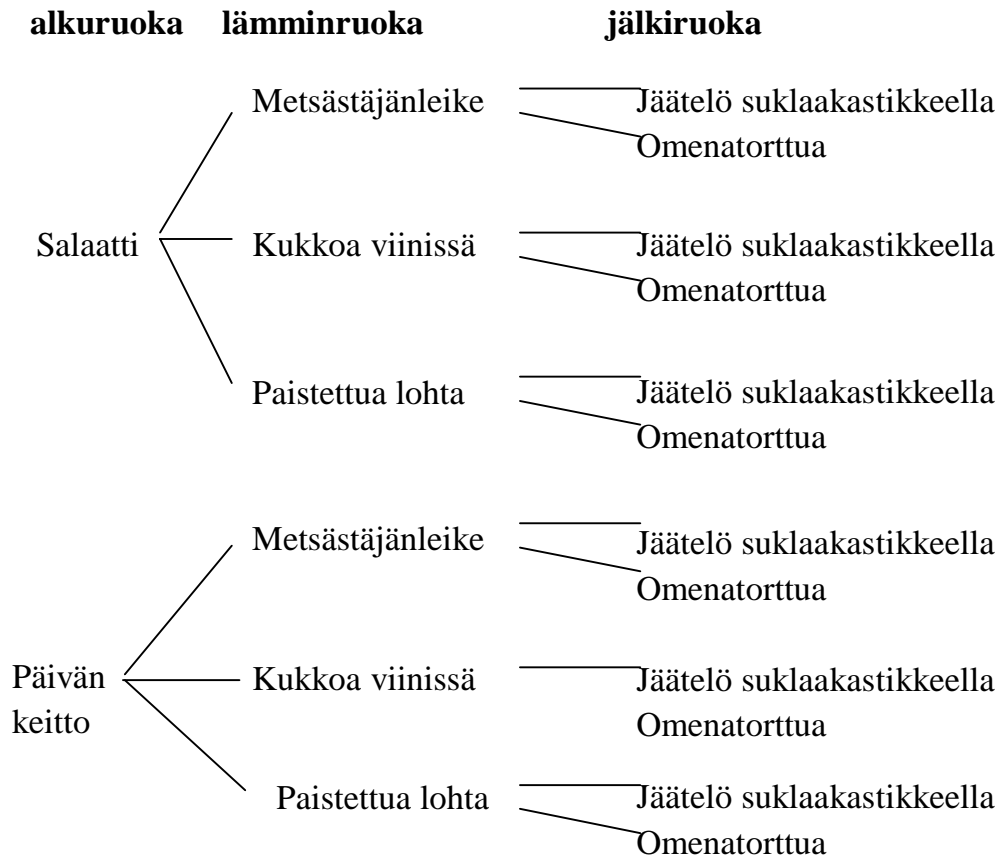
Todennäköisyyslaskennan ongelmissa joudutaan usein laskemaan erilaisten tulomahdollisuuksien lukumääriä. Apuna tässä voidaan käyttää *kombinatoriikkaa*, jota tarkastellaan seuraavaksi.

1.3.1 Kombinatoriikkaa

Kombinatoriikan laskusäännöt koostuvat lähinnä erilaisista periaatteista tai menetelytavoista, joita voidaan soveltaa aina tietyntyyppisiin tehtäviin.

Tuloperiaate

Esimerkki 25: Ravintolan lounaslistalta voi valita alkuruoan kahdesta, lämpimän ruoan kolmesta ja jälkiruoan kahdesta vaihtoehdosta. Montako erilaista kolmen ruokalajin lounasvaihtoehtoa saadaan?



Ratkaisu: Edellä oleva kaavio valaisee ratkaisuperiaatetta. Alkuruoka voidaan valita kahdella eri tavalla. Kumpaankin alkuruokaan voidaan liittää kolme erilaisista lämminruokavaihtoehtoa. Lopuksi jälkiruoka voidaan valita kahdella eri tavalla. Ateriavaihtoehtoja on siis kaikkiaan $2 \cdot 3 \cdot 2 = 12$.

Esimerkin menettelytapa voidaan muotoilla yleisemmin seuraavasti. Kun suoritetaan valinta k kertaa siten, että ensimmäinen valinta voidaan tehdä n_1 tavalla, toinen valinta n_2 tavalla, . . . ja k :s valinta n_k tavalla, on erilaisia tulostulomahdollisuuksia kaikkiaan

$$n_1 \cdot n_2 \cdot \dots \cdot n_k .$$

Permutaatio

Esimerkki 26: Montako erilaista “sanaa” voidaan muodostaa kirjaimista I , S ja O ?

Ratkaisu: Ajatellaan, että sanan muodostamiseksi varataan kolme peräkkäistä paikkaa, joista jokainen täytetään yhdellä kirjaimella. Ensimmäinen paikka voidaan täyttää millä hyvänsä annetuista kolmesta eri kirjaimesta, eli kolmella eri tavalla. Kun ensimmäinen paikka on täytetty jollakin kirjaimella, tätä ei voi enää käyttää, joten toinen paikka voidaan täyttää enää vain kahdella eri tavalla. Kolmatta paikkaa varten jää jäljelle yksi kirjain, joten viimeisen paikan täyttömahdollisuuksia on vain yksi. Mahdollisia tapoja eli erilaisia kolmikirjaimisia sanoja on tuloperiaatteen mukaan kaikkiaan $3 \cdot 2 \cdot 1 = 6$.

Joukon *permutaatioksi* sanotaan mitä tahansa kyseisen joukon alkioden järjestettyä jonoa. Joukon kaikkien mahdollisten permutaatioiden lukumäärä siis kertoo, kunka moneen eri järjestykseen joukon alkiot voidaan asettaa. Jos joukon keskenään erilaisten alkioden lukumäärä on n , joukon permutaatioiden lukumäärä on

$$n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1 = n!$$

Merkintä $n!$ luetaan “ n :n kertoma”. Määritelmää voidaan täydentää vielä sopimuksella

$$0! = 1.$$

Jos n alkioista n_1 on keskenään samanlaisia, n_2 keskenään samanlaisia, \dots ja n_k keskenään samanlaisia, alkioden erilaisia järjestyksiä on määrä

$$\frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}.$$

Variaatio

Esimerkki 27: Lasketaan, montako *järjestettyä paria* voidaan muodostaa kirjaimista A , B , C ja D , siten että sama kirjain ei esiinny parissa kahta kertaa. Järjestetty pari tarkoittaa, että alkioden keskinäisellä järjestyksellä on merkitystä, esim. pari AB on eri kuin pari BA .

Ratkaisu: Parin ensimmäinen alkio voidaan valita neljästä kirjaimesta ja toinen kolmesta jäljelläolevasta, joten tuloperiaatteen mukaan pareja on $4 \cdot 3 = 12$.

Joukon k -*variaatioksi* sanotaan mitä hyvänsä n -alkioisen joukon k :sta alkioista muodostettua järjestettyä jonoa. Tällaisen joukon erilaisten k -variaatioiden ($1 < k < n$) lukumäärä on

$$n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-k+1),$$

tai lyhyemmin

$$\frac{n!}{(n-k)!}.$$

Kombinaatio

Esimerkki 28: Montako erilaista järjestämätöntä paria voidaan muodostaa kirjaimista A , B ja C ? Järjestämättömyys merkitse, että esim. pari AB katsotaan samaksi kuin pari BA ?

Ratkaisu: Kirjainjoukon järjestettyjen parien lukumäärä on $3 \cdot 2 = 6$. Koska nyt samat kirjaimet eri järjestyksessä sisältävät parit katsotaan samoiksi, on parien lukumäärä jaettava alkioiden keskinäisten järjestysten lukumäärällä eli permutaatioiden määrällä. Koska kaksi alkioita voidaan permutoida kahdella tavalla, on erilaisten järjestämättömien parien lukumäärä $6/2 = 3$.

Joukon, joka sisältää n alkioita, k -*kombinaatioiksi* sanotaan kaikkia järjestämättömiä k :n alkion osajoukkoja. Kaikkien k -kombinaatioiden lukumäärä siis kertoo, monellako eri tavalla joukon alkioista voidaan muodostaa k :n alkion osajoukko. Jos joukossa on n alkioita, sen kaikkien erilaisten k -kombinaatioiden lukumäärä on

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

Merkintä $\binom{n}{k}$ luetaan “ n yli k :n”.

1.3.2 Klassinen todennäköisyys

Olkoon tarkasteltavan satunnaisilmiön perusjoukossa n alkeistapahtumaa, jotka ovat *symmetrisiä* eli kaikki yhtä mahdollisia, ja olkoon A :lle suotuisten alkeistapahtumien lukumäärä k . Silloin tapahtuman A todennäköisyys on

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

Suotuisien tapahtumien lukumäärä on aina välillä $0 \leq k \leq n$, joten $0 \leq P(A) \leq 1$. Jos A :lle suotuisia alkeistapahtumia ei ole yhtään, on $P(A) = 0$, ja jos kaikki alkeistapahtumat ovat A :lle suotuisia, on $P(A) = 1$.

Klassisien todennäköisyyden määritelmä edellyttää, että perusjoukon alkioden lukumäärä on äärellinen ja että jokaisen alkeistapahtuman esiintymismahdollisuus on sama eli $1/n$. Tällaisissa tapauksissa klassinen todennäköisyys onkin käyttökelpoinen todennäköisyyksien laskemiseen.

Esimerkki 29: Heitetään arpakuutiota. Perusjoukko $E = \{1, 2, 3, 4, 5, 6\}$ sisältää 6 alkeistapahtumaa, ja jos arpakuutio on virheetön, jokaisen tuloksen tapahtumismahdollisuus on yhtä suuri. Tapahtuman $A = \text{“saadaan kuutonen”}$, eli $A = \{6\}$ todennäköisyys on siis $P(A) = 1/6$ ja samoin $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = 1/6$. Tapahtumalle $B = \text{“saadaan parillinen silmäluku”}$ eli $B = \{2, 4, 6\}$ suotuisia alkeistapahtumia on kolme, alkeistapahtumat $\{2\}$, $\{4\}$ ja $\{6\}$, joten $P(B) = 3/6 = 1/2$.

1.3.3 Aksiomaattinen todennäköisyys

Aksiomaattisessa lähestymistavassa todennäköisyys määritellään joukkofunktiona, joka liittää perusjoukon eli *otosavaruuden* jokaiseen tapahtumaan reaaliluvun, jota sanotaan ko. tapahtuman todennäköisyydeksi. Aksiomat edellyttävät todennäköisyysfunktiolta seuraavia ominaisuuksia:

1. $0 \leq P(A) \leq 1$ kaikille tapahtumille $A \subset E$
2. $P(E) = 1$
3. $P(A \cup B) = P(A) + P(B)$ kaikille tapahtumille $A, B \subset E$, joille $A \cap B = \emptyset$ (ts. *toisensa poissulkeville* eli *erillisille* tapahtumille)

Sekä klassinen että tilastollinen todennäköisyys toteuttavat nämä aksiomat. Aksiomista voidaan johtaa monia seurausominaisuuksia ja laskusääntöjä.

1. Mahdottoman tapahtuman todennäköisyys

Mahdottomalle tapahtumalle \emptyset ja mille hyvänsä tapahtumalle A on voimassa $A \cup \emptyset = A$ ja $A \cap \emptyset = \emptyset$. Aksiooman 3 nojalla on $P(A) = P(A \cup \emptyset) = P(A) + P(\emptyset)$, eli $P(\emptyset) = 0$. Saadaan siis sääntö

$$P(\emptyset) = 0$$

2. Komplementtitapahtuman todennäköisyys

Tapahtuma A ja sen komplementtitapahtuma A^* ovat toisensa poissulkevia, ts. $A \cap A^* = \emptyset$, ja lisäksi $A \cup A^* = E$. Aksioomien 2 ja 3 perusteella on $P(A) + P(A^*) = 1$, joten

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

3. Toisensa poissulkevien tapahtumien yhteenlaskusääntö

Aksiooman 3 ominaisuus voidaan yleistää koskemaan usean toisensa poissulkevan tapahtuman unionia. Jos tapahtumille A_1, A_2, \dots, A_k on voimassa $A_i \cap A_j = \emptyset$ aina, kun $i \neq j$, on voimassa

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k) = \sum_{i=1}^k P(A_i).$$

4. Yleinen yhteenlaskusääntö

Jos tapahtumat A ja B eivät ole toisensa poissulkevia, niin

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

Sääntö voidaan yleistää kuten aksiooma 3 koskemaan useampiakin tapauksia. Esimerkiksi kolmelle tapahtumalle A, B ja C on voimassa

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

Yhdistetyt satunnaisilmiöt

Monissa todennäköisyyyslaskennan ongelmissa voidaan käyttää hyväksi lähestymistapaa, jossa satunnaisilmiön ajatellaan muodostuvan kahden tai useamman osailmiön yhdisteenä. Esimerkiksi satunnaisilmiö “heitetään rahaa kaksi kertaa” voidaan ajatella yhdistetyksi ilmiöistä “heitetään rahaa ensimmäisen kerran” ja “heitetään rahaa toisen kerran”, ja tapahtuma, esimerkiksi “saadaan kaksi klaavaa”, muodostuu niinkään osailmiöiden tapahtumien (“saadaan klaava ensimmäisellä heitolla” ja “saadaan klaava toisella heitolla”) yhdistelmänä. Jos tapahtumien todennäköisyyksiä lasketaan tällä menettelyllä, on ensin tutkittava, ovatko osailmiöt toisistaan *riippuvia* vai *riippumattomia*.

Satunnaisilmiöt ovat riippumattomia, jos toisessa ilmiössä esiintyvä tulos ei millään lailla vaikuta siihen, mikä toisen ilmiön tulos tulee olemaan. Muutoin ilmiöt ovat toisistaan riippuvia. Edellä mainitussa rahanheitossa ensimmäinen ja toinen rahanheitto ovat toisistaan riippumattomat. Esimerkki riippuvista ilmiöistä on kahden kortin vetäminen peräkkäin korttipakasta (vedettyä korttia palauttamatta). Esimerkiksi tapahtuman “toinen kortti on pata” esiintymistodennäköisyys riippuu siitä, oliko ensiksi vedetty kortti pata vai ei.

Jos kaksi satunnaisilmiötä ovat toisistaan riippumattomia ja tapahtuma A liittyy ensimmäiseen ilmiöön sekä tapahtuma B toiseen ilmiöön, tapahtumat A ja B ovat toisistaan riippumattomia. Vastaavasti toisistaan riippuvien ilmiöiden yhteydessä puhutaan toisistaan riippuvista tapahtumista.

5. Riippumattomien tapahtumien kertolaskusääntö

Jos tapahtumat A ja B ovat toisistaan riippumattomia, niin todennäköisyys sille, että molemmat tapahtumat tapahtuvat, on

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

Esimerkki 30: Heitetään rahaa kaksi kertaa peräkkäin. Merkitään

A: saadaan ensimmäisellä heitolla klaava

B: saadaan toisella heitolla klaava.

Nyt $P(A) = P(B) = \frac{1}{2}$ ja todennäköisyys, että saadaan kummallakin heitolla klaava on

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Kertolaskusääntö voidaan yleistää koskemaan n :ää toisistaan riippumatonta tapahtumaa A_1, A_2, \dots, A_n :

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n).$$

Esimerkki 31: Heitetään rahaa kolme kertaa peräkkäin. Merkitään A_i : saadaan i :nnellä heitolla klaava, $i = 1, 2, 3$.

Todennäköisyys, että saadaan jokaisella heitolla klaava on

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

Ehdollinen todennäköisyys

Jos yhdistetyssä satunnaisilmiössä tapahtumat A ja B riippuvat toisistaan, ja tapahtuman A mahdollinen esiintyminen havaitaan ensin, tapahtuman B todennäköisyys riippuu siitä, onko A tapahtunut vai ei. Tämä ilmaistaan ehdollisen todennäköisyyden avulla. Merkintä

$P(B/A)$, luetaan “tapahtuman B ehdollinen todennäköisyys” tai “ B :n todennäköisyys ehdolla A ”

tarkoittaa B :n tapahtumisen todennäköisyyttä ehdolla, että A on tapahtunut. Jos $P(A) > 0$, tapahtuman B ehdollinen todennäköisyys on

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

Vastaavasti, jos $P(B) > 0$, on tapahtuman A ehdollinen todennäköisyys

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

Esimerkki 32: Erään tavaratalon alennusmyynissä oli sekaisessa laatikossa jäljellä naisten T-paitoja kahta väriä ja kahta kokoa seuraavat määrät:

	väri		
	valkoinen	musta	yht.
koko 42	12	8	20
36	9	17	26
yht.	21	25	46

Merkitään V: valkoinen paita
M: musta paita
L: koko 42
S: koko 36

Taulukosta voidaan laskea seuraavat todennäköisyydet:

	väri		
	V	M	yht.
L	$P(V \cap L) = \frac{12}{46}$	$P(M \cap L) = \frac{8}{46}$	$P(L) = \frac{20}{46}$
S	$P(V \cap S) = \frac{9}{46}$	$P(M \cap S) = \frac{17}{46}$	$P(S) = \frac{26}{46}$
yht.	$P(V) = \frac{21}{46}$	$P(M) = \frac{25}{46}$	1

Kokoa 42 oleva henkilö nostaa laatikosta mustan paidan. Millä todennäköisyydellä se on sopiva?

$P(\text{“saadaan kokoa 42 oleva paita”} \mid \text{“saatiin musta paita”}) = P(L \mid M) =$

$$\frac{P(L \cap M)}{P(M)} = \frac{P(M \cap L)}{P(M)} = \frac{8/46}{25/46} = \frac{8}{25}$$

Ehdolliset todennäköisyydet voidaan esittää taulukkona

	väri		
	V	M	yht.
L	$P(L \mid V) = \frac{12}{21}$	$P(L \mid M) = \frac{8}{25}$	$P(L) = \frac{20}{46}$
S	$P(S \mid V) = \frac{9}{21}$	$P(S \mid M) = \frac{17}{25}$	$P(S) = \frac{26}{46}$
yht.	1	1	1

6. Yleinen kertolaskusääntö

Kaikille tapahtumille soveltuva kertolaskusääntö, ns. yleinen kertolaskusääntö, seuraa suoraan ehdollisen todennäköisyyden määrittelystä:

$$P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B).$$

Tapahtumien A ja B riippumattomuus voidaan määritellä myös ehdollisten todennäköisyyksien kautta. Tällöin riippumattomuutta (ja riippuvuutta) ei tarvitse ajatella ainoastaan yhdistettyihin satunnaisilmiöihin vaan myös saman satunnaisilmiön eri tapahtumiin liittyvänä ominaisuutena:

Olkoot A ja B saman satunnaisilmiön tapahtumia. Tapahtumien A ja B sanotaan olevan riippumattomia, jos

$$P(A/B) = P(A).$$

Riippumattomien tapahtumien kertolaskusääntö saadaan nyt yleisen kertolaskusäännön erikoistapauksena. Jos A ja B ovat riippumattomia, on

$$P(A \cap B) = P(B) \cdot P(A/B) = P(B) \cdot P(A) = P(A) \cdot P(B).$$

Esimerkin 32 tapauksessa esim. $P(L|V) = \frac{12}{21} \approx 0.57$ ja $P(L) = \frac{20}{46} \approx 0.43$, eli

$P(L|V) \neq P(L)$, joten tapahtumat L ja V ovat riippuvia.

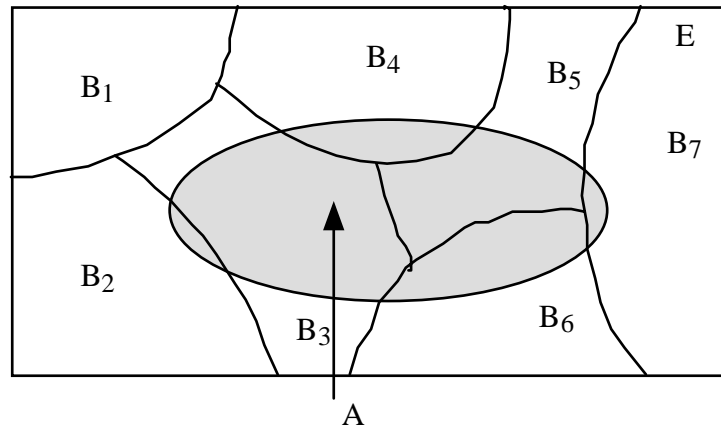
7. Kokonaistodennäköisyys ja Bayesin kaava

Seuraavassa esiteltävät tulokset koskevat tilannetta, jossa perusjoukko E on jaettu toisensa poissulkeviin tapahtumiin B_i , $i = 1, \dots, n$, ts.

$$B_i \cap B_j = \emptyset \text{ kaikilla } i \neq j, \text{ ja}$$

$$B_1 \cup B_2 \cup B_3 \cup \dots \cup B_n \equiv \bigcup_{i=1}^n B_i = E.$$

Lisäksi oletetaan tunnetuiksi todennäköisyydet $P(B_i) > 0$. Olkoon myös A tarkasteltavan satunnaisilmiön tapahtuma (ks. seuraava kuvio).



Joukko A voidaan nyt kirjoittaa muodossa

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n),$$

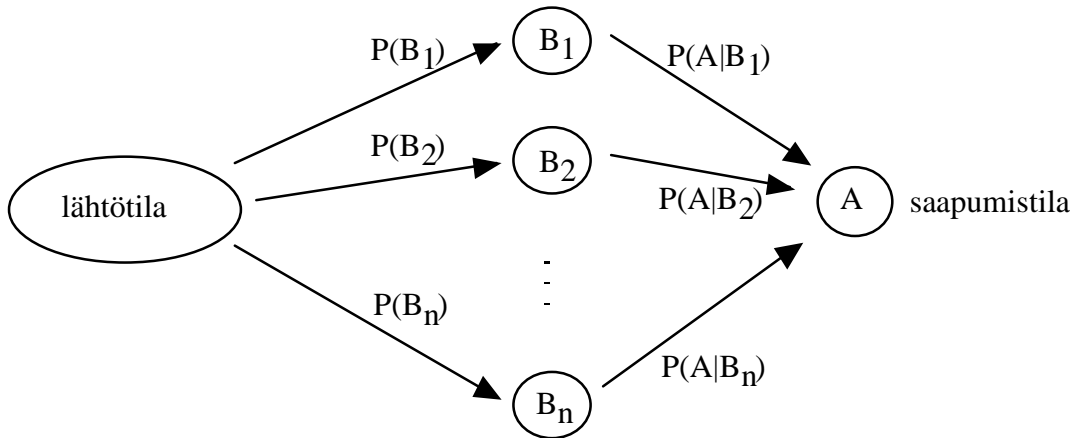
missä joukot $A \cap B_i$ ja $A \cap B_j$ ovat toisensa poissulkevia kaikilla $i \neq j$. Siis

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n).$$

Yleisen kertolaskusäännön mukaan $P(A \cap B_i) = P(B_i) \cdot P(A/B_i)$ kaikilla $i = 1, \dots, n$, joten

$$P(A) = P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) + \dots + P(B_n) \cdot P(A/B_n).$$

Tätä lauseketta sanotaan *kokonaistodennäköisyyden* kaavaksi. Kaavaa voidaan tulkita mm. seuraavasti: “Tiloilla” B_i on tunnetut todennäköisyydet $P(B_i)$, ja tilaan A päästään vain jonkin tilan B_i kautta. Ehdollinen todennäköisyys ilmaisee ns. siirtymätodennäköisyyden tilasta B_i tilaan A . Kokonaistodennäköisyys $P(A)$ on siirtymätodennäköisyyksien painotettu keskiarvo painojen ollessa luvut $P(B_i)$ (joiden summa on 1). Seuraava kaavio havainnollistaa tätä tulkintaa.



Kokonaistodennäköisyyden kaavaa käytetään laskettaessa tapahtuman A todennäköisyys, kun tunnetaan ne eri reitit, joiden kautta tilaan A päädytään. Jos halutaan vastaus käänteiseen kysymykseen eli halutaan tietää, millä todennäköisyydellä tilaan A on tultu tietyn tilan B_i kautta, voidaan apuna käyttää Bayesin (“käänteistodennäköisyyden”) kaavaa

$$\begin{aligned}
 P(B_i/A) &= \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i) \cdot P(A/B_i)}{P(B_1) \cdot P(A/B_1) + P(B_2) \cdot P(A/B_2) + \dots + P(B_n) \cdot P(A/B_n)} \\
 &= \frac{P(B_i) \cdot P(A|B_i)}{\sum_{j=1}^n P(B_j) \cdot P(A|B_j)}.
 \end{aligned}$$

1.4 Todennäköisyysjakaumat

Satunnaisilmiöitä voidaan kuvata todennäköisyysmallien avulla. Edellä käsitelty laskusäännöt soveltuvat yksittäisten tapahtumien todennäköisyyksien määrittämiseen. Jos halutaan tarkastella satunnaisilmiön ominaisuuksia kokonaisvaltaisesti, muodostetaan ilmiöstä todennäköisyyslaskennallinen malli. Mallin tarkastelussa voidaan käyttää apuvälineenä todennäköisyysteorian yleisiä tuloksia ja laskusääntöjä. Keskeisiä todennäköisyysmalleihin liittyviä käsitteitä ovat *satunnaismuuttuja* ja sen *todennäköisyysjakauma*.

1.4.1 Satunnaismuuttuja ja todennäköisyysjakauma

Satunnaismuuttuja liittyy tarkasteltavan satunnaisilmiön (eli satunnaiskokeen) alkeistapahtumiin numeerisia arvoja. Kahden arpakuution heitossa satunnaismuuttujaksi voidaan valita esimerkiksi kuutioiden yhteenlaskettu silmäluku. Tämä satunnaismuuttuja liittyy tapahtumaan “saadaan kaksi ykköstä” luvun 2, tapahtumaan “saadaan kakkonen ja ykkönen” luvun 3, jne., ja lopuksi tapahtumaan “saadaan kaksi kuutosta” luvun 12. Tarkemmin, satunnaismuuttuja on *funktio*, joka liittyy tarkasteltavan satunnaisilmiön perusjoukon jokaiseen tulokseen (alkeistapahtumaan) jonkin yksikäsitteisen reaalityluvun. Eri tuloksiin liittyviä lukuja sanotaan satunnaismuuttujan *arvoiksi*. Esimerkin satunnaismuuttujan mahdolliset arvot ovat 2, 3, . . . , 12. Tässä monisteessa merkitään satunnaismuuttujia isoilla kirjaimilla X, Y, Z , tms. ja muuttujan arvoja vastaavilla pienillä kirjaimilla x, y, z , jne.

Tarkasteltavaan satunnaisilmiöön voidaan useimmiten liittää useita erilaisia satunnaismuuttujia. Edellisessä esimerkissä satunnaismuuttujaksi voitaisiin valita myös vaikkapa kuutosten lukumäärä, jolloin satunnaismuuttujan mahdolliset arvot olisivat 0, 1 ja 2.

Satunnaismuuttujat jaetaan diskreetteihin ja jatkuviin sen mukaan, saavatko ne diskreettejä vai jatkuvia arvoja. Seuraavassa esimerkissä 1. ja 2. tapaus edustavat diskreettiä ja 3. jatkuvaa satunnaismuuttujaa.

Esimerkki 33: Esimerkkejä satunnaismuuttujista:

satunnaisilmiö	satunnaismuuttuja	satunnaismuuttujan mahdolliset arvot
tarkastetaan 100 kpl:n tuote-erä	viallisten tuotteiden lukumäärä	0, 1, 2, . . . , 100
tutkitaan tavara-taloa	asiakkaiden lukumäärä	0, 1, 2, . . .
tutkitaan tavara-talon asiakkaita	asiakkaan ostoksiin kuluttama aika	[0, a] a= aukioloaika

Satunnaismuuttujaan liittyvä satunnaisuus on peräisin itse satunnaisilmiöstä: satunnaismuuttujan arvoa ei voi varmuudella tietää ennen ilmiön tuloksen realisoitumista. Satunnaismuuttuja sinänsä, sääntönä, jolla tuloksiin liitetään lukuja, on täysin määrätty.

Satunnaismuuttujan eri arvojen todennäköisyydet määräytyvät vastaavista satunnaisilmiön tapahtumien todennäköisyyksistä. Satunnaismuuttujan arvot ja niihin liittyvät todennäköisyydet muodostavat satunnaismuuttujan todennäköisyysjakauman. Todennäköisyysjakaumaa voidaan ajatella *todennäköisyysmassana*, jonka arvo on yksi ja joka on jakaantunut satunnaismuuttujan eri arvojen kesken näiden arvojen todennäköisyyksien mukaisesti. Satunnaismuuttujan ja sen jakauman muodollinen käsittely jatkuvan ja diskreetin muuttujan kohdalla eroavat toisistaan, joten nämä muuttujatyypit esitellään jäljempänä erikseen.

Kertymäfunktio

Satunnaismuuttujan X *kertymäfunktio* F on funktio, joka liittää jokaiseen reaali-lukuun x todennäköisyyden, että muuttuja saa arvokseen x :n tai jonkin sitä pienemmän arvon. Muodollisesti kertymäfunktio kirjoitetaan

$$F(x) = P(X \leq x) , \text{ kaikilla reaaliluvuilla } x.$$

Kertymäfunktioilla on seuraavat ominaisuudet:

1. $P(X \leq a) = F(a)$
2. $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
3. $P(a < X \leq b) = F(b) - F(a)$.

Kertymäfunktion määrittely ja yllämainitut ominaisuudet ovat voimassa minkä tyypiselle satunnaismuuttujalle hyvänsä. Kertymäfunktio määrää täysin muuttujan X todennäköisyysjakauman, mutta ei auta yksikäsitteisesti spesifioimaan muuttujaa X , sillä kahdella eri muuttujalla voi olla sama kertymäfunktio.

Diskreetit satunnaismuuttujat

Diskreetti satunnaismuuttuja X voi saada vain tiettyjä erillisiä arvoja, joita tässä merkitään (suuruusjärjestyksessä) x_1, x_2, x_3, \dots , ja joiden todennäköisyydet ovat $P(X=x_1), P(X=x_2), P(X=x_3), \dots$. Diskreetin muuttujan X todennäköisyysjakauma voidaan ilmoittaa *pistetodennäköisyysfunktion* $p(x)$ avulla:

$$p(x) = \begin{cases} P(X = x), & \text{kun } x \in \{x_1, x_2, \dots\} \\ 0, & \text{muulloin} \end{cases}$$

Todennäköisyysjakauman massa keskittyy diskreetin muuttujan tapauksessa pisteisiin x_i , ja näiden massojen suuruuden ilmoittavat todennäköisyydet $p_i \equiv p(x_i) = P(X=x_i)$. Pistetodennäköisyysfunktiolle on siis oltava voimassa $\sum_i p_i \equiv \sum_i p(x_i) = 1$, missä summaus suoritetaan kaikkien mahdollisten X :n arvojen yli. Pistetodennäköisyysfunktiota voi havainnollistaa graafisesti janadiagrammin avulla.

Diskreetin satunnaismuuttujan kertymäfunktion arvot $F(x)$ saadaan laskemalla yhteen todennäköisyydet $p_i = P(X=x_i)$ alkaen pienimmästä arvosta x_1 aina siihen arvoon x_k , jolle viimeiseksi on voimassa $x_k \leq x$. Siis

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p_i.$$

Kertymäfunktion kuvaaja on tässä tapauksessa porraskuvio, joka on vakio niillä väleillä, joihin ei kuulu pisteitä x_i ja joka tekee jokaisen arvon x_i kohdalla todennäköisyyden p_i suuruisen hyppäyksen.

Esimerkki 34: Heitetään rahaa kaksi kertaa ja olkoon X = klaavojen lukumäärä. Alkeistapahtumat ja niihin liittyvät satunnaismuuttujan X arvot ovat $X(kr,kr) = 0$, $X(kr,kl) = 1$, $X(kl,kr) = 1$, $X(kl,kl) = 2$.

Vastaavat todennäköisyydet ovat siis

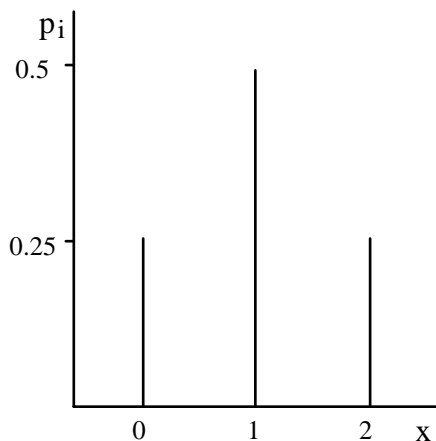
$$p_1 = P(X = 0) = 1/4$$

$$p_2 = P(X = 1) = 1/2$$

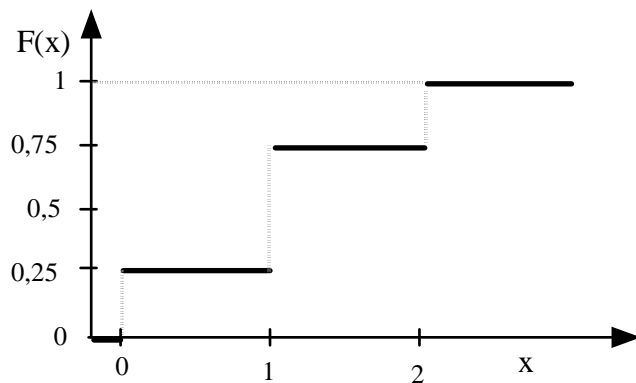
$$p_3 = P(X = 2) = 1/4$$

ja kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.25, & 0 \leq x < 1 \\ 0.75, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$



Satunnaismuuttujan X
pistetodennäköisyysfunktio

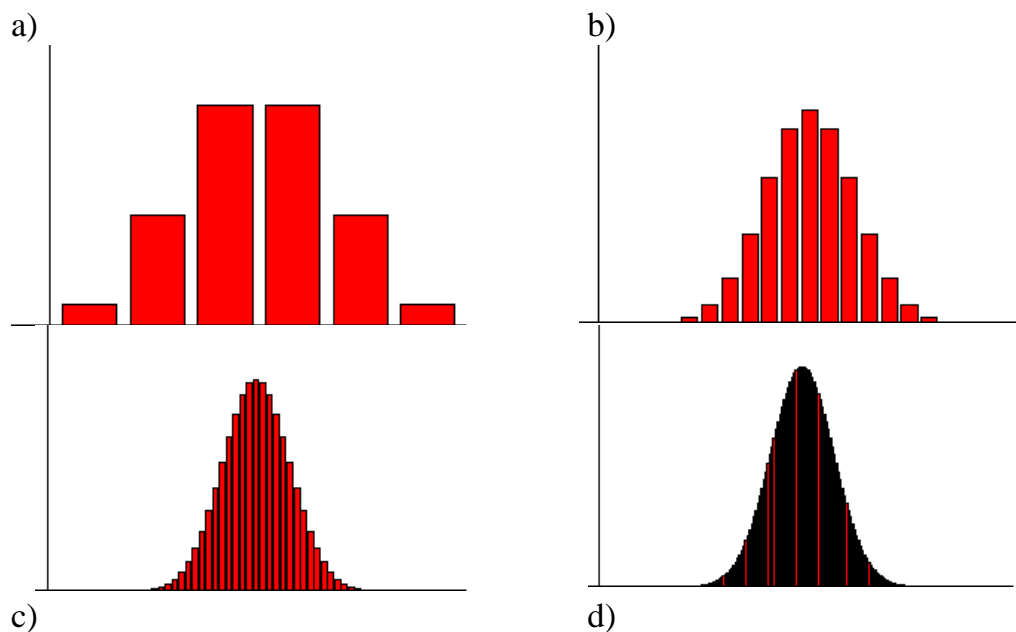


Satunnaismuuttujan X
kertymäfunktio

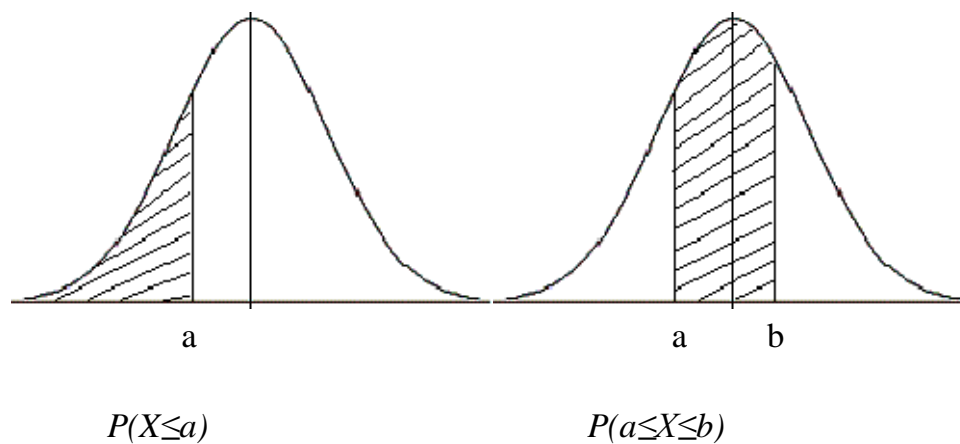
Jatkuvat satunnaismuuttujat

Jatkuva satunnaismuuttuja voi saada joltain tietyltä reaalilukuväliltä periaatteessa minkä hyvänsä arvon. Koska muuttujan arvoja on tällöin ääretön määrä, yksittäisen arvon $X = a$ todennäköisyydeksi tulee aina nolla, ts. $P(X = a) = 0$. Jatkuvan satunnaismuuttujan jakaumaa ei siis ole mielekäästä kuvata samoin kuin diskreetin muuttujan.

Jatkuvan satunnaismuuttujan todennäköisyysjakauman antaa jatkuva tai ainakin paloittain jatkuva funktio, jota sanotaan *tiheysfunktio*ksi. Tiheysfunktion voidaan ajatella muodostuvan seuraavasti. Tarkastellaan luokitellun, jatkuvan muuttujan X histogrammia, joka on piirretty siten, että kunkin pylvään pinta-ala vastaa kyseisen X :n luokan todennäköisyyttä (kuvio 35 a). Kun luokkajakoa tihennetään äärettömän tiheäksi (kuvio 35 b-d), saadaan lopulta X :n tiheysfunktio (kuvio 35 d). Jakauman todennäköisyysmassaa edustaa pinta-ala, joka jää muodostuvan käyrän ja vaaka-akselin väliin, ja joka on suuruudeltaan yksi.

Kuvio 35: *Tiheysfunktion “muodostuminen”.*

Tiheysfunktion avulla voidaan määrittää todennäköisyys sille, että satunnaismuuttujan X arvo kuuluu jollekin tietylle arvovälille. Näitä todennäköisyyksiä edustaa tiheysfunktion kuvaajan, vaaka-akselin sekä ko. arvovälin ala- ja/tai ylärajan rajoittaman alueen pinta-ala. Tätä havainnollistaa seuraava kuvio.



Pinta-alan täsmällinen suuruus voidaan laskea integroimalla tiheysfunktio tarkasteltavan arvoalueen yli, esimerkiksi

$$P(a \leq X \leq b) = \int_a^b f(x) dx.$$

Jatkuvan satunnaismuuttujan kertymäfunktio määritellään samoin kuin diskreetin muuttujan tapauksessa, ts. $F(x) = P(X \leq x)$. Siis esimerkiksi

$$F(a) = P(X \leq a) = \int_{-\infty}^a f(x) dx.$$

Tiheysfunktio on siis kertymäfunktion derivaatafunktio,

$$f(x) = \frac{d}{dx} F(x).$$

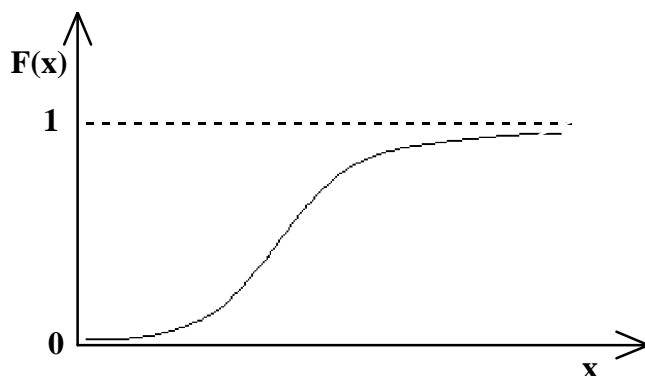
Koska X :n yksittäisen arvon todennäköisyys oli nolla, on myös

$$P(X < a) = F(a)$$

ja samoin esimerkiksi

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

Kertymäfunktion kuvaaja on nouseva käyrä, joka saa arvot 0:sta 1:een:



Tiheysfunktion ominaisuuksista voidaan mainita mm.

$$\text{i) } f(x) \geq 0$$

ja

$$\text{ii) } \int_{-\infty}^{\infty} f(x) dx = 1.$$

On huomattava, että tiheysfunktion arvo pisteessä a ei edusta todennäköisyyttä, että $X=a$, kuten pistetodennäköisyysfunktion tapauksessa.

Jakauman tunnusluvut

Kuten empiiristen jakaumien tapauksessa, todennäköisyysjakaumiakin voidaan luonnehtia tiivistetysti tunnuslukujen avulla. Tunnusluvuista tärkeimmät ovat *odotusarvo* (merk. E) ja *varianssi* (merk. D^2 tai Var).

Diskreetin muuttujan (ja jakauman) odotusarvo vastaa empiirisen jakauman keskiarvoa. Odotusarvo diskreetille satunnaismuuttujalle määritellään seuraavasti.

Olkoon X diskreetti satunnaismuuttuja, jonka mahdolliset arvot ovat x_1, x_2, \dots joiden todennäköisyydet ovat $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$, \dots . Satunnaismuuttujan X odotusarvo EX on silloin

$$EX = \sum_i p_i x_i = p_1 x_1 + p_2 x_2 + \dots$$

Odotusarvo voidaan tulkita muuttujan arvojen x_1, x_2, \dots painotetuksi keskiarvoksi, kun painot ovat todennäköisyydet p_1, p_2, \dots

Jatkuvan satunnaismuuttujan X odotusarvo on

$$EX = \int_{-\infty}^{\infty} x \cdot f(x) dx,$$

missä $f(x)$ on X :n tiheysfunktio.

Diskreetin satunnaismuuttujan varianssi vastaa niinkään empiirisen jakauman varianssia. Diskreetin satunnaismuuttujan varianssi $D^2 X$ (tai $Var(X)$) määritellään kaavalla

$$D^2 X = E[(X - EX)^2] = \sum_i p_i (x_i - EX)^2$$

eli se on X :n arvojen odotusarvosta laskettujen poikkeamien neliöiden odotusarvo. Varianssi voidaan laskea myös kaavalla

$$D^2 X = E(X^2) - (EX)^2, \text{ missä } E(X^2) = \sum_i p_i x_i^2.$$

Jatkuvan satunnaismuuttujan varianssi määritellään kaavalla

$$D^2 X = \int_{-\infty}^{\infty} (x - EX)^2 \cdot f(x) dx.$$

Jos satunnaismuuttujan varianssi tunnetaan, sen keskihajonta D voidaan laskea kaavalla

$$DX = \sqrt{D^2 X}.$$

Esimerkki 36: Heitetään arpakuutiota ja määritellään satunnaismuuttuja X = saatu silmäluku. Nyt X saa arvot $1, 2, \dots, 6$, ja jokaiseen arvoon liittyvä todennäköisyys on $1/6$. Satunnaismuuttujan X odotusarvo on siis

$$EX = \frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) = 3.5,$$

varienssi

$$D^2 X = \frac{1}{6} [(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2] = 2.92$$

ja keskihajonta

$$DX = \sqrt{2.92} = 1.7.$$

Satunnaismuuttujien muunnoksista

Jos X on satunnaismuuttuja ja g on jokin funktio reaalilukujen joukolta reaalilukujen joukolle, on $Y = g(X)$ myös satunnaismuuttuja. Merkintä tarkoittaa, että Y saa arvon $y = g(x)$ silloin kun X saa arvon x .

Eräs yksinkertaisimmista muunnoksista on lineaarinen muunnos $Y = a + bX$, missä a ja b ovat reaalilukuvakioita. Esimerkiksi jos X on satunnaisesti valitun henkilön pituus metreinä, voidaan määritellä lineaarinen muunnos $Y = 100 X$, eli ko. henkilön pituus senttimetreinä.

Muunnoksien tunnuslukujen laskemisessa voidaan käyttää apuna mm. seuraavia yleisiä laskusääntöjä.

Olkoon a ja b reaalilukuvakioita sekä X ja Y satunnaismuuttujia. Silloin

- 1) $E(a) = a$ ja $D^2(a) = 0$
- 2) $E(a + bX) = a + bEX$ ja $D^2(a + bX) = b^2D^2X$
- 3) $E(X+Y) = EX + EY$
- 4) jos satunnaismuuttujat X ja Y ovat keskenään riippumattomia, niin $D^2(X + Y) = D^2X + D^2Y$.

Standardointimuunnos on eräs lineaarisen muunnoksen erikoistapaus, joka saadaan valitsemalla $a = -EX/DX$ ja $b = 1/DX$. Jos merkitään satunnaismuuttujan X standardoitua muuttujaa Z :lla, on

$$Z = \frac{X - EX}{DX}.$$

Standardoidun muuttujan odotusarvo on aina $EZ = 0$ ja varianssi $D^2Z = 1$.

1.4.2 Diskreettejä todennäköisyysjakaumia

Edellä käsiteltyjen esimerkkien yhteydessä nähtiin, että diskreetti jakauma voidaan määritellä luettelemalla kaikki satunnaismuuttujan arvot ja niihin liittyvät todennäköisyydet. Jos kuitenkin muuttujan arvoja on hyvin suuri tai jopa ääretön määrä, ei tällainen menettely käy päinsä. Yleisempi tapa määritellä todennäköisyysjakauma on esittää yleinen sääntö, jonka avulla liitetään toisiinsa muuttujan arvot ja niiden todennäköisyydet. Diskreetin jakauman tapauksessa muodostetaan pistetodennäköisyysfunktion ja jatkuvan jakauman tapauksessa tiheysfunktion lauseke.

Todennäköisyyslaskennassa on muodostettu eräitä yleisiä todennäköisyysjakaumia, jotka soveltuvat useisiin käytännön ongelmiin. Tällainen jakauma esitetään lausekkeena, joka sisältää yhden tai useamman tuntemattoman vakion, ns. *parametrin*. Parametrien arvot määräytyvät tarkasteltavasta tapauksesta riippuen.

Keskeisimpiä diskreettejä jakaumia ovat *binomijakauma* ja *Poisson-jakauma*. Näiden lisäksi seuraavassa esitellään myös *hypergeometrinen* jakauma.

Binomijakauma

Binomijakaumaa voidaan käyttää sellaisten satunnaisilmiöiden kohdalla, joissa ilmiö toistuu tai toistetaan n kertaa ja kustakin toistosta havaitaan, esiintyykö tapahtuma A vai ei. Lisäksi oletetaan, että tapahtuman A todennäköisyys on sama jokaisessa toistossa ja kysytään, mikä on todennäköisyys sille, että A esiintyy tasan k kertaa kun ilmiö toistuu n kertaa.

Esimerkki 37: Heitetään arpakuutiota 3 kertaa ($n=3$). Millä todennäköisyydellä saadaan tasan kaksi kertaa kuutonen ($k=2$)?

Määritellään satunnaismuuttuja X = kuutosten määrä kolmessa heitossa. Koska heitot ovat toisistaan riippumattomia, on esim. heittosarjan tuloksen “saadaan ensin kuusi, sitten kuusi ja sitten jokin muu kuin kuusi” todennäköisyys $\left(\frac{1}{6}\right)^2 \cdot \frac{5}{6}$.

Erilaisia mahdollisuuksia saada kaksi kuutosta kolmen heiton sarjassa on kombinatoriikan sääntöjen mukaan $\binom{n}{k} = \binom{3}{2}$, joten

$$P(X = 2) = \binom{3}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \frac{5}{6}.$$

Esimerkin päättely voidaan yleistää seuraavasti:

Tarkastellaan satunnaisilmiötä, jolla on tulosmahdollisuudet A ja A^* . Olkoon A :n todennäköisyys $P(A) = p$, jolloin $P(A^*) = 1 - p$. Oletetaan, että ilmiö toistuu n riippumattomasti n kertaa, ja määritellään satunnaismuuttuja X = A :n esiintymisten lukumäärä n toistossa. Silloin todennäköisyys, että A esiintyy tasan k kertaa, eli $P(X=k)$, on

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

Sanotaan, että muuttujan X jakauma on binomijakauma parametreina n ja p , ja merkitään $X \sim \text{Bin}(n, p)$.

Binomijakautuneen satunnaismuuttujan kertymäfunktio on

$$F(k) = P(X \leq k) = \sum_{i=0}^k \binom{n}{i} \cdot p^i \cdot (1 - p)^{n-i},$$

ja sen odotusarvo on

$$EX = np$$

ja varianssi on

$$D^2X = np(1-p).$$

Hypergeometrinen jakauma

Binomijakauman tapauksessa oli oleellista, että tapahtuman A todennäköisyys pysyi samana jokaisella toistokerralla. Binomijakauman kuvaama tilanne voi esiintyä otannan yhteydessä, jos poimittu yksikkö palautetaan havaintojen teon jälkeen takaisin perusjoukkoon. Jos sama otanta suoritetaan siten, että poimitut yksiköt jätetään palauttamatta, käytetään ilmiön kuvaamiseen hypergeometrista jakaumaa.

Esimerkki 38: (*Helenius s.237*) Olkoon 15 tuotteen joukossa 10 virheetöntä ja 5 virheellistä tuotetta. Valitaan tästä joukosta satunnaisesti 3 tuotetta

- i) palauttamalla poimittu tuote takaisin ennen seuraavan valintaa
- ii) palauttamatta tuotetta

ja määritetään todennäköisyys, että valittujen tuotteiden joukossa on korkeintaan yksi virheellinen. Määritellään satunnaismuuttuja X = virheellisten lukumäärä kolmen tuotteen joukossa.

Tapauksessa i) X on binomijakautunut, parametreina $n = 3$ ja $p = \frac{5}{15} = \frac{1}{3}$. Siis

$$P(\text{korkeintaan yksi virheellinen}) = P(X \leq 1) = \binom{3}{0} \left(\frac{1}{3}\right)^0 \cdot \left(\frac{2}{3}\right)^3 + \binom{3}{1} \left(\frac{1}{3}\right)^1 \cdot \left(\frac{2}{3}\right)^2 \approx 0.741.$$

Tapauksessa ii) X noudattaa hypergeometrista jakaumaa. Tapahtuman $\{X=1\}$ todennäköisyys lasketaan klassisen todennäköisyyden avulla seuraavasti. Alkeistapauksina ovat 15 tuotteen joukosta valitut 3:n tuotteen joukot, joita on

$$\binom{15}{3}$$

kappaletta. Tapahtumalle suotuisat alkeistapaukset ovat ne 3 tuotteen joukot, joissa on 1 virheellinen ja 2 virheetöntä. Tällaiset joukot voidaan ajatella muodostuneeksi kahdessa vaiheessa: Ensin poimitaan 1 virheellinen kaikkien viiden

virheellisen joukosta, mikä voidaan tehdä $\binom{5}{1}$ tavalla. Sitten poimitaan $3-1 = 2$

virheetöntä loppujen $15-5 = 10$ virheettömän joukosta, mikä puolestaan voidaan tehdä $\binom{15-5}{3-1} = \binom{10}{2}$ tavalla. Tulosäännön mukaan on siis suotuisia tapauksia yhteensä $\binom{5}{1} \cdot \binom{10}{2}$ kpl ja todennäköisyys $P(X=1)$ on

$$P(X=1) = \frac{\binom{5}{1} \binom{10}{2}}{\binom{15}{3}} = \frac{225}{455}.$$

Todennäköisyys $P(X=0)$ on vastaavasti

$$P(X=0) = \frac{\binom{5}{0} \binom{10}{3}}{\binom{15}{3}} = \frac{120}{455},$$

joten

$$P(X \leq 1) = P(X=0) + P(X=1) = \frac{120 + 225}{455} \approx 0.758.$$

Edellinen voidaan yleistää seuraavasti. Olkoon N = populaation koko, n = otoskoko, K = virheellisten tuotteiden lukumäärä populaatiossa ja k = virheellisten tuotteiden lukumäärä otoksessa. Silloin

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, 2, \dots, n.$$

Jos satunnaismuuttujan pistetodennäköisyysfunktio on tätä muotoa, sanotaan, että X noudattaa hypergeometrista jakaumaa parametrein N , K ja n , ja merkitään $X \sim \text{Hyperg}(N, K, n)$.

Hypergeometrista jakaumaa noudattavan satunnaismuuttujan X odotusarvo ja varianssi ovat

$$EX = \frac{nK}{N} \text{ ja } D^2X = \frac{nK}{N} \left(1 - \frac{K}{N}\right) \frac{N-n}{N-1}.$$

Poisson-jakauma

Poisson-jakaumaa sovelletaan binomijakaumaa vastaavissa tilanteissa, joissa toistojen lukumäärä on hyvin suuri ja tarkasteltavan tapahtuman esiintymistodennäköisyys pieni, esimerkiksi liikenneonnettomuuksien lukumäärä tietyllä aikavälillä. Poisson-jakaumaa siis käytetään harvinaisten tapahtumien todennäköisyyksien laskemiseen. Suurilla toistokertojen määrillä binomijakauman ja Poisson-jakauman antamat todennäköisyydet vastaavat likimain toisiaan. Poisson-jakaumaa käytetäänkin usein binomitodennäköisyyksien likimääräiseen määrittämiseen.

Olkoon X diskreetti satunnaismuuttuja, jonka mahdolliset arvot ovat $0, 1, 2, \dots$. Jos X :n pistetodennäköisyysfunktio on muotoa

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

sanotaan, että X on Poisson-jakautunut parametrina λ , ja merkitään $X \sim \text{Po}(\lambda)$.

Tarkasteltavassa ongelmassa X voi olla esimerkiksi harvinaisen tapahtuman esiintymiskertojen lukumäärä tietyn ajanjakson aikana, jolloin parametri λ edustaa tapahtuman keskimääräistä esiintymisfrekvenssiä kyseisellä ajanjaksolla. Tällöin edellytetään, että

- 1) Tapahtuman esiintyminen tietyllä ajanjaksolla on riippumaton sen esiintymisestä millä tahansa toisella ajanjaksolla.
- 2) Tapahtuman esiintymistaajuutta eli esiintymisten lukumäärää aikayksikössä voidaan pitää vakiona, joka ei muutu ajan kuluessa.

Esimerkki 39: (*Helenius s. 241*) Puhelinlaitokselle tulee satunnaisesti vikailmoituksia, keskimäärin kolme vikailmoitusta viikossa. Millä todennäköisyydellä viikon aikana tulee täsmälleen yksi vikailmoitus?

X = vikailmoitusten määrä viikon aikana, $X \sim \text{Po}(3)$.

$$P(X=1) = \frac{3^1 e^{-3}}{1!} = 3e^{-3} \approx 0.149.$$

Poisson-jakautuneen satunnaismuuttujan X odotusarvo ja varianssi ovat

$$EX = \lambda \text{ ja } D^2X = \lambda.$$

1.4.3 Jatkuvia todennäköisyysjakaumia

Tärkein jatkuva jakauma on *normaalijakauma*. Muista jatkuvista jakaumista tässä esitellään *tasainen (tasa-) jakauma*.

Tasainen jakauma

Tarkastellaan satunnaismuuttujaa X , joka voi saada minkä hyvänsä reaalilukuarvon väliltä $[a, b]$, eli jonka arvojoukko on väli $[a, b]$. Jos tässä joukossa vallitsee symmetria, siten että jokaisen samanpituisen osavälin todennäköisyys on yhtä suuri, on X tasaisesti jakautunut välillä $[a, b]$.

Todennäköisyys, että X saa arvon osaväliltä $[a, c]$, $a \leq c \leq b$, on osavälin pituuden suhde koko arvovälin $[a, b]$ pituuteen, eli $P(a \leq X \leq c) = \frac{c-a}{b-a}$. Lisäksi huoma-

mataan, että $P(X < c) = 0$ aina, kun $c < a$ ja $P(X < c) = 1$ aina, kun $c > b$. Tasaisesti jakautuneen satunnaismuuttujan kertymäfunktioksi saadaan siis

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

Tasaisesti jakautuneen satunnaismuuttujan jakaumassa todennäköisyysmassa on tasaisesti levinnyt yli koko arvojoukon, mistä seuraa, että muuttujan tiheysfunktio on vakio. Tiheysfunktio $f(x)$ saadaan kertymäfunktioista $F(x)$ derivoimalla, joten

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x > b \end{cases}$$

Tasaisesti jakautuneen satunnaismuuttujan odotusarvo on

$$EX = \frac{b+a}{2}$$

ja varianssi on

$$D^2X = \frac{(b-a)^2}{12}.$$

Esimerkki 40: Valitaan sattumanvaraisesti piste väliltä $[1,4]$. Millä todennäköisyydellä valittu piste sijaitsee 2:n ja 3:n välissä?

Merkitään X = pisteen etäisyys arvovälin alarajasta eli 1:stä. X :n kertymäfunktio on

$$F(x) = \begin{cases} 0, & x < 1 \\ \frac{x-1}{3}, & 1 \leq x \leq 4 \\ 1, & x > 4 \end{cases}$$

$$\text{ja todennäköisyys } P(2 \leq X \leq 3) = F(3) - F(2) = \frac{3-1}{3} - \frac{2-1}{3} = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}.$$

Normaalijakauma

Normaalijakaumaa kutsutaan toisen kehittäjänsä mukaan myös Gaussin jakaumaksi. Normaalijakauma soveltuu monien empiiristen ilmiöiden kuvaamiseen ja on jakaumista ehkä käytetyin. Se on myös tilastotieteen tärkein jakauma; mm. monien tilastollisten päättelymenetelmien yhteydessä lähtökohtana on normaalijakaumaoletus. Normaalijakaumaa voidaan lisäksi käyttää diskreettien muuttujien jakaumien approksimointiin.

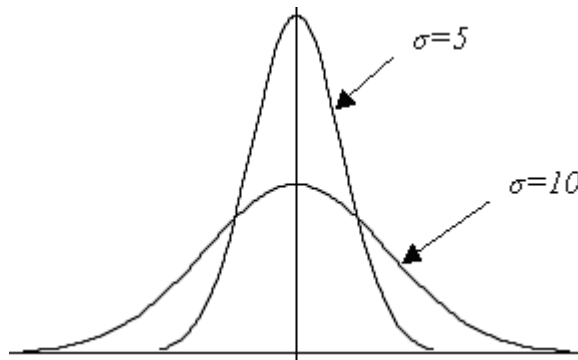
Olkoon X satunnaismuuttuja, joka voi saada kaikki reaalilukuarvot. Sanotaan, että X on normaalisti jakautunut parametrein μ ja σ^2 , merkitään $X \sim N(\mu, \sigma^2)$, jos sen tiheysfunktio on muotoa

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

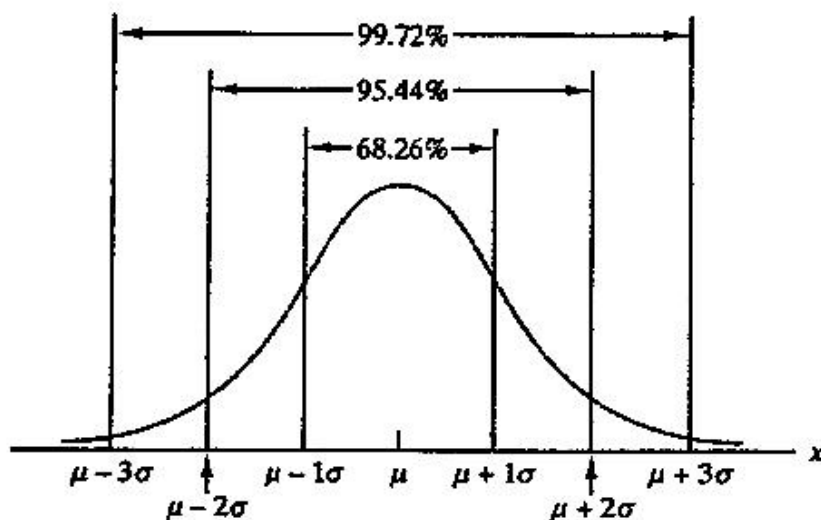
Normaalisti jakautuneen satunnaismuuttujan odotusarvo ja varianssi ovat

$$EX = \mu \text{ ja } D^2X = \sigma^2.$$

Nämä määräävät jakauman sijainnin ja muodon. Normaalijakauman tiheysfunktion kuvaajan huippu sijaitsee keskiarvon μ kohdalla. Kuvaaja on muodoltaan symmetrinen: keskiarvon vasemmalle puolelle jäävä käyrän osa on oikean puolen peilikuva. Hajonta σ määrää, kuinka leveälle satunnaismuuttujan arvot ovat hajonneet. Mitä suurempi σ , sitä litteämpi ja leveämpi on tiheysfunktion kuvaajan muoto.



Todennäköisyyden keskittyminen tietyn hajonnanmitan etäisyydelle keskiarvon ympärille on sama kaikille normaalijakaumille parametrien arvosta riippumatta. Seuraava kuvio esittää eräiden välien todennäköisyydet (edellinen ja seuraava kuvio *Anderson & al* s.185).



Normaalijakaumaan liittyvät todennäköisyydet lasketaan, kuten muillakin jatkuvilla jakaumilla, kertymäfunktion avulla. Kertymäfunktion lauseketta ei kuitenkaan voida muodostaa, koska tiheysfunktion integraalia ei pystytä määrittämään muutoin kuin numeerisesti. Useimmissa tilasto-ohjelmissa ja joissain laskimissa on toiminnot normaalijakauman kertymäfunktion laskemiseksi. Arvoja on myös taulukoitu jakaumalle $N(0,1)$ eli *standardoidulle* normaalijakaumalle.

Parametrein μ ja σ^2 normaalisti jakautunut satunnaismuuttuja on helppo muuntaa $N(0,1)$ -jakautuneeksi standardointimuunnoksen avulla. Jos $X \sim N(\mu, \sigma^2)$, sen standardoitu muuttuja

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

on normaalisti jakautunut parametrein 0 ja 1, eli $Z \sim N(0,1)$. Standardoidun normaalijakauman kertymäfunktiota merkitään $\Phi(z)$ ja tiheysfunktiota $\phi(z)$,

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

Koska jakauma on symmetrinen, kertymäfunktion arvojen taulukointi on suoritettu vain positiivisille Z :n arvoille. Symmetriasta seuraa, että $\Phi(0) = 0.5$ ja $\Phi(-z) = 1 - \Phi(z)$, joten kertymäfunktion arvot negatiivisilla Z :n arvoilla voidaan laskea vastaavan positiivisen arvon avulla. Esimerkiksi $\Phi(-0.5) = 1 - \Phi(0.5)$.

2 TILASTOLLINEN PÄÄTTELY

2.1 Johdanto

Monisteen alussa tutustuttiin tilastollisen tutkimuksen lähtökohtiin. Tutkimuksen kohteena on tilastoyksiköistä muodostuva perusjoukko eli populaatio ja tutkimuksessa tarkastellaan tilastoyksiköihin liittyviä ominaisuuksia, tilastollisia muuttujia. Jos tutkitaan kaikki populaatioon kuuluvat yksiköt, on kyseessä kokonaistutkimus. Tavallisempaa kuitenkin on otantatutkimuksen suorittaminen. Tällöin populaatiosta poimitaan jollakin otantamenetelmällä satunnaisotos, tutkitaan otokseen kuuluvat tilastoyksiköt ja yleistetään tutkimuksen tulokset koskemaan koko perusjoukkoa. Tämänkaltaisen menettely edustaa ns. *induktiivista päättelyä*.

Koska otos muodostaa vain suppean osan populaatiosta, otoksen perusteella tehtäviin johtopäätöksiin sisältyy aina epävarmuutta. Tilastollisessa päättelyssä tutkitaankin tätä epävarmuutta ja pyritään arvioimaan sen vaikutusta yleistyksiin. Tässä voidaan käyttää tehokkaasti hyväksi todennäköisyyslaskennan menetelmiä ja tuloksia, kunhan ongelma-alueesta on ensin muodostettu sopiva matemaattinen malli.

Tilastollisen päättelyn taustalla olevat mallit eli *tilastolliset mallit* muodostuvat todennäköisyysjakaumista. Malleihin liittyy väistämättä erilaisia oletuksia, joiden voimassaolo sovellustilanteessa on tärkeää. Käytännön tutkimustilanteessa valitaan sellainen malli, jonka oletuksia tutkittava ilmiö parhaiten vastaa.

Tilastollinen päättely voidaan jakaa kahteen toisiaan täydentävään lähestymistapaan. Nämä ovat (1) *estimointi* ja (2) *hypoteesien testaus* eli *tilastollinen merkitsevyystestaus*.

Estimoinnissa on kyse perusjoukon tuntemattomien parametrien arvioimisesta eli estimoinnista otoksen antaman informaation perusteella. Estimointi voidaan jakaa *piste-estimointiin* ja *väliestimointiin*.

Tilastollisen merkitsevyystestauksen lähtökohtana on jokin populaatiota koskeva ennakko-oletus eli *hypoteesi*, joka voi olla esimerkiksi väittämä populaation tunnusluvun suuruudesta. Testauksessa selvitetään, tukeeko otoksesta saatu havaintoaineisto kyseistä hypoteesia. Testin tulosten nojalla hypoteesi sitten hylätään tai hyväksytään.

2.2 Estimointi

Estimoinnissa pyritään otoksen perusteella antamaan mahdollisimman hyvä arvio tutkimuksen kohteena olevalle populaation parametrille. Parametri on usein jonkin populaatiosta mitattavan ominaisuuden empiirisen jakauman tunnusluku, kuten keskiarvo, mutta se voi olla jokin muukin suure, esim. tietyn ominaisuuden omaavien yksiköiden suhteellinen osuus populaatiossa. Tässä monisteessa estimoitavaa parametria sanotaan yleisesti tunnusluvuksi ja merkitään symbolilla T .

Estimoinnin lähtökohtana on populaatiosta poimittu satunnaisotos. Mittausten avulla otoksesta hankitaan tarvittava informaatio, jota sitten käytetään populaation tunnusluvun arviointiin. Satunnaisotannassa sattuma määrää otokseen poimittavat tilastoyksiköt, joten samasta populaatiosta poimitut eri otokset antavat käytännössä aina hiukan toisistaan poikkeavan datan. Estimointiteorian tehtävänä onkin selvittää, millä tavoin sattuman vaikutus ilmenee otoksen perusteella tehdyissä arvioissa. Tästä voidaan sitten päätellä mm., miten luotettavia arviot ovat.

Jos tutkittavaa tilastollista muuttujaa merkitään X :llä ja otoksen koko on n , niin havaintoja, eli otoksesta mitattuja muuttujan X arvoja merkitään x_1, x_2, \dots, x_n . Esimerkiksi, jos tutkitaan tietynikäisten, vaikkapa 25-60-vuotiaiden, suomalaisten pituutta, populaationa on kyseisen ikäiset suomalaiset ja muuttujana X henkilön pituus. Arvot x_1, x_2, \dots, x_n saadaan mittaamalla otokseen poimitujen n henkilön pituudet; esim. $x_1 = 178$ cm, $x_2 = 166$ cm, $\dots, x_n = 181$ cm.

Estimointiteoria käsittelee otoksesta saatavia havaintoja satunnaismuuttujina. Estimointiteorian näkökulmasta poimittavaan n kappaleen otokseen liittyy n satunnaismuuttujaa X_1, X_2, \dots, X_n . Edellisen esimerkin tapauksessa satunnaismuuttujat ovat $X_1 =$ otokseen ensimmäiseksi poimitun henkilön pituus, $X_2 =$ otokseen toiseksi poimitun henkilön pituus, jne. Kyse on todellakin satunnaismuuttujasta, sillä jokaisen muuttujan arvo määräytyy otannassa satunnaisesti. Kun otanta on suoritettu, mittauksella selvitetään satunnaismuuttujien saamat arvot.

Arvio eli *estimaatti* populaation tunnusluvulle lasketaan mitatuista arvoista x_1, x_2, \dots, x_n tapaukseen soveltuvan laskukaavan eli *estimaattorin* avulla. Estimaattorit ovat ns. *otossuureita*, eli funktioita, jotka riippuvat vain otoksen havainnoista (ja tätä kautta myös otoskoosta).

Käytetään seuraavia merkintöjä:

T = populaation tunnusluku

\hat{T} = tunnusluvun T estimaattori, eli laskukaava, jonka avulla estimoidaan T .

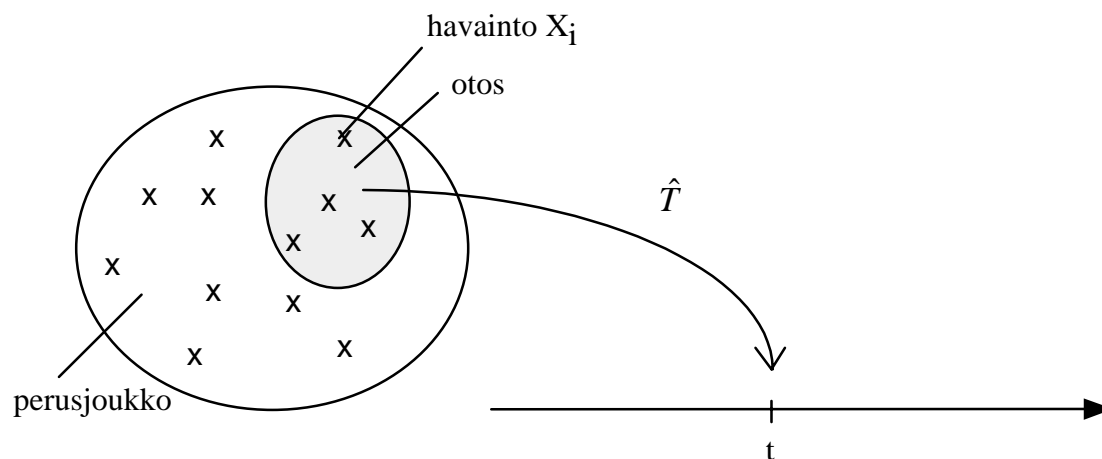
$$\hat{T} = \hat{T}(X_1, X_2, \dots, X_n).$$

t = tunnusluvun T estimaatti, eli estimaattorista saatava lukuarvo, kun siihen sijoitetaan tietyistä otoksesta mitatut havaintoarvot x_1, x_2, \dots, x_n , ts.

$$t = \hat{T}(x_1, x_2, \dots, x_n).$$

Huomaa, että estimaattorin arvo riippuu paitsi mitatuista havaintoarvoista, myös otoskoosta.

Kuvio 41: Estimaatin t laskeminen estimaattorin \hat{T} avulla otoksesta.



Esimerkiksi populaation keskiarvon $T = \mu$ estimaattorina voidaan käyttää otoskeskiarvon kaavaa

$$\hat{T} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

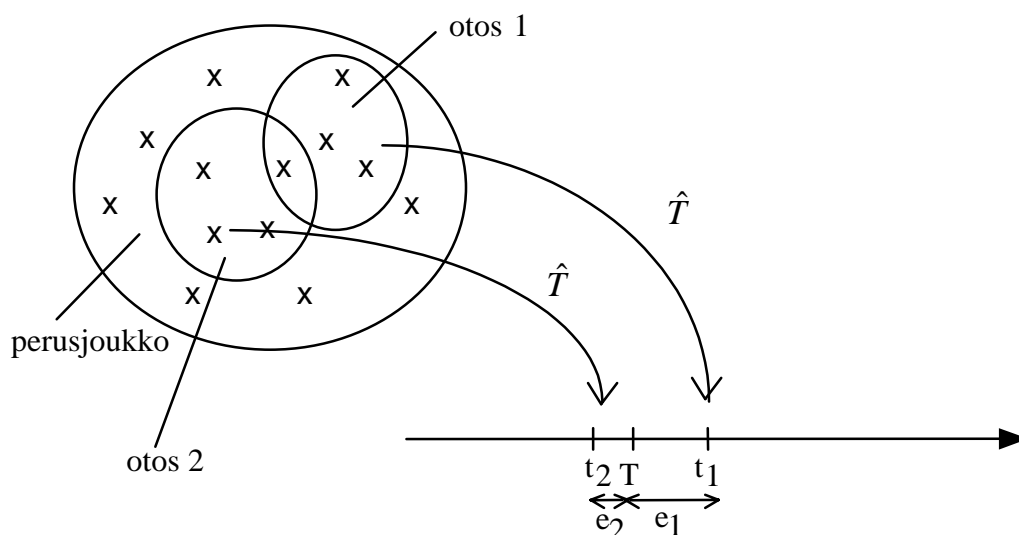
Koska estimaattori on satunnaismuuttujien funktio, on se itsekin satunnaismuuttuja.

Otantavirheeksi sanotaan otoksesta lasketun arvion ja populaation parametrin oikean arvon erotusta. Laskukaava otantavirheelle e annetaan muodossa

$$e = \hat{T} - T.$$

Otantavirheen suuruus vaihtelee otoksesta riippuen. Jos otoksesta j laskettu estimaatti on t_j , on tämän otoksen kohdalla otantavirheen suuruus $e_j = t_j - T$.

Kuvio 42: Estimaattorin \hat{T} avulla saadaan otoksesta 1 estimaatti t_1 ja otoksesta 2 estimaatti t_2 . Otokseen 1 liittyvä otantavirhe e_1 on estimaatin t_1 poikkeama populaation tunnusluvusta T ja otokseen 2 liittyvä otantavirhe e_2 on estimaatin t_2 poikkeama T :stä.



Esimerkki 43 (Manninen s. 145): 10 eri säähavaintoaseman eräänä päivänä mitaamat päivän maksimilämpötilat ($^{\circ}\text{C}$) ovat seuraavat:

asema	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
lämpötila ($^{\circ}\text{C}$)	9	18	15	13	11	17	14	15	10	13

Tutkimuksen kohteena on säähavaintoasemien kattaman alueen keskimääräinen maksimilämpötila kyseisenä päivänä, eli mitattujen maksimilämpötilojen keskiarvo. Tarkasteltava tilastollinen muuttuja X on siis kunkin aseman mittaama maksimilämpötila. Tässä populaationa on koko 10 aseman joukko, ja populaation

tilastoyksiköt ovat asemat a_1, \dots, a_{10} . Näin pienestä populaatiosta on helppo laskea koko populaation keskiarvo (kokonaistutkimus), joksi saadaan $\mu = 13,5$ °C.

Estimaattien ja otantavirheen käyttäytymisen selventämiseksi poimitaan nyt tästä populaatiosta 8 kpl kolmen yksikön suuruista otosta. Otanta suoritetaan yksinkertaisena satunnaisotantana satunnaislukutaulukon avulla. Käytetään populaation keskiarvon μ estimaattorina otoskeskiarvoa. Seuraavassa taulukossa esitetään kustakin otoksesta mitatut X :n arvot, näistä lasketut estimaatit sekä kuhunkin otokseen liittyvät otantavirheet. Otoksella O_i tarkoitetaan i :nnestä otoksesta mitattuja havaintoarvoja (x_1, x_2, x_3) .

Otoksen järjestysnumero i	otos O_i (X :n arvot)	otoskeskiarvo $\bar{x}_i = \hat{T}(O_i)$	otantavirhe $e_i = \bar{x}_i - \mu$
1	13, 10, 11	11.3	-2.2
2	11, 9, 17	12.3	-1.2
3	10, 17, 9	12.0	-1.5
4	15, 15, 11	13.7	0.2
5	15, 11, 9	11.7	-1.8
6	18, 15, 13	15.3	1.8
7	13, 14, 13	13.3	-0.2
8	13, 14, 15	14.0	0.5

Taulukosta nähdään, että estimaatit eli \bar{X} :n arvot vaihtelevat populaation keskiarvon μ ympärillä, jolloin otantavirhe vaihtelee nollan ympärillä. Lisäksi huomataan, että \bar{X} :n vaihtelu otoksesta toiseen on pienempää kuin X :n arvojen vaihtelu populaatiossa.

Jos halutaan tietää, kuinka luotettavia estimaatteja tietyllä estimaattorilla saadaan, on tunnettava ko. estimaattorin *otantajakauma*. Otantajakaumaksi sanotaan yleisesti otossuureen jakaumaa. Jos tuntisimme edellisen esimerkin tapauksessa otoskeskiarvon otantajakauman, voisimme sen perusteella arvioida vaikkapa sitä, millä todennäköisyydellä otoksesta laskettu keskiarvo poikkeaa populaation keskiarvosta korkeintaan 0.5 °C. Tässä tapauksessa populaatio on pieni, ja siksi olisi periaatteessa mahdollista määrittää otoskeskiarvon otantajakauma laskemalla otoskeskiarvot kaikista mahdollisista otoksista, jotka populaatiosta voidaan poimia. Käytännössä tällainen menettely on harvoin mielekäästä tai edes mahdollista. Otantatutkimuksissa populaatiot ovat tyypillisesti suuria ja joissain tapauksissa myös tuntemattomia.

Estimaattorien otantajakaumien määrittämiseen voidaan käyttää eräitä todennäköisyyslaskennan tuloksia. Täsmällisen tiheysfunktioilausekkeen johtaminen estimaattorille kuitenkin edellyttää, että tunnetaan tutkittavan muuttujan X jakauma populaatiossa. Näin ei käytännössä kovin usein tapahdu. Estimaattorien lausekkeet muodostuvat kuitenkin usein havaintojen summalausekkeista, jolloin otantajakauman likimääräiseen arviointiin voidaan käyttää ns. *keskeistä raja-arvolauseetta*.

Lause 44 (*keskeinen raja-arvolause*):

Olkoon X_1, X_2, \dots, X_n joukko samoin jakautuneita, riippumattomia satunnaismuuttujia, joiden yhteinen keskiarvo μ ja yhteinen varianssi σ^2 ovat äärellisiä. Silloin standardoidun muuttujan

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma}} = \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

jakauma lähestyy standardoitua normaalijakaumaa $N(0,1)$, kun n kasvaa rajatta

Käytännössä jo melko pienillä otoskoon n arvoilla Z on likimain $N(0,1)$ -jakautunut. Jos populaation jakauma on yksihiippuinen ja symmetrinen, Z on jotakuinkin normaali jo kun $n=3$. Jos taas populaation jakauma on kovin vino, voidaan Z :n jakaumaa pitää normaalina vasta kun $n \leq 30$. Jos populaation jakaumaa ei tunneta, riittävän suurena otoskokona onkin pidettävä jälkimmäistä.

Edellä esitettiin keskeinen raja-arvolause yksinkertaisessa muodossa. Tässä voidaan vielä huomauttaa, että yleisemmin, muutamien lisäoletusten vallitessa, riippumattomien satunnaismuuttujien summa on likimain normaalisti jakautunut, kun n on riittävän suuri, olivatpa näiden satunnaismuuttujien jakaumat millaiset hyvänsä.

Keskeinen raja-arvolause on suurten otosten teorian tärkeimpiä perustuloksia. Riittävän suurissa otoksissa otoskeskiarvo on suurin piirtein normaalisti jakautunut, riippumatta siitä, millainen X :n jakauma perusjoukossa on. Kun tarkastellaan pieniä otoksia, joudutaan usein olettamaan, että populaation jakauma on normaali. Tällöin voidaan käyttää hyväksi seuraavaa tulosta:

Lause 45 (*normaalisti jakautuneiden muuttujien yhteenlaskuominaisuus*):

Jos X_1, X_2, \dots, X_n ovat riippumattomia normaalisti jakautuneita satunnaismuuttujia, ja $X_i \sim N(\mu_i, \sigma_i^2)$, niin $Y = \sum_{i=1}^n X_i$ on normaalisti jakautunut, $Y \sim N(\mu, \sigma^2)$,

missä $\mu = \sum_{i=1}^n \mu_i$ ja $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Tästä seuraa erityisesti, että jos tutkittavan muuttujan X jakauma populaatiossa on normaali $N(\mu, \sigma^2)$ ja tästä populaatiosta poimitaan yksinkertainen satunnaisotos X_1, X_2, \dots, X_n , on otoskeskiarvon \bar{X} otantajakauma $N(\mu, \sigma^2/n)$.

Jos satunnaisotoksessa X_1, X_2, \dots, X_n muuttujat ovat keskenään riippumattomat ja X :n jakauma populaatiossa on $N(\mu, \sigma^2)$, sanotaan, että X_1, X_2, \dots, X_n on *riippumaton (satunnais)otos jakaumasta* $N(\mu, \sigma^2)$.

Tarkastellaan nyt lähemmin otoskeskiarvon otantajakaumaa. Edellisistä tuloksista voidaan koota yhteenvedoksi seuraavaa: jos populaation jakauman keskiarvo on μ ja varianssi σ^2 , niin n suuruudessa otoksessa

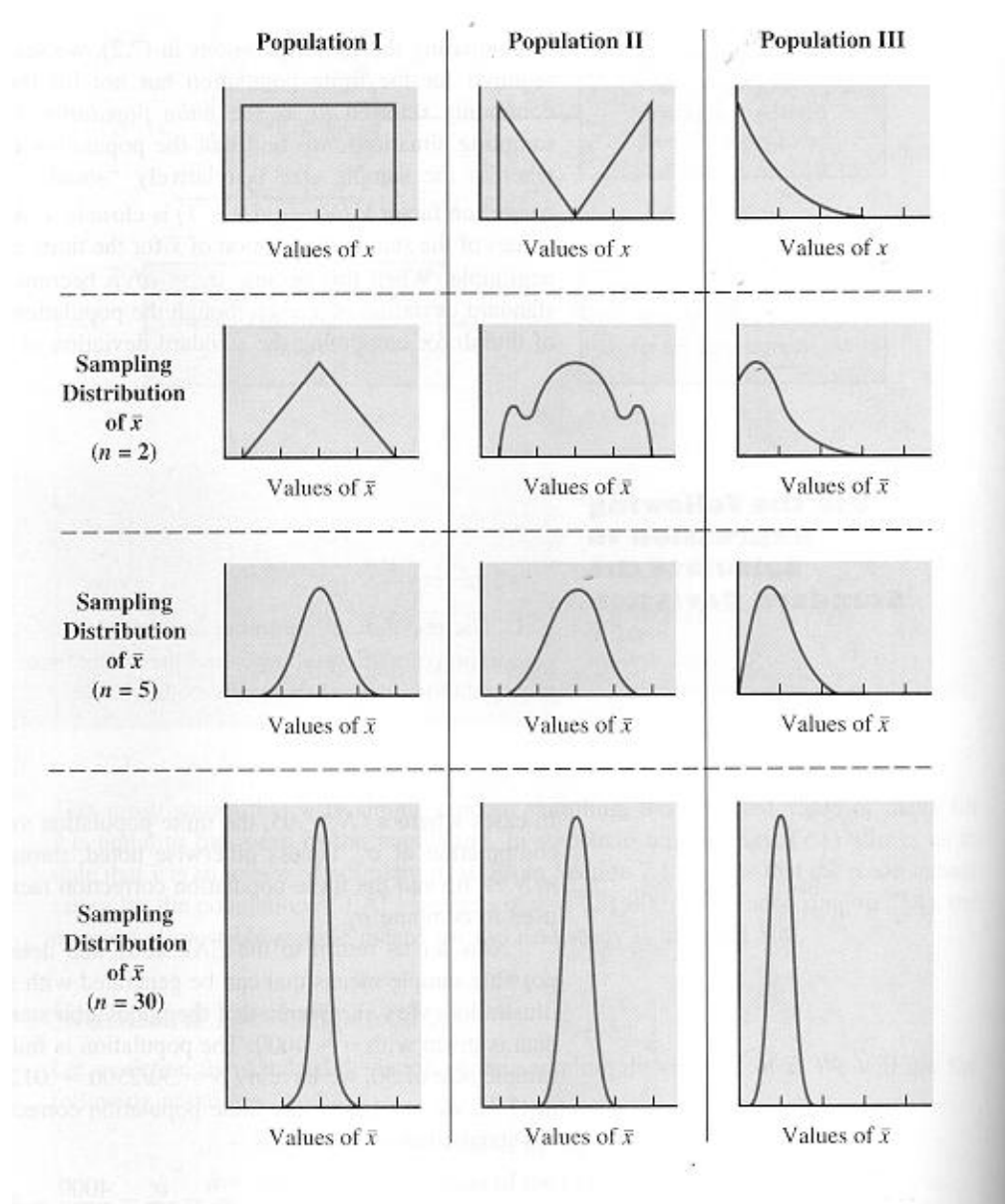
1) otoskeskiarvon otantajakauma on täsmälleen $N(\mu, \sigma^2/n)$ aina, kun populaation jakauma on normaali. ja

2) otoskeskiarvon otantajakauma on likimain $N(\mu, \sigma^2/n)$, mikäli otoskoko $n \geq 30$ ja populaation jakauma ei ole normaali.

Tarkastellaan seuraavaksi tapausta, jossa populaation jakauma on normaali, $N(\mu, \sigma^2)$. Koska otoskeskiarvon \bar{X} odotusarvo on μ , otoskeskiarvon jakauma keskittyy populaation keskiarvon ympärille. Tämä ominaisuus huomattiinkin esimerkin 43 yhteydessä. Otoskeskiarvon varianssi on σ^2/n , eli se on X :n varianssia pienempi. Mitä suurempi otoskoko on, sitä pienempi on \bar{X} :n varianssi, eli sitä tiiviimmin \bar{X} :n jakauma keskittyy μ :n lähelle.

Seuraavan sivun kuvio havainnollistaa keskiarvon otantajakauman käyttäytymistä kun otoskoko kasvaa. Kuviossa nähdään myös populaation jakauman vaikutus otantajakaumaan.

Kuvio 46 (Anderson & al., s. 226): *Otoskeskiarvon \bar{X} otantajakauma erilaisilla otoskoon n arvoilla kolmen eri populaation jakauman tapauksessa.*



Estimaatin t luotettavuutta arvioidaan usein ilmoittamalla estimaatin *keskivirhe* σ_t . Tällä tarkoitetaan estimaattorin jakauman keskihajontaa. Otoskeskiarvon keskivirhe on siis

$$\sigma_{\bar{x}} = D(\bar{X}) = \sqrt{D^2(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Samoin kuin varianssi keskiarvon keskivirhe pienenee eli otoskeskiarvon luotettavuus kasvaa otoskoon kasvaessa.

Edellä esitetyt tulokset koskevat tarkkaan ottaen vain otantaa palauttaen, koska tällöin muuttujat X_1, X_2, \dots, X_n ovat riippumattomia. Jos suoritetaan otanta palauttamatta, on keskiarvon keskivirheen kaava

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$$

missä $(N-n)/(N-1)$ on äärellisen populaation korjauskerroin. Jos populaatio ei ole äärettömän suuri, kerroin $(N-n)/(N-1)$ ja samalla myös termi $\sqrt{\frac{N-n}{N-1}}$ on

ykköistä pienempi aina, kun otoskoko $n > 1$. Jos populaation koko on hyvin suuri ja otoskoko siihen nähden kohtalaisen pieni, kerroin on hyvin lähellä ykköstä ja voidaan jättää huomiotta. Käytännössä rajana pidetään usein otantasuhteen n/N arvoa 0.05. Äärettömän populaation tapauksessa kerroin on tasan yksi, ja keskivirhe on sama kuin otannassa palauttaen. Siis suoritettaessa otanta palauttamatta, jos $n/N \leq 0.05$ tai populaatio on ääretön, voidaan keskiarvon keskivirhe laskea kaavalla $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Jäljempänä monisteessa oletetaan, jollei erikseen toisin

mainita, että kyseessä on otanta palauttaen tai että populaation koko on riittävän suuri otoskokoon verrattuna.

2.2.1 Piste-estimointi

Piste-estimoinnissa etsitään populaation parametrille sellaista estimaattoria, jonka arvot olisivat keskimäärin mahdollisimman lähellä parametrin oikeaa arvoa. Edellä esimerkissä käytettiin populaation keskiarvon piste-estimaattorina otoskeskiarvoa, jonka tuottamien estimaattien keskiarvo on täsmälleen populaation keskiarvon suuruinen. Populaation keskiarvolle voidaan löytää muitakin estimaattoreita, joilla on tämä sama ominaisuus. Jos populaation jakauma on

$N(\mu, \sigma^2)$, suurissa otoksissa myös otosmediaanin odotusarvo on μ . Samoin kaikki havaintojen X_1, X_2, \dots, X_n painotetut keskiarvot ja jopa jokainen yksityinen havainto X_i täyttävät tämän kriteerin.

Nyt voidaankin kysyä, että jos tietylle populaation tunnusluvulle voidaan konstruoida monia piste-estimaattoreita, niin mikä näistä mahdollisista estimaattoreista sitten on paras? Tämän ratkaisemiseen käytetään useampaakin kriteeriä.

Estimaattorin hyvyys

Saman tunnusluvun eri estimaattoreiden hyvyyden vertaileminen perustuu niiden otantajakaumiin, sillä otantajakauma kuvaa estimaattorin arvon vaihtelua otoksesta toiseen. Tarvittava tieto saadaan tiivistetyssä muodossa otantajakauman tunnusluvuista, esim. odotusarvosta ja varianssista. Seuraavassa esitellään eräitä piste-estimaattorin toivottavia ominaisuuksia, joiden perusteella estimaattorin hyvyyttä arvioidaan.

1. Harhattomuus

Satunnaisotoksesta estimaattorin avulla laskettu estimaatti voi vain harvinaisen sattuman kautta antaa populaation tunnusluvulle täsmälleen oikean arvon. Sattuman aiheuttamaa poikkeamaa oikeasta arvosta eli otantavirhettä ei voi mitenkään välttää, mutta hyvältä estimaattorilta edellytetään, että se ei anna *systemaattisesti* virheellistä arviota. Otantavirhe ei ole systemaattinen, mikäli sen odotusarvo on nolla, ts. kaikista mahdollisista otoksista laskettujen estimaattien joukon positiiviset ja negatiiviset otantavirheet kumoavat toisensa.

Estimaattoria \hat{T} sanotaan tunnusluvun T *harhattomaksi* estimaattoriksi, mikäli estimaattorin odotusarvo on sama kuin tunnusluvun arvo, eli

$$E(\hat{T}) = T.$$

Tällöin otantavirheen $e = \hat{T} - T$ odotusarvo on

$$E(e) = E(\hat{T} - T) = E(\hat{T}) - T = 0.$$

Jos estimaattori ei ole harhaton, sitä sanotaan *harhaiseksi*. Estimaattorin \hat{T} *harha* $B(\hat{T})$ onkin juuri otantavirheen odotusarvo ja siis määritellään

$$B(\hat{T}) = E(\hat{T}) - T,$$

eli se on estimaattorin odotusarvon ja populaation tunnusluvun arvon erotus.

Harhattomuus on tärkein estimaattorin hyvyyden kriteereistä. Jos tunnusluvulle ei löydy harhatonta estimaattoria, vaaditaan ainakin *asymptoottista harhatto-
muutta*. Asymptoottisesti harhattoman estimaattorin odotusarvo lähestyy tunnus-
luvun arvoa, kun otoskoko n kasvaa, ts.

$$E(\hat{T}) \rightarrow T, \text{ kun } n \rightarrow \infty.$$

Esimerkki 47: Aiemmin todettiin, että otoskeskiarvon otantajakauman odotusarvo on sama kuin populaation keskiarvo, joten otoskeskiarvo on populaation keskiarvon harhaton estimaattori. Tämä seuraa suoraan odotusarvon ominaisuuksista:

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} n\mu = \mu.$$

Samalla tavoin voidaan näyttää, että otosvarianssi s^2 on populaation varianssin σ^2 harhaton estimaattori, eli että $E(s^2) = \sigma^2$.

2. Varianssin pienuus

Harhattomuus yksinään ei riitä estimaattorien vertailun kriteeriksi. Joissain tapauksissa tunnusluvulle voidaan löytää useampia harhattomia estimaattoreita. Esim. jos tutkittavan muuttujan arvojen jakauma populaatiossa on symmetrinen (kuten normaalijakauma), otoskeskiarvon lisäksi myös mediaani on populaation keskiarvon harhaton estimaattori.

Kahdesta saman tunnusluvun harhattomasta estimaattorista parempi on se, jolla on pienempi varianssi. Tämä merkitsee, että eri otoksista saatavat estimaatit ovat keskimäärin lähempänä populaation tunnusluvun oikeaa arvoa. Jos \hat{T}_1 ja \hat{T}_2 ovat populaation tunnusluvun T harhattomia estimaattoreita ja \hat{T}_1 :n varianssi on pienempi kuin \hat{T}_2 :n, eli $D^2(\hat{T}_1) < D^2(\hat{T}_2)$, sanotaan, että \hat{T}_1 on *tehokkaampi* kuin \hat{T}_2 . *Tehokkain* eli *minimivarianssiestimaattori* on sellainen tunnusluvun T harhaton

estimaattori, jonka varianssi on pienempi kuin T :n minkään muun harhattoman estimaattorin.

Voidaan osoittaa, että otoskeskiarvo on tehokkain populaation keskiarvon estimaattori.

3. Tarkentuvuus

Tarkentuvuus yleisesti tarkoittaa sitä, että otoskokoa kasvattamalla estimaattori saadaan yhä tarkemmaksi. Tämä voidaan tilanteesta riippuen määritellä eri tavoin. Esimerkiksi estimaattorin \hat{T} sanotaan olevan *kvadraattisesti tarkentuva*, jos otantavirheen neliön odotusarvo pienenee nollaan, kun otoskoko n kasvaa äärettömän suureksi, eli

$$E(\hat{T} - T)^2 \rightarrow 0, \text{ kun } n \rightarrow \infty.$$

Tarkentuva estimaattori antaa tunnusluvulle keskimäärin sitä paremman arvion, mitä suurempaan otokseen se perustuu.

Esimerkiksi otoskeskiarvo on tarkentuva populaation keskiarvon estimaattori.

2.2.2 Väliestimointi

Väliestimoinnissa otetaan eksplisiittisesti huomioon se epävarmuus, joka estimointiin väistämättä liittyy. Yksittäisen estimaattiarvon sijasta määritetäänkin sellainen populaation tunnusluvun T arvoväli, joka sisältää oikean T :n arvon halutun suuruisella todennäköisyydellä. Siis määritetään sellainen väli $V = (T_1, T_2)$, että

$$P(T_1 \leq T \leq T_2) = 1 - \alpha,$$

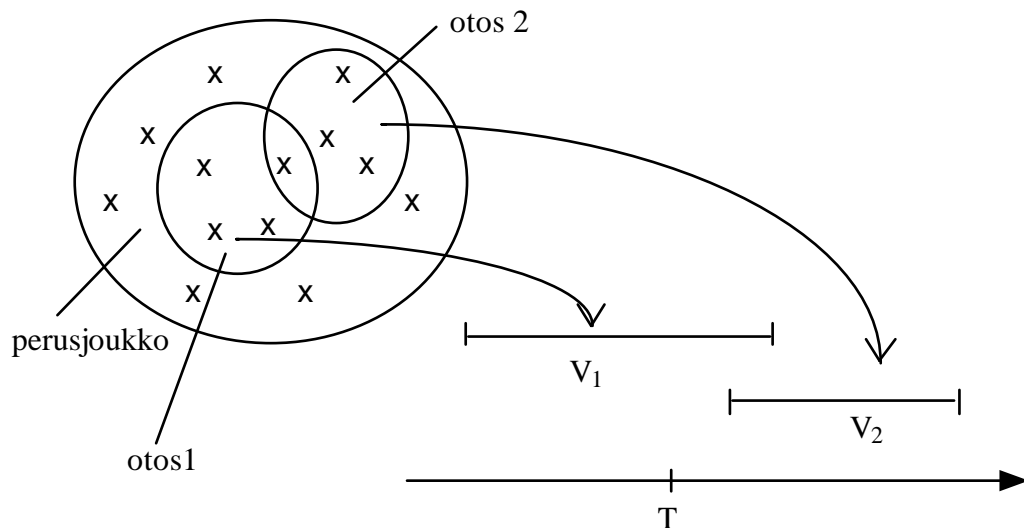
missä α on jokin pieni etukäteen määrätty luku, esim. 0.01 tai 0.05 ($0 < \alpha < 1$). Väliä V sanotaan tunnusluvun T $100(1-\alpha)$ %:n luottamusväliksi. Tässä $1-\alpha$ on se todennäköisyys, jolla V sisältää T :n. Esimerkiksi jos $\alpha = 0.05$, V sisältää T :n todennäköisyydellä 0.95 ja V on silloin T :n 95 %:n luottamusväli. Todennäköisyyttä $1-\alpha$ sanotaan myös *luottamustasoksi*.

Luottamusvälin määrittämiseksi tunnusluvulle T käytetään T :n pisteestimaattoria. Tehokkain estimaattori antaa yleensä mahdollisimman lyhyen luot-

tamusvälin. Piste-estimaattorin otantajakaumaa hyväksi käyttäen muodostetaan laskukaavat luottamusvälin päätepisteille.

Otoksen perusteella laskettava väliestimaatti saadaan sijoittamalla otoksen havainnoista lasketut tarvittavat suureet näihin kaavoihin. Tyypillisesti päätepisteet lasketaan piste-estimaatin keskivirheen avulla, jolloin piste-estimaatti sijaitsee luottamusvälin keskipisteessä. Väliestimaatti on siis otoksesta riippuvainen ja sen arvo vaihtelee otoksesta toiseen.

Kuvio 48: Otoksesta 1 on laskettu tunnusluvulle T luottamusväli V_1 ja otoksesta 2 luottamusväli V_2 .



Edellä olevassa kuviossa havainnollistetaan myös sitä, että yksittäisestä otoksesta laskettu luottamusväli ei välttämättä sisällä tunnusluvun T oikeaa arvoa, ts. on mahdollista, että otoksessa i $T \notin V_i$. Kuitenkin minkä hyvänsä otoksen kohdalla on voimassa, että $P(T \in V_i) = 1 - \alpha$, eli tunnusluku sisältyy väliin V_i todennäköisyydellä $1 - \alpha$.

Esimerkki 49: Määritetään $100(1 - \alpha)\%$:n luottamusväli populaation keskiarvolle μ .

Käytetään välin määrittämiseen otoskeskiarvoa \bar{X} . Jos oletetaan, että populaation jakauma on $N(\mu, \sigma^2)$ tai otoskoko $n \geq 30$, niin otoskeskiarvon jakauma on ainakin likimain $N(\mu, \sigma^2/n)$. Normaalijakauman ominaisuuksista tiedetään, että esim. sen tiheysfunktioista odotusarvosta molemmiin puolin kahden hajonnanmitan pää-

hän rajatulle alueelle, eli välille $(\mu-2\sigma, \mu+2\sigma)$ sijoittuu 95.44 % todennäköisyysmassasta. Tämä tarkoittaa, että satunnaiskoetta suoritettaessa normaalisti jakautunut muuttuja saa arvokseen jonkin tällä alueella sijaitsevan arvon todennäköisyydellä 0.9544.

Otoskeskiarvon hajonta on $D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, joten edellä mainittu väli on $\left(\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right)$. Siis satunnaisesti poimitusta otoksesta laskettu otoskeskiarvo sijaitsee 95.44%:n todennäköisyydellä välillä $\left(\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right)$. Tämä voidaan tulkita myös siten, että otoskeskiarvoa laskettaessa otantavirhe on todennäköisyydellä 0.9544 korkeintaan $2\frac{\sigma}{\sqrt{n}}$ suuruinen. Kun laskettu otoskeskiarvo on \bar{X} , niin edellisen nojalla voidaan sanoa, että todennäköisyydellä 0.9544 populaation keskiarvo kuuluu väliin $\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)$. Tämä väli on siis populaation keskiarvon 95.44 %:n luottamusväli.

Luottamustason suuruus riippuu siitä, monenko hajonnanmitan päähän odotusarvosta raja vedetään. Kolmen hajonnanmitan päässä saadaan luottamustasoksi 0.9974, eli 99.74 %:n luottamusväli populaation keskiarvolle on $\left(\bar{X} - 3\frac{\sigma}{\sqrt{n}}, \bar{X} + 3\frac{\sigma}{\sqrt{n}}\right)$.

Jos halutaan määrittää luottamusväli tietynsuuruuisella luottamustasolla $1-\alpha$, merkitään vastaavaa hajonnanmittaa $z_{\alpha/2}$:lla. Tätä lukua sanotaan luottamustason $1-\alpha$ kriittiseksi arvoksi. Populaation keskiarvon μ 100(1- α) %:n luottamusvälin päätepisteet ovat siis

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right).$$

Kriittinen arvo halutulle luottamustasolle saadaan standardoidun normaalijakauman kertymäfunktion avulla kaavasta $\Phi(z_{\alpha/2}) = 1-\alpha/2$. Jos luottamustasoksi valitaan vaikkapa $1-\alpha = 0.95$, niin

$$\begin{aligned} \alpha &= 0.05, \\ \alpha/2 &= 0.025 \text{ ja} \end{aligned}$$

$$1-\alpha/2 = 0.975.$$

Taulukosta nähdään, että $\Phi(1.96) = 0.975$, joten luottamustason 0.95 kriittinen arvo on $z_{0.025} = 1.96$. Tämän avulla saadaan 95 %:n luottamusväliksi

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right).$$

Muita usein käytettyjä luottamusvälejä ovat 90 %:n ja 99 %:n välit. Näiden kriittiset arvot $z_{\alpha/2}$ ovat seuraavat:

$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.10	0.05	1.65
0.99	0.01	0.005	2.58

Keskiarvon keskivirheen ja sen myötä keskiarvon luottamusvälin laskemiseen tarvitaan populaation keskihajontaa. Monissa käytännön tutkimustilanteissa ei tätä kuitenkaan tunneta, vaan keskihajonta on estimoitava otoksesta. Populaation keskihajonnan σ

estimaattorina käytetään usein otoskeskihajontaa $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$. Tä-

män tunnusluvun estimointi tuo lisää epävarmuutta populaation keskiarvon arviointiin, mistä seuraa, että ainoastaan suurten otosten tapauksessa voidaan otoskeskiarvoa pitää likimain $N(\mu, s^2/n)$ -jakautuneena.

Jos populaation jakauma on ainakin likimain normaali ja populaation hajontaa ei tunneta, noudattaa otoskeskiarvo ns. *t-jakaumaa*. Tämä, kuten eräät muutkin jäljempänä eteen tulevat jakaumat, on johdettu normaalijakaumasta. Jakauma riippuu otoskoon suuruudesta ja otoskoon kasvaessa se lähenee normaalijakaumaa, joten t-jakauma onkin tärkeä nimenomaan pienten otosten ($n < 30$) tapauksessa. Suurten otosten tapauksessa normaalijakauma antaa riittävän tarkan approksimaation. T-jakaumaan palataan myöhemmin tilastollisen päättelyn yhteydessä.

2.3 Tilastollinen merkitsevyystestaus

Tilastollisessa merkitsevyystestauksessa, kuten estimoinnissakin, tehdään perusjoukkoa koskevia päätelmiä otoksen tuottaman havaintoaineiston pohjalta, mutta

asetelma on erilainen kuin estimoinnissa. Merkitsevyystestaus lähtee liikkeelle ns. *nollahypoteesin* asettamisesta. Nollahypoteesi on jokin populaatiota koskeva väittäjä tai ennakko-oletus. Testi sisältää päättelysäännön, jonka perusteella nollahypoteesi joko hylätään tai hyväksytään.

Tarkkaan ottaen hypoteesin taustalla on oletus siitä, että populaatio on jonkin tietyn tilastollisen mallin mukainen. Testauksessa selvitetään, onko otoksesta saatu havaintoaineisto sopusoinnussa tämän mallin kanssa.

Nollahypoteesi määräytyy tutkimusongelman perusteella, mutta lisäksi on otettava huomioon, millaisia tilastollisia testejä on käytettävissä. Testejä on kehitetty esimerkiksi muuttujan jakauman sijaintia ja hajontaa ja muuttujien välistä riippuvuutta koskevaa päättelyä varten.

Sopivan testin valintaan vaikuttavat mm. tutkittavien muuttujien mittaustaso, otoskoko ja se informaatio, joka populaation jakaumasta on saatavilla. Karkein jako testien välillä voidaan tehdä 1) *parametrisiin* ja 2) *ei-parametrisiin* eli *jakaumasta vapaisiin* testeihin. Parametrisia testejä voidaan tehdä vähintään välimatka-asteikon tasoisille muuttujille, kategorisille muuttujille taas soveltuvat ei-parametriset testit.

Parametrisissa testeissä tehdään päätelmiä populaation jakauman parametreista, jolloin testissä joudutaan usein tekemään oletuksia populaation jakaumatyypistä. Näitä testejä onkin syytä käyttää vain silloin, kun jakaumaoletusten paikkansapitävyys voidaan edes suurin piirtein arvioida. Jos populaation jakaumasta ei ole riittävästi tietoa, suositellaan käytettäväksi ei-parametrisia testejä. Tässä monisteessa käsiteltävistä testeistä χ^2 -riippumattomuus- ja χ^2 -yhteensopivuustestit ovat ei-parametrisia testejä, muut ovat parametrisia.

2.3.1 Hypoteesien testaaminen

Hypoteesien asettaminen

Kun testattava nollahypoteesi on asetettu, voidaan sille valita tilanteeseen soveltuva *vastahypoteesi*. Vastahypoteesilta edellytetään, että se ei voi missään tilanteessa olla nollahypoteesin kanssa yhtäaikaan tosi. Nollahypoteesista käytetään yleisesti merkintää H_0 ja vastahypoteesista merkintää H_1 .

Esimerkki 50: Ohrasen leipomo valmistaa ryssänlimppuja, joiden keskipainon tulisi olla tasan 1 kg. Tämän tarkistamiseksi valittiin limpuista 40 kappaleen otos,

josta laskettiin limppujen keskipainoksi 1.020 kg. Ennalta tiedetään, että limppujen painon keskihajonta (koko populaatiossa) on 0.060 kg. Voidaanko otoksen perusteella tehdä johtopäätös, että limput ovat keskimäärin eripainoisia kuin on tarkoitettu?

Nollahypoteesina tässä ongelmassa toimii ennakko-oletus, jonka mukaan limppujen keskipaino on 1 kg. Nollahypoteesi siis väittää, että limppupopulaation (kaikki leipomon valmistamat limput) painojen keskiarvo μ on 1 kg. Vastahypoteesi on muotoiltu jo esitetyssä kysymyksessä: limppujen keskipaino poikkeaa 1 kg:sta. Hypoteesit voidaan kirjoittaa

$$H_0: \mu = 1 \text{ kg}$$

$$H_1: \mu \neq 1 \text{ kg}$$

Esimerkissä suoritetaan ns. *kaksisuuntainen testi*. Jos leipomo olisi kiinnostunut vain siitä, poikkeako keskipaino halutusta vain jompaan kumpaan suuntaan, tulisi kyseeseen *yksisuuntainen testi*. Yksisuuntainen testi eroaa kaksisuuntaisesta vastahypoteesin muotoilun perusteella. Jos leipomolle on tärkeää vain se, *alittavatko* limput halutun painon vai eivät, suoritetaan testi

$$H_0: \mu = 1 \text{ kg}$$

$$H_1: \mu < 1 \text{ kg},$$

ja jos halutaan päätellä, *ylittävätkö* limput halutun painon, testataan hypoteesit

$$H_0: \mu = 1 \text{ kg}$$

$$H_1: \mu > 1 \text{ kg}.$$

Tutkimustilanteessa hypoteesit asetetaan *ennen* otoksen poimimista. Jos tutkijalla ei tässä vaiheessa ole ennakkokäsitystä siitä, mihin suuntaan oletetusta arvosta populaation parametri mahdollisesti poikkeaa, on vastahypoteesi muotoiltava kaksisuuntaiseksi. Yksisuuntaista testiä käytetään vain, jos tutkijaa ennakolta kiinnostaa ainoastaan poikkeama toiseen suuntaan.

Hypoteesin testauksen periaatteet

Hypoteesin testaus perustuu otoksen perusjoukosta antamaan informaatioon. Tätä informaatiota verrataan nollahypoteesin mukaiseen perusjoukkoa koskevaan malliin. Jos otoksesta saatu tieto ei ole selvästi ristiriidassa oletetun mallin kanssa,

nollahypoteesi hyväksytään. Muussa tapauksessa nollahypoteesi hylätään, ja vastahypoteesin katsotaan olevan tosi.

Esimerkissä 50 nollahypoteesia vastaavan mallin mukaan limppujen populaation jakauman odotusarvo on 1 kg. Otoksen perusteella laskettu otoskeskiarvo poikkesi tästä 20 gramman verran. Kuitenkin tiedetään, että otantaan vaikuttava sattuuma tuottaa otoksia, joiden keskiarvot poikkeavat populaation keskiarvosta. Voidaanko 20 gramman suuruista poikkeamaa pitää normaalina sattuman aiheuttamana otantavirheenä, jos populaatio on oletuksen mukaisesti jakautunut? Vai onko poikkeama niin suuri, että on epäuskottavaa, että otos olisi peräisin oletuksen mukaisesti jakautuneesta populaatiosta?

Kysymykseen vastaamiseksi on ensiksikin määritettävä, mitä tarkoitetaan uskottavuudella ja sitten sovittava uskottavuuden ja epäuskottavuuden välisestä rajasta. Uskottavuutta voidaan arvioida todennäköisyyksien avulla. Tässä tapauksessa tarkastellaan erisuuruisten otantavirheiden todennäköisyyksiä. Jos oletuksen mukaisesta jakaumasta poimitun otoksen kohdalla keskiarvon korkeintaan 20 g suuruisen otantavirheen todennäköisyys on riittävän suuri, voidaan pitää uskottavana, että otos on todella mainitusta jakaumasta peräisin. Sille, mikä on ”riittävän suuri” todennäköisyys, ei löydy täsmällistä objektiivista kriteeriä. Testauksen yhteydessä tutkija määrää etukäteen tämän uskottavuusrajan, ja se toimii mittarina riskille, että testin perusteella tehty johtopäätös on väärä.

Tilastollisen testauksen avulla ei siis voi koskaan *varmistaa*, että esitetty hypoteesi on oikea tai väärä. Testin tulos on päätössääntö, jonka nojalla nollahypoteesi hyväksytään tai hylätään, mutta johtopäätöksen virheettömyydestä ei voi olla täysin varma.

Verrattaessa testin perusteella tehtyä johtopäätöstä siihen, mikä tutkittavan ilmiön tila todellisuudessa on, seuraavat neljä tapausta ovat mahdollisia:

		Päätös	
Todellinen tila	H_0 tosi	H_0 hyväksytään	H_1 hyväksytään
	H_1 tosi	oikein	I lajin virhe
		II lajin virhe	oikein

Virheellinen johtopäätös voi siis syntyä kahdella eri tavalla: joko hylätään nollahypoteesi, vaikka se on oikea, eli tehdään *I lajin virhe* tai sitten hyväksytään nollahypoteesi, vaikka se on väärä, eli tehdään *II lajin virhe*. I lajin virhettä sanotaan

myös *hylkäämisvirheeksi* ja II lajin virhettä *hyväksymisvirheeksi*. Näistä kahdesta I lajin virhettä pidetään vakavampana.

Testauksessa asetetaan hylkäämisvirheen todennäköisyydelle yläraja. Tätä rajaa sanotaan *testin* merkitsevyystasoksi ja merkitään α :lla. Testin päätössäännöt nojautuvat merkitsevyystasoon: jos otoksen perusteella laskettu hylkäämisvirheen todennäköisyys on korkeintaan α , niin nollahypoteesi hylätään.

Hypoteesien testaukseen liittyvät päätössäännöt voidaan määritellä seuraavasti:

Tapa 1:

Lasketaan todennäköisyys sille, että otantavirhe on vähintään otoksesta lasketun suuruinen. Tätä todennäköisyyttä sanotaan *havaituksi merkitsevyystasoksi* tai *p-arvoksi*. Jos kyseinen todennäköisyys on p , ja vähintään otoksesta lasketun suuruiset poikkeamat johtavat nollahypoteesin hylkäämiseen, niin hylättäessä nollahypoteesi tehdään virhe korkeintaan todennäköisyydellä p .

Tapa 2:

Merkitsevyystason α avulla lasketaan raja suurimmalle “uskottavalle” otantavirheelle: rajaluku määritetään siten, että tätä suuremman otantavirheen todennäköisyys on α . Jos siis otoksen perusteella laskettu poikkeama nollahypoteesista on suurempi kuin mainittu rajaluku, nollahypoteesi hylätään. Ko. rajalukua sanotaan kriittiseksi arvoksi.

Hyväksymis- ja hylkäämisvirheiden suuruudet liittyvät toisiinsa. Jos hylkäämisvirheen riskiä pienennetään, hyväksymisvirheen riski kasvaa. Jos hylkäämisvirheen raja on jo määritetty, hyväksymisvirheen suuruutta ei voi säädellä vastavalla tavalla, mikäli käytettävää testiä tai otoskokoa ei muuteta. Otoskokoa kasvattamalla voidaan pienentää kummankin virheen riskiä.

Testin merkitsevyystasoa ei siis kannata valita liian pieneksi, jotta hyväksymisvirhe pysyisi kohtuullisissa rajoissa. Usein merkitsevyystasoksi valitaan $\alpha = 0.05$. Muita yleisimmin käytettyjä merkitsevyystasoja ovat $\alpha = 0.01$ ja $\alpha = 0.001$.

Jos testin merkitsevyystaso on α ja nollahypoteesi hylätään, niin sanotaan, että havaintoaineisto on *tilastollisesti merkitsevä* testattavan hypoteesin vastainen todiste (*merkitsevyys*)tasolla α tai 100α %:n (*merkitsevyys*)tasolla. Toisen käytännön mukaan edellä mainittuja yleisimpiä merkitsevyystasoja luonnehditaan kuta-

kin omalla ilmaisullaan seuraavasti: Jos testin merkitsevyystaso on α ja nollahypoteesi hylätään, niin tasolla $\alpha = 0.05$ on *tilastollisesti melkein merkitsevä*, tasolla $\alpha = 0.01$ *tilastollisesti merkitsevä* ja tasolla $\alpha = 0.001$ *tilastollisesti erittäin merkitsevä* ero havaintojen ja nollahypoteesin välillä.

Testisuure ja päätössäännöt

Edellä päädyttiin siihen, että nollahypoteesin hyväksymis- tai hylkäämissääntö voidaan liittää vähintään havaitunsuuruuden otantavirheen todennäköisyyteen. Seuraavassa tarkastellaan, miten kyseistä todennäköisyyttä voidaan arvioida.

Kaksisuuntainen testaus

Esimerkissä 50 nollahypoteesi väitti, että tutkittavan limppupopulaation jakauman odotusarvo (eli populaation keskiarvo) on 1 kg. Väittämä siis koskee populaation jakauman tunnuslukua, jolloin tiettyjen oletusten vallitessa testiksi voidaan valita parametrisiin testeihin kuuluva *keskiarvotesti*. Hypoteesit olivat

$$H_0: \mu = 1 \text{ kg}$$

$$H_1: \mu \neq 1 \text{ kg}.$$

Otoksesta oli laskettu otoskeskiarvo 1.020 kg ja otoskoko oli $n = 40$, jota voidaan pitää suurena otoksena. Merkitään nyt nollahypoteesissa oletettua populaation keskiarvoa eli populaation jakauman odotusarvoa μ_0 :lla. Esimerkissä siis $\mu_0 = 1$ kg. Jos populaation jakauma todella on nollahypoteesin mukainen, niin otoskeskiarvon otantajakauma on likimain normaali, parametreina populaation keskiarvo μ_0 ja otosvarianssi σ^2/n , ts. $\bar{X} \sim N(\mu_0, \sigma^2/n)$. Tämän tiedon nojalla voidaankin arvioida otantavirheiden todennäköisyyksiä. Väliestimoinnin yhteydessä todettiin, että otoskeskiarvon otantavirhe on todennäköisyydellä $1-\alpha$ korkeintaan $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ suuruinen, missä $z_{\alpha/2}$ oli luottamustason $1-\alpha$ kriittinen arvo. Testauksen yhteydessä tätä lukua sanotaan *kaksisuuntaisen testin merkitsevyystason α kriittiseksi arvoksi*. Siis

$$P(|e| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1-\alpha,$$

missä otantavirhe $e = \mu_0 - \bar{X}$. Koska

$$|e| = |\mu_0 - \bar{X}| = |\bar{X} - \mu_0|,$$

epäyhtälö

$$|e| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

voidaan kirjoittaa

$$|\bar{X} - \mu_0| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

eli

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2}.$$

Edellä mainittu todennäköisyys voidaan ilmaista muodossa

$$P(|e| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = P\left(\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2} \right) = 1 - \alpha,$$

eli

$$P\left(\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \right) = \alpha.$$

Jos siis testin merkitsevyystasoksi on valittu α , niin *testisuureen* $z = \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right|$ ne

arvot, jotka ovat pienempiä kuin $-z_{\alpha/2}$ tai suurempia kuin $z_{\alpha/2}$, vastaavat "liian" suuria poikkeamia keskiarvosta ja näin ollen toimivat nollahypoteesin vastaisena todisteena. Kriittinen arvo määritetään, kuten väliestimoinnin yhteydessä mainittiin, standardoidun normaalijakauman kertymäfunktion taulukosta kaavalla $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Testisuure z voidaan yksinkertaisemmin muodostaa standardoimalla otoskeskiarvo. Koska $\bar{X} \sim N(\mu_0, \sigma^2/n)$, on sen standardoitu muuttuja

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

Lähtökohtana tässä päättelyssä oli oletus, että populaation jakauma oli nollahypoteesin mukainen, esimerkiksi $50 \mu_0 = 1 \text{ kg}$. Testisuureen arvon laskemisessa tarvitaan lisäksi populaation hajonta $\sigma = 0.060$ ja otoskoko $n = 40$. Testisuure on

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{X} - 1}{0.06 / \sqrt{40}}.$$

Jos merkitsevyystasoksi valitaan vaikkapa $\alpha = 0.05$, kriittinen arvo on $z_{\alpha/2} = z_{0.025} = 1.96$, joten päätössäännöksi tässä tapauksessa saadaan seuraava:

$$\begin{aligned} \text{jos } \frac{\bar{X} - 1}{0.06 / \sqrt{40}} < -1.96 \text{ tai } \frac{\bar{X} - 1}{0.06 / \sqrt{40}} > 1.96, \quad H_0 \text{ hylätään} \\ \text{jos } -1.96 < \frac{\bar{X} - 1}{0.06 / \sqrt{40}} < 1.96, \quad H_0 \text{ hyväksytään.} \end{aligned}$$

Kun testisuureen kaavaan sijoitetaan vielä otoksesta laskettu otoskeskiarvo $\bar{X} = 1.020 \text{ kg}$, saadaan testisuureen *havaittu arvo* z_{hav} :

$$z_{hav} = \frac{\bar{X} - 1}{0.06 / \sqrt{40}} = \frac{1.020 - 1}{0.06 / \sqrt{40}} = 2.108.$$

Koska

$$z_{hav} = 2.108 > 1.96,$$

niin päätössäännön mukaan nollahypoteesi hylätään. Otoksesta laskettu keskiarvo $\bar{X} = 1.020 \text{ kg}$ poikkeaa oletetusta populaation keskiarvosta $\mu_0 = 1 \text{ kg}$ niin paljon, että on epäuskottavaa, että otos olisi poimittu sellaisesta populaatiosta kuin nollahypoteesi väittää. Otoksen antaman informaation perusteella voidaan pitää uskottavana, että otos on peräisin jostakin muusta kuin nollahypoteesin mukaisesta populaatiosta.

Jos havaittu merkitsevyystaso on α ja testisuureen havaittu arvo on z_{hav} , voidaan päätössäännöt kaksisuuntaisessa testissä

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

muotoilla yleisesti:

Kaksisuuntaisen testin päätössäännöt (I):

$$\begin{aligned} \text{jos } z_{\text{hav}} < -z_{\alpha/2} \text{ tai } z_{\text{hav}} > z_{\alpha/2}, H_0 \text{ hylätään,} \\ \text{jos } -z_{\alpha/2} \leq z_{\text{hav}} \leq z_{\alpha/2}, H_0 \text{ hyväksytään.} \end{aligned}$$

Päätössääntöä voi havainnollistaa seuraavalla kuviolla.

Kuvio 51: Hyväksymis- ja hylkäämisalueet kaksisuuntaisessa testissä

Hyväksymisalue: H_0 hyväksytään

$$P(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) = \phi(z_{\alpha/2}) - \phi(-z_{\alpha/2}) = 1 - \alpha$$

Hylkäämisalue:

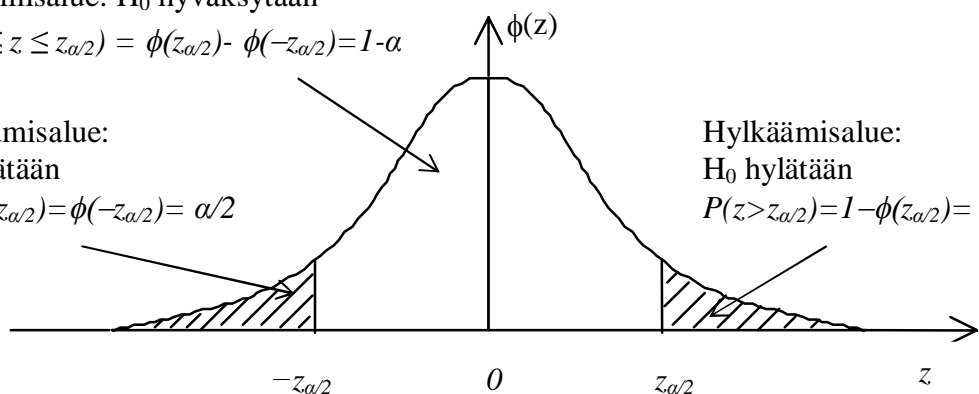
H_0 hylätään

$$P(z < -z_{\alpha/2}) = \phi(-z_{\alpha/2}) = \alpha/2$$

Hylkäämisalue:

H_0 hylätään

$$P(z > z_{\alpha/2}) = 1 - \phi(z_{\alpha/2}) = \alpha/2$$



Kuvioon merkitty hyväksymisalue edustaa niitä testisuureen z havaittuja arvoja, jotka päätössäännön mukaan johtavat nollahypoteesin hyväksymiseen ja hylkäämisalueet niitä testisuureen z havaittuja arvoja, joilla nollahypoteesi hylätään. Kriittiset arvot rajaavat testisuureen jakauman molempiin häntiin todennäköisyysmassaltaan yhteensä α :n suuruisen hylkäämisalueen. Koska testisuureen z jakauma on normaali, tiheysfunktio on symmetrinen ja kummankin häntäpäähän hylkäämisalue on yhtä suuri eli $\alpha/2$. Hylkäämisalueen rajat eli kriittiset arvot määräytyvät siis siten, että todennäköisyys, että testisuureen arvo on kriittistä arvoa kauempana nolasta, on α , ts.

$$\begin{aligned} P(|z| > z_{\alpha/2}) &= P(z < -z_{\alpha/2}) + P(z > z_{\alpha/2}) = 2P(z > z_{\alpha/2}) \\ &= 2(1 - \phi(z_{\alpha/2})) = \alpha. \end{aligned}$$

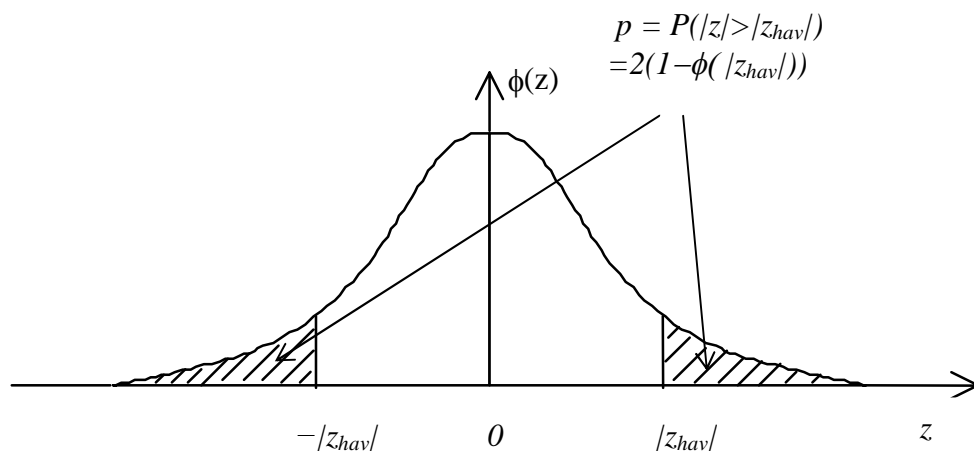
Edellä johtopäätökset tehtiin kriittiseen arvon perusteella. Toinen, vaihtoehtoinen tapa on tehdä johtopäätökset p-arvon nojalla.

Testisuureen havaitun arvon p-arvo eli havaittu merkitsevyystaso on todennäköisyys

$$p = P(|z| \geq |z_{hav}|) = 2(1 - \phi(|z_{hav}|)),$$

eli todennäköisyys, että testisuureen arvo on nolasta eli odotusarvosta (kumpaan hyvänsä suuntaan) kauempana kuin havaittu arvo. Seuraava kuvio havainnollistaa p-arvoa.

Kuvio 52: Havaittu merkitsevyystaso eli p-arvo kaksisuuntaisessa testissä.



Jos testin merkitsevyystaso on α , päätössäännöt ovat seuraavat:

Kaksisuuntaisen testin päätössäännöt (II):

jos $p < \alpha$, H_0 hylätään

jos $p \geq \alpha$, H_0 hyväksytään.

Esimerkin 50 tapauksessa $|z_{hav}| = z_{hav} = 2.108$, joten p-arvoksi saadaan

$$p = P(|z| \geq 2.11) = 2(1 - \phi(2.11)) = 2(1 - 0.9826) = 0.0348.$$

Jos merkitsevyystasona on edelleen 0.05, on $p = 0.0348 < 0.05$, joten päätössäännön mukaan nollahypoteesi hylätään.

Huomaa, että annetulla merkitsevyystasolla sekä kriittisen arvon että p-arvon lähestymistavat tuottavat sisällöltään täsmälleen saman päätössäännön. Testin perusteella tehtävät johtopäätökset ovat siis samat, käytettiinpä kumpaa lähestymistapaa hyvänsä.

Yksisuuntainen testaus

Edellä esitetyt päätössäännöt soveltuvat vain kaksisuuntaiseen testaukseen. Yksisuuntaisessa testauksessa säännöt ovat hieman erilaiset ja merkitsevyystason α kriittistä arvoa merkitään nyt z_α :lla. Esimerkin 50 yhteydessä esiteltiin mahdolliset yksisuuntaiset hypoteesit

$$\begin{array}{ll} (1) & H_0: \mu = 1 \text{ kg} \\ & H_1: \mu < 1 \text{ kg} \end{array} \qquad \begin{array}{ll} (2) & H_0: \mu = 1 \text{ kg} \\ & H_1: \mu > 1 \text{ kg}. \end{array}$$

Merkitään seuraavassa nollahypoteesissa oletettua parametriarvoa yleisemmin μ_0 :lla. Testisuure yksisuuntaisessa testauksessa on aivan sama kuin kaksisuuntais-

$$\text{sessä, eli } z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}.$$

Yksisuuntaisessa testauksessa tarkastellaan poikkeamaa populaation keskiarvosta vain toiseen suuntaan. Testissä (1) riittävän suuret poikkeamat populaation keskiarvosta *alaspäin*, eli populaation keskiarvoa riittävän paljon *pienemmät* otoskeskiarvon arvot, tulkitaan nollahypoteesin vastaisiksi. Testin kriittinen arvo z_α (merkitsevyystasolla α) määräytyy säännöstä

$$P(z < z_\alpha) = \Phi(z_\alpha) = \alpha,$$

eli todennäköisyys, että testisuure saa z_α :aa pienemmän arvon, on α .

Testissä (2) riittävän suuret poikkeamat populaation keskiarvosta *ylöspäin* eli populaation keskiarvoa riittävän paljon *suuremmat* otoskeskiarvon arvot edustavat nollahypoteesin vastaista todistetta. Merkitsevyystason α kriittinen arvo z_α , määräytyy säännöstä

$$P(z > z_\alpha) = 1 - P(z \leq z_\alpha) = \Phi(z_\alpha) = \alpha,$$

eli todennäköisyys, että testisuure saa suuremman arvon kuin z_α on α .

Päätössäännöt yksisuuntaisissa testeissä (1) ja (2) merkitsevyystasolla α kriittisten arvojen avulla lausuttuna ovat seuraavat:

Yksisuuntaisen testin päätössäännöt (I):**Testi (1)**

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Testi (2)

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

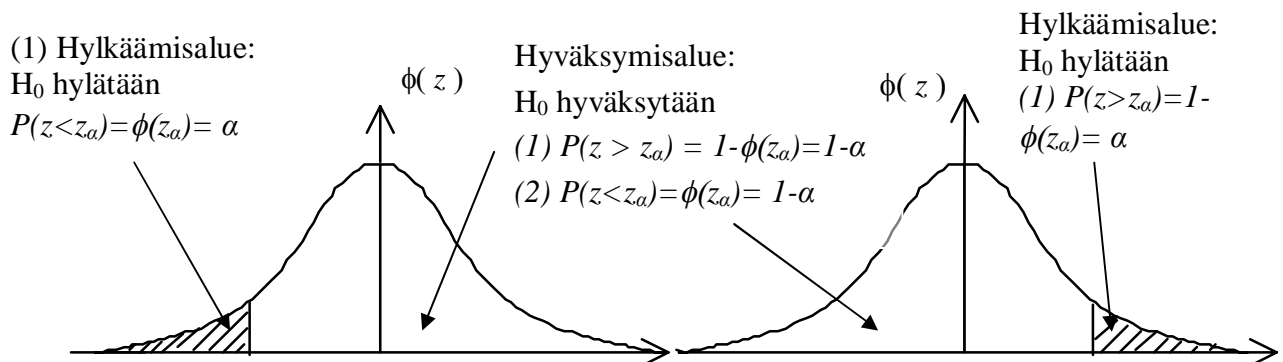
jos $z_{\text{hav}} < z_\alpha$, H_0 hylätään

jos $z_{\text{hav}} \geq z_\alpha$, H_0 hyväksytään

jos $z_{\text{hav}} > z_\alpha$, H_0 hylätään

jos $z_{\text{hav}} \leq z_\alpha$, H_0 hyväksytään

Kuvio 53: Hyväksymis- ja hylkäämisalueet yksisuuntaisessa testissä.



$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Yksisuuntaisessa testissä hylkäämisalue sijaitsee testisuureen z tiheysfunktion toisessa häntäpäässä. Kummassakin tapauksessa hylkäämisalue käsittää kriittisen arvon z_α rajaaman todennäköisyysmassaltaan α :n suuruisen alueen.

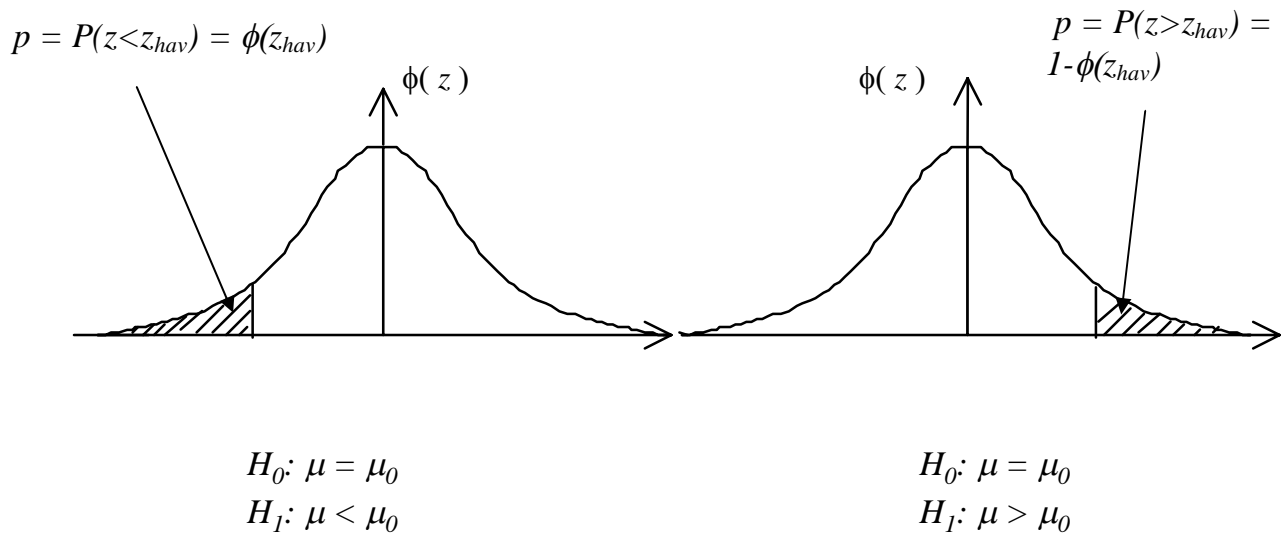
Havaittu merkitsevyystaso yksisuuntaisissa testeissä on testin (1) tapauksessa

$$p = P(z \leq z_{\text{hav}}) = \phi(z_{\text{hav}})$$

ja testin (2) tapauksessa

$$p = P(z \geq z_{\text{hav}}) = 1 - \phi(z_{\text{hav}}).$$

Kuvio 54: Havaittu merkitsevyystaso eli p -arvo yksisuuntaisessa testissä.



Merkitsevyystasolla α p -arvon avulla lausuttu päätössääntö on nyt kummassakin testissä sama kuin kaksisuuntaisen testin kohdalla, eli:

Yksisuuntaisen testin päätössäännöt (II):

jos $p < \alpha$, H_0 hylätään
 jos $p \geq \alpha$, H_0 hyväksytään.

Huomaa, että yksi- ja kaksisuuntaisen testin välinen ero tulee tehdyksi jo itse p -arvoa laskettaessa.

Päätössäännöt on tässä esitelty keskiarvotestin yhteydessä, mutta säännöt soveltuvat yleisesti kaikkiin testeihin, joissa testisuureen jakauma on *symmetrinen* (normaalijakauma tai t -jakauma). Eri testeissä käytettävät testisuureet ovat erilaisia ja voivat noudattaa jotakin muuta kuin normaalijakaumaa. Tällöin on testisuureen kriittinen arvo tai p -arvo etsittävä asianomaisen jakauman taulukosta. Kun testisuureen havaittu arvo ja valitun merkitsevyystason kriittinen arvo tai vaihtoehtoisesti havaittu merkitsevyystaso on määritetty, voi edellä esiteltyjä päätössääntöjä soveltaa sellaisenaan.

Epäsymmetristen jakaumien (jäljempänä käsiteltävät F - ja χ^2 -jakaumat) tapauksessa yksisuuntaisen testin päätössäännöt ovat analogiset edellä esitettyjen kanssa. Kaksisuuntaisen testin yhteydessä säännöt muotoillaan ja p -arvo lasketaan hieman eri tavalla. Tähän palataan tarkemmin kyseisten testien yhteydessä.

Jäljempänä käsiteltävissä testeissä ei enää johdeta yksityiskohtaisesti testisuureita. Periaatteet, joilla testisuure muodostetaan ja sen jakauma selvitetään, ovat kuitenkin samat kuin edellä on esitelty. Lähtökohtana tyypillisesti on tarkasteltavan tunnusluvun estimaattori ja sen otantajakauma, joiden avulla päätellään, kuinka suuri poikkeama oletetusta arvosta on uskottavuuden rajoissa. Testisuureiden ja niiden jakaumien johtamisesta löytyy lisätietoa kirjallisuudesta (ks. lähdeluettelo monisteen lopussa).

Esitetään vielä seuraavassa yhteenvedona hypoteesin testauksen vaiheet pääpiirteissään.

Hypoteesin testauksen vaiheet

1. Määritä nollahypoteesi ja tilanteeseen soveltuva vastahypoteesi.
2. Valitse käytettävä testi ja sen mukainen testisuure.
3. Valitse testissä käytettävä merkitsevyystaso α .
4. Muodosta valitusta merkitsevyystasosta riippuva päätössääntö, joka kertoo, millä testisuureen arvoilla nollahypoteesi hylätään. Jos päätössääntö muotoillaan testisuureen *kriittisen arvon* avulla, on kriittinen arvo luonnollisesti määritettävä tässä.
5. Kerää otoksesta mitattu data ja laske sen avulla testisuureen arvo (eli määritä testisuureen *havaittu arvo*).
6. **joko:** Käytä kohdan 4. päätössääntöä eli vertaa testisuureen havaittua ja kriittistä arvoa toisiinsa ja tämän perusteella joko hyväksy tai hylkää nollahypoteesi
- tai** (jos käytät mieluummin p-arvoa): Määritä testisuureen havaitun arvon p-arvo eli *havaittu merkitsevyystaso*. Käytä kohdan 4. päätössääntöä eli vertaa havaittua merkitsevyystasoa kohdassa 3 valittuun merkitsevyystasoon ja tämän perusteella joko hyväksy tai hylkää nollahypoteesi.
7. Tulkitse testauksen tulokset selväkielisesti.

2.3.2 Testisuureiden jakaumista

Tällä kurssilla käsitellään normaalisti jakautuneiden testisuureiden lisäksi t -, F - ja χ^2 -jakautuneita testisuureita. Tarkastellaan seuraavassa lyhyesti näiden jakaumien ominaisuuksia.

Normaalisti jakautuneet testisuureet ovat valmiiksi standardoituja, eli ne noudattavat standardoitua normaalijakaumaa $N(0,1)$. Näiden kohdalla parametreja (standardoimattoman muuttujan jakauman parametrin) tarvitaan vain testisuureen arvon laskemisessa. Myös muut, eli t -, F - ja χ^2 -jakaumat, riippuvat yhdestä tai useammasta parametrista, joita yleisesti sanotaan ko. jakauman *vapausasteiksi*. Kuten normaalijakauman tapauksessa, kullakin eri parametrin eli vapausasteen arvolla saadaan oma jakaumansa tarkasteltavasta jakaumatyypistä. Täten testisuureen jakauma tunnetaan täysin vasta sitten, kun sen vapausaste/vapausasteet on määritetty. Vapausastelukua merkitään usein symboleilla k, f tai df .

t-jakauma (*Studentin jakauma*)

Aiemmin todettiin, että keskeisen raja-arvolauseen nojalla otoskeskiarvon otanta-jakauma on likimain normaali, mikäli otoskoko n on riittävän suuri. Standardoitu otoskeskiarvo

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

noudattaa siis standardoitua normaalijakaumaa. Jos populaation hajontaa ei tunneta, se estimoidaan otoskeskihajonnalla, ja jos populaation jakauma voidaan olettaa normaaliksi, suuren otoksen tapauksessa myös standardoitu otoskeskiarvo

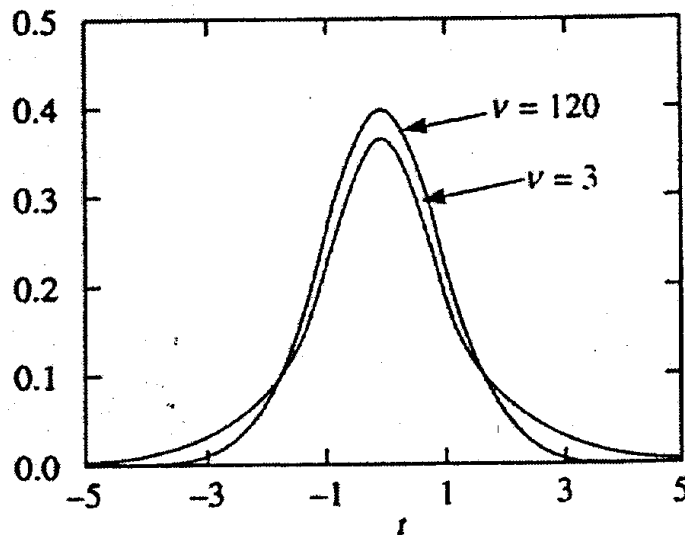
$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

on likimain $N(0,1)$ -jakautunut. Jos otoskoko on pieni, ei keskeinen raja-arvolause sovellu, ja tässä tapauksessa kyseinen suure onkin *t-jakautunut vapausasteena $n-1$* , missä n on otoskoko, ja merkitään $t \sim t(n-1)$, missä

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}.$$

T-jakauma muistuttaa muodoltaan normaalijakaumaa, eli sen tiheysfunktion kuvaaja on “kellokäyrä”. Samoin t-jakauman tiheysfunktion kuvaaja, millä hyvänsä vapausasteella k on symmetrinen nollan suhteen. Kuvaaja on tyypillisesti leveämpi ja litteämpi kuin standardoidun normaalijakauman. Tämä merkitsee, että t-jakauman varianssi on suurempi kuin standardoidun normaalijakauman varianssi, olipa k mikä hyvänsä. Jakauman täsmällinen muoto riippuu vapausasteluvusta: mitä suurempi k (eli otoskeskiarvon tapauksessa mitä suurempi otoskoko), sitä lähempänä standardoitua normaalijakaumaa t-jakauma on.

Kuvio 55 (Neter & al. s.910): Studentin t-jakauman tiheysfunktion kuvaaja, kun vapausaste ν on i) $\nu = 3$ ja ii) $\nu = 120$.



χ^2 -jakauma

χ^2 -jakaumaa käytetään mm. varianssi- ja riippumattomuustesteissä. Se on yhdestä parametrasta riippuva jakauma, jonka tiheysfunktion kuvaaja on oikealle vino. Jakauma on sitä vinompi, mitä pienempi vapausasteluku on.

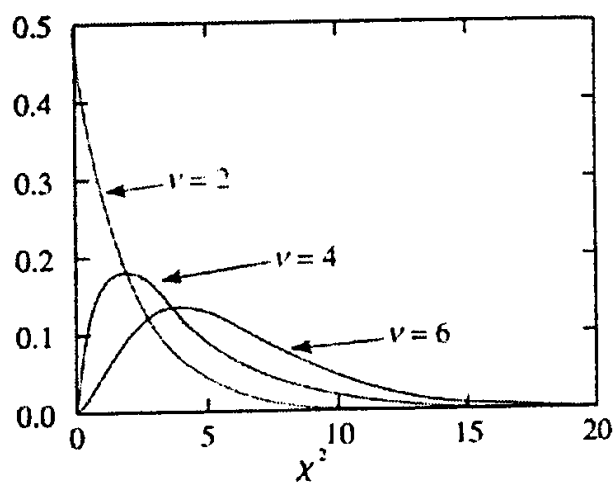
χ^2 -jakautunut suure voi saada vain ei-negatiivisia arvoja. Jos tarkasteltava satunnaismuuttuja (esim. testisuure) χ^2 on χ^2 -jakautunut vapausasteena k , merkitään $\chi^2 \sim \chi^2(k)$, saadaan sen odotusarvo kaavasta

$$E\chi^2 = k$$

ja varianssi kaavasta

$$D^2\chi^2 = 2k.$$

Kuvio 56 (Neter & al. s.910): χ^2 -jakauman tiheysfunktion kuvaaja, kun vapausaste ν on i) $\nu = 2$, ii) $\nu = 4$ ja iii) $\nu = 6$.



F-jakauma (Fisherin jakauma)

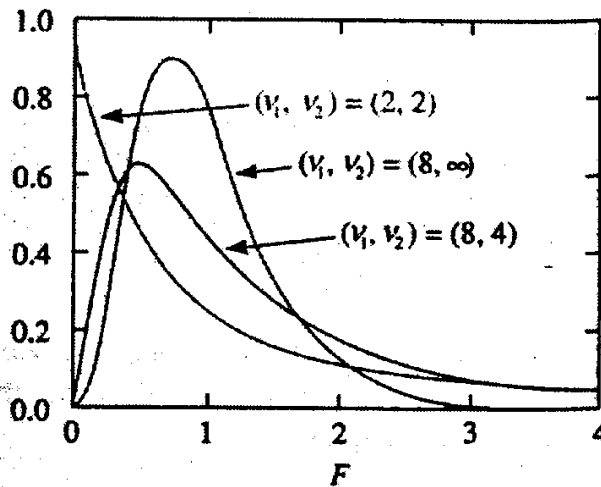
Kuten χ^2 -jakauma, F-jakaumakin on oikealle vino, ja F-jakautunut suure saa vain ei-negatiivisia arvoja. F-jakauma on kahdesta parametrasta riippuva, ja vapausasteilla k_1 ja k_2 F-jakautuneelle satunnaismuuttujalle F käytetään merkintää $F \sim F(k_1, k_2)$.

Eräs monesti tarpeellinen F-jakauman ominaisuus on seuraava: jos satunnaismuuttuja X on F-jakautunut vapausasteilla k_1 ja k_2 , eli $X \sim F(k_1, k_2)$, niin satunnaismuuttuja $Y = \frac{1}{X}$ on F-jakautunut vapausasteilla k_2 ja k_1 , ts.

$$Y = \frac{1}{X} \sim F(k_2, k_1).$$

F-jakaumaa käytetään mm. kahden jakauman varianssien vertailussa ja regressioanalyysin yhteydessä.

Kuvio 57 (Neter & al. s.910): *F*-jakauman tiheysfunktion kuvaaja, kun vapausasteet (v_1, v_2) ovat i) $(v_1, v_2) = (2, 2)$, ii) $(v_1, v_2) = (8, 4)$, ja iii) $(v_1, v_2) = (8, \infty)$.



2.3.3 Keskiarvotestit

Keskiarvotesteissä tehdään johtopäätöksiä populaation keskiarvosta. Tässä monisteessa tarkasteltavat testit voidaan jakaa kahteen ryhmään:

- 1) yhden otoksen testit, joissa testataan populaation keskiarvoa yhden otoksen perusteella
- 2) kahden otoksen testit, joissa testataan kahden keskiarvon yhtäsuuruutta.

Keskiarvotestejä voidaan tehdä vähintään välimatka-asteikon tasoisille muuttujille.

Yhden otoksen keskiarvotestit

Yhden otoksen keskiarvotesteissä testataan, onko populaation keskiarvo μ väitetyn suuruinen, eli nollahypoteesi on muotoa $\mu = \mu_0$. Vastahypoteesin muotoilu riippuu siitä, onko testaus yksi- vai kaksisuuntainen. Hypoteesit voivat siis olla seuraavanlaiset:

kaksisuuntainen testaus	yksisuuntainen testaus	
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	tai $H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$

Sopivan testin valintaan vaikuttavat:

- 1) otoskoko n : onko otos pieni ($n < 30$) vai suuri ($n \geq 30$) ?
- 2) mikä on muuttujan X jakauma populaatiossa (eli mikä on *populaation jakauma*): onko se normaali vai ei ?
- 3) tunnetaanko σ eli X :n keskihajonta perusjoukossa (eli tunnetaanko *populaation hajonta*) vai ei ?

Suuret otokset

Jos otoksen koko on riittävän suuri (sääntönä voidaan pitää kokoa $n \geq 30$), on keskeisen raja-arvolauseen nojalla X :n otoskeskiarvon \bar{X} otantajakauma ainakin likimain normaali, oli X :n jakauma populaatiossa mikä hyvänsä. Jos X :n odotusarvo on $\mu = \mu_0$ ja keskihajonta σ , testisuure

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

noudattaa (ainakin likimain) standardoitua normaalijakaumaa, ts. $z \sim N(0,1)$.

Kun testisuure on $N(0,1)$ -jakautunut, sanotaan, että *suoritetaan z -testi*. Z -testissä testisuureen kriittinen arvo (tai vaihtoehtoisesti havaittu p -arvo) etsitään standardoidun normaalijakauman taulukosta.

Jos testi johtaa nollahypoteesin hylkäämiseen, voidaan populaation keskiarvolle laskea otoksen perusteella estimaatti. Piste-estimaatiksi käy otoskeskiarvo \bar{X} ja $100(1-\alpha) \%$:n luottamusväliksi

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

missä $z_{\alpha/2}$ toteuttaa ehdon $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Jos populaation hajontaa σ ei tunneta, sen estimaattina käytetään otoskeskihajontaa s . Tällöin testisuure on muotoa

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}},$$

ja suurten otosten tapauksessa se on likimain normaalisti jakautunut, ts. $z \sim N(0,1)$. Jos nollahypoteesi hylätään, $100(1-\alpha) \%$:n luottamusväli populaation keskiarvolle μ on

$$(\bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}}),$$

ja piste-estimaatti on otoskeskiarvo \bar{X} .

Yhteenvedona edellisestä voidaan todeta, että jos otoskoko on riittävän suuri ($n \geq 30$), niin yhden otoksen keskiarvotestiksi voidaan aina valita z-testi.

Pienet otokset

Pienten otosten ($n < 30$) tapauksessa on merkityksellistä se, onko muuttujan X jakauma populaatiossa normaali vai ei. Jos populaation hajonta on tunnettu ja populaation jakauma on normaali, $X \sim N(\mu, \sigma^2)$, on lauseen 45 nojalla \bar{X} normaalisti jakautunut, $\bar{X} \sim N(\mu, \sigma^2/n)$. Jos $\mu = \mu_0$, on testisuure

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

Tällöin siis kysymykseen tulee z-testi.

Jos populaation hajonta on tuntematon, ja populaatio on ainakin likimain normaalisti jakautunut, noudattaa testisuure t-jakaumaa vapausasteella $n-1$. Nyt testisuureta merkitään

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t(n-1),$$

missä s on otoksesta estimoitu keskihajonta, ja sanotaan, että *suoritetaan t-testi*. T-testissä testisuureen kriittinen arvo etsitään t-jakauman taulukosta. Taulukossa on esitettyinä kriittisiä arvoja eri vapausasteilla yleisimmin käytetyille merkitsevyystasojen. Jos testissä halutaan mieluummin käyttää p-arvoa, sille saa taulukosta useimmissa tapauksissa vain karkean arvion. Tilasto-ohjelmat laskevat tarkan p-arvon t-jakauman kertymäfunktioista.

Populaation keskiarvon μ $100(1-\alpha) \%$:n luottamusväli on nyt

$$(\bar{X} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}),$$

missä $t_{\alpha/2}(n-1)$ toteuttaa ehdon $P(t(n-1) > t_{\alpha/2}(n-1)) = \alpha/2$. Piste-estimaatti μ :lle on populaation keskiarvo \bar{X} .

Jos populaation hajontaa eikä jakaumaa tunneta, sekä t- että z-testi ovat epäluotettavia, eikä niitä siis pidä käyttää.

Seuraavassa taulukossa esitetään yhteenveto yhden otoksen keskiarvotesteistä.

σ tunnettu, $\mu = \mu_0$:	$n < 30$	$n \geq 30$
X:n jakauma populaatiossa on normaali	$\bar{X} \sim N(\mu, \sigma^2/n)$ $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ z-testi	$\bar{X} \sim N(\mu, \sigma^2/n)$ $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ z-testi
X:n jakauma populaatiossa ei ole normaali	-	$\bar{X} \sim N(\mu, \sigma^2/n)$ $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ (likimäärin) z-testi
σ tuntematon, $\mu = \mu_0$:	$n < 30$	$n \geq 30$
X:n jakauma populaatiossa on normaali	$\bar{X} \sim N(\mu, s^2/n)$ $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$ (likimäärin) t-testi	$\bar{X} \sim N(\mu, s^2/n)$ $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$ (likimäärin) z-testi
X:n jakauma populaatiossa ei ole normaali	-	$\bar{X} \sim N(\mu, s^2/n)$ $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$ (likimäärin) z-testi

HUOM! Tässä käytetty suuren ja pienen otoksen kokoraja 30 on ainoastaan ohjeellinen. Tätä voidaan pitää ”turvallisena” rajana otoskoolle, jotta otoskeskiarvo olisi likimäärin normaalisti jakautunut. Jos populaation jakauma on ”siistin”

muotoinen (yksihiippuinen, symmetrinen), niin otoskeskiarvon jakauma lähestyy normaalia jo huomattavasti pienemmällä otoskoollla (ks. kuvio 46).

Tehtävä 60 (*Karjalainen-Ruuskanen s. 155*):

Erään oppilaitoksen pääsykokeisiin osallistui 852 henkilöä. Kokeessa saadut pistemäärät noudattivat likimain normaalijakaumaa, siten että $\mu = 84.1$ ja $\sigma = 7.9$. Osa pyrkijöistä oli osallistunut preppauskursseille ja kurssin pitäjä halusi tietää, poikkesiko kurssille osallistuneiden pistemäärä kaikkien pyrkijöiden pistemäärästä. 30 satunnaisesti valitun kurssille osallistuneen pistemäärien keskiarvo oli 86.2. Testaa 5 % merkitsevyystasolla, oliko kurssin käyneiden pistemäärä parempi.

Tehtävä 61 (*Karjalainen-Ruuskanen s. 156*):

Urheiluvälinetehdas ilmoittaa valmistamiensa 41-numeroisten kenkien painoksi 160 g. Asiaa tarkistettaessa saatiin 16 kengän otoksesta keskiarvoksi 164 g ja keskihajonnaksi 6 g. Testaa tehtaan ilmoittaman keskipainon paikkansapitävyyttä 1% merkitsevyystasolla (kun oletetaan, että kenkien painon jakauma populaatiossa on ainakin likimain normaali).

Kahden otoksen keskiarvotestit

Tässä käsiteltävät kahden otoksen keskiarvotestit voidaan jakaa kahteen ryhmään sen perusteella, ovatko poimitut otokset toisistaan riippumattomia vai eivät. *Kahden riippumattoman otoksen* testeissä päättely koskee kahden populaation jakauman sijainnin vertailua. Kummastakin populaatiosta poimitaan (toisistaan riippumatta) otos ja niiden nojalla päätellään, ovatko populaatioiden keskiarvot samat. Kahteen toisistaan riippuvan otokseen perustuva testaustilanne esiintyy esim. kokeellisessa tutkimuksessa, jossa halutaan verrata kahden erilaisen käsittelyn tai menetelmän vaikutuksia tilastoyksikköjen johonkin ominaisuuteen. Tällöin puhutaan *parittaisesta t-testistä*.

1. Kaksi riippumatonta otosta

Tarkastellaan kahta populaatiota, joissa muuttujan X keskiarvoa ja keskihajontaa merkitään seuraavasti:

	keskiarvo	keskihajonta
populaatio 1	μ_1	σ_1
populaatio 2	μ_2	σ_2

Testauksen kohteena on se, eroavatko populaatioiden keskiarvot μ_1 ja μ_2 toisistaan vai ovatko ne samat. Testausta varten poimitaan populaatiosta 1 n_1 suuruisen ja populaatiosta 2 n_2 suuruisen otos. Merkitään näistä otoksista laskettuja muuttujan X keskiarvoa ja keskihajontaa seuraavasti:

	otos- koko	otos- keskiarvo	otos- keskihajonta
populaatio 1	n_1	\bar{X}_1	s_1
populaatio 2	n_2	\bar{X}_2	s_2

Testaus voidaan suorittaa yksi- tai kaksisuuntaisena. Hypoteesit ovat tällöin seuraavaa muotoa:

<i>kaksisuuntainen testaus</i>	<i>yksisuuntainen testaus</i>	
$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_2$	tai $H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$	$H_1: \mu_1 < \mu_2$	$H_1: \mu_1 > \mu_2$

Nollahypoteesissa esiintyvä väittämä populaatioiden keskiarvojen yhtäsuuruudesta, $\mu_1 = \mu_2$ voidaan myös ilmaista muodossa $\mu_1 - \mu_2 = 0$.

Johtopäätökset populaatiojakaumien sijainnista perustuvat nyt otoskeskiarvojen erotukseen $\bar{X}_1 - \bar{X}_2$: mitä pienempi (lähempänä nollaa) erotus on, sitä uskottavampaa on, että populaatiokeskiarvot ovat yhtä suuria. Uskottavuusrajojen tarkempi määrittäminen perustuu taas erotuksen $\bar{X}_1 - \bar{X}_2$ otantajakaumaan. Kuten yhden otoksen keskiarvotesteissä, tähän vaikuttavat otoskoot n_1 ja n_2 , muuttujan X jakauma kummassakin populaatiossa sekä se, tunnetaanko populaatioiden hajonnat σ_1 ja σ_2 vai ei. Jos kumpikin otoksista on riittävän suuri, on erotuksen $\bar{X}_1 - \bar{X}_2$ jakauma ainakin likimain normaali ja testinä voidaan käyttää z-testiä. Otoskoon ”riittävä” suuruus riippuu tapauksesta, joita käsitellään tarkemmin seuraavassa.

Kaksi riippumatonta otosta: populaatioiden hajonnat tunnettuja

Jos populaatioiden 1 ja 2 jakaumat ovat normaaleja $N(\mu_1, \sigma_1^2)$ ja $N(\mu_2, \sigma_2^2)$, niin voidaan osoittaa, että otoskeskiarvojen erotus $\bar{X}_1 - \bar{X}_2$ on normaalisti jakautunut odotusarvona

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

ja varianssina

$$D^2(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Mikäli otokset ovat peräisin nollahypoteesin mukaisista populaatioista, on satunnaismuuttujan $\bar{X}_1 - \bar{X}_2$ odotusarvo $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 = 0$, joten testisuure eli erotuksen $\bar{X}_1 - \bar{X}_2$ standardoitu muuttuja on

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1),$$

eli käytetään z-testiä. Jos populaation jakauma ei ole normaali, suurten otosten ($n_1 \geq 30$ ja $n_2 \geq 30$) tapauksessa testisuure on kuitenkin likimain $N(0,1)$ -jakautunut, joten z-testi soveltuu myös tähän tapaukseen.

Populaatiokeskiarvojen erotuksen $\mu_1 - \mu_2$ piste-estimaatti on otoskeskiarvojen erotus $\bar{X}_1 - \bar{X}_2$, ja $100(1-\alpha)$ %:n luottamusväli on

$$(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

missä $z_{\alpha/2}$ toteuttaa ehdon $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Tehtävä 62 (*Liski-Puntanen (II) ss. 29, 66*): Liukuhihnalta A tulee ruuveja, joiden läpimitta pitäisi olla sama kuin liukuhihnalta B tulevien ruuvien. Tiedetään, että A- ja B-ruuvien läpimitan hajonnat ovat kumpikin 8. Kummaltakin hihnalta valittiin 25 ruuvia, joiden läpimitan keskiarvot olivat A-hihnalla 19 ja B-hihnalla 22.

a) Testaa läpimittojen keskiarvojen yhtäsuuruusoletusta 5 % merkitsevyystasolla.

b) Mikä on pienin merkitsevyystaso, jolla nollahypoteesi voidaan hylätä kaksisuuntaisessa testissä $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$?

Jos populaatioiden hajontoja ei tunneta, ne on estimoitava otoskeskihajontoja apuna käyttämällä. Estimointiin vaikuttaa se, voidaanko tuntemattomia hajontoja pitää yhtä suurina vai ei. Jos hajontojen yhtäsuuruudesta ei ole tietoa, yhtäsuuruus voidaan testata jäljempänä käsiteltävällä F-testillä.

Kaksi riippumatonta otosta: populaatioiden hajonnat tuntemattomat ja yhtäsuuret

Jos populaatioiden hajonnat σ_1 ja σ_2 ovat yhtä suuret, voidaan merkitä $\sigma_1 = \sigma_2 = \sigma$. Tuntemattoman yhteisen varianssin σ^2 estimaattorina käytetään otosvarianssien s_1^2 ja s_2^2 painotettua keskiarvoa s^2

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Populaatioiden jakaumien ollessa normaaleja, $N(\mu_1, \sigma_1^2)$ ja $N(\mu_2, \sigma_2^2)$, nollahypoteesin testaamiseen voidaan käyttää t-testiä. Mikäli otokset ovat peräisin nollahypoteesin mukaisista populaatioista, testisuure on

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

missä

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}},$$

eli testisuure t noudattaa t-jakaumaa vapausasteella $n_1 + n_2 - 2$.

Erotuksen $\mu_1 - \mu_2$ $100(1 - \alpha)\%$:n luottamusväli on nyt

$$(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}(n_1 + n_2 - 2) s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}(n_1 + n_2 - 2) s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}),$$

missä $t_{\alpha/2}(n_1 + n_2 - 2)$ toteuttaa ehdon $P(t(n_1 + n_2 - 2) > t_{\alpha/2}(n_1 + n_2 - 2)) = \alpha/2$. Pisteestimaatiksi käy, kuten edellä, otoskeskiarvojen erotus $\bar{X}_1 - \bar{X}_2$.

Suurten otosten tapauksessa ($n_1 \geq 30$ ja $n_2 \geq 30$) edellämainittu testisuure on likimain $N(0,1)$ -jakautunut, vaikka populaatioiden jakaumat eivät olisikaan normaaleja. Siis jos otoskoot ovat riittävän suuret, voidaan käyttää z-testiä populaatiojakaumista riippumatta.

Tehtävä 63 (*Karjalainen-Ruuskanen s. 157*): Verrattaessa poikien ja tyttöjen menestymistä eräässä testissä saatiin 13 tytön otoksesta testipisteiden keskiarvoksi 32.62 ja keskihajonnaksi 6.54 sekä 14 pojan otoksesta keskiarvoksi 27.93 ja keskihajonnaksi 5.89. Oletetaan testipisteiden jakauma sekä poikien että tyttöjen populaatiossa normaaliksi ja pistemäärien varianssit populaatioissa yhtä suuriksi. Testaa 5 % merkitsevyystasolla, onko poikien ja tyttöjen välillä eroa testissä menestymisen suhteen.

Kaksi riippumatonta otosta: populaatioiden hajonnat tuntemattomat ja erisuuret

Kun populaatioiden hajonnat ovat tuntemattomat, ja niitä ei perustellusti voi pitää yhtäsuurina, käytetään populaation varianssien σ_1^2 ja σ_2^2 estimaatteina vastaavia otosvariانسseja s_1^2 ja s_2^2 . Tässä tapauksessa ei normaalistikaan jakautuneiden populaatioiden kohdalla voida testisuurelle johtaa tarkkaa jakaumaa. Jos otokset ovat peräisin nollahypoteesin mukaisista populaatioista, testisuure on muotoa

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

ja suurten otosten tapauksessa ($n_1 \geq 50$ ja $n_2 \geq 50$) se on likimain $N(0,1)$ -jakautunut (z-testi). Jos otoskoot ovat kohtalaisen suuria ($n_1 \geq 20$ ja $n_2 \geq 20$), ja populaatioiden jakaumat normaaleja, noudattaa testisuure likimain t-jakaumaa vapausasteena f (t-testi). Vapausasteen f määrittämiseen käytetään ns. Welchin menetelmää:

$$\frac{1}{f} = \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1},$$

missä

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

T-testin tapauksessa erotuksen $\mu_1 - \mu_2$ $100(1-\alpha)$ %:n luottamusväli on

$$(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}(f) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}(f) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}),$$

missä $t_{\alpha/2}(f)$ toteuttaa ehdon $P(t(f) > t_{\alpha/2}(f)) = \alpha/2$. Jos testinä on z-testi, em. luottamusvälin kaavassa kriittisen arvon $t_{\alpha/2}(f)$ sijaan käytetään kriittistä arvoa $z_{\alpha/2}$, jolle on voimassa $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Piste-estimaattina toimii kummassakin tapauksessa otoskeskiarvojen erotus $\bar{X}_1 - \bar{X}_2$.

2. Kaksi riippuvaa otosta: parittainen t-testi

Tarkastellaan koejärjestelyä, jossa halutaan verrata kahden eri menetelmän tai käsittelyn vaikutusta tilastoyksiköiden johonkin ominaisuuteen. Kokeen suorittamiseen tarvitaan kaksi tilastoyksikköjen ryhmää, joista toiseen sovelletaan käsittelyä A ja toiseen käsittelyä B. Käsittelyvaikutuksen selville saamiseksi on tärkeää eristää muiden samaan ominaisuuteen vaikuttavien muuttujien vaikutus kokeen aikana.

Jos koeryhmät muodostetaan jakamalla käytettävissä olevien tilastoyksiköiden joukko satunnaisesti kahteen ryhmään, voidaan tulosten analysointiin käyttää riippumattomien otosten keskiarvotestiä. Nyt ongelmaksi voi kuitenkin muodostua tilastoyksiköiden keskinäinen heterogeenisuus, joka luonnollisista syistä aiheuttaa eri ryhmistä tehtävien havaintojen jakaumiin vaihtelua ja haittaa täten käsittelyvaikutusten mittaamista. Tämänkaltaisten häiriötekijöiden vaikutusta voidaan vähentää kahdella tavalla:

1. Tilastoyksiköistä muodostetaan yhteensovitettuja pareja siten, että kunkin parin yksiköt ovat keskenään mahdollisimman samankaltaisia niiden ominaisuuksiensa suhteen, joilla saattaisi olla vaikutusta tutkittavaan ominaisuuteen. Tämän jälkeen kunkin parin kohdalla toiseen sovelletaan käsittelyä A ja toiseen käsittelyä B.

2. Samoilte tilastoyksiköille suoritetaan kumpikin käsittelyistä A ja B (jokainen tilastoyksikkö on “itse itsensä pari”).

Käsittelyjen erojen selvittämiseksi mitataan tilastoyksiköistä käsittelyvaikutuksen indikaattorina olevan muuttujan arvo ja lasketaan kustakin parista arvojen parittaiset erotukset. Päättely käsittelyjen välisistä eroista perustuu näiden *parittaisten erotusten keskiarvoon*. Jos erotusten keskiarvo esimerkiksi on hyvin pieni (lähellä nollaa), voidaan pitää uskottavana, että käsittelyillä ei ollut eroa.

Olkoon yhteensovitettujen parien havaintoarvot seuraavan taulukon mukaiset:

pari	1	2	...	n
käsittely A	X_{A1}	X_{A2}	...	X_{An}
käsittely B	X_{B1}	X_{B2}	...	X_{Bn}
<hr/>				
erotus	D_1	D_2	...	D_n

Tulkitaan, kuten aiemminkin, havaintoarvot satunnaismuuttujiksi. Oletetaan lisäksi, että parit $(X_{A1}, X_{B1}), (X_{A2}, X_{B2}), \dots, (X_{An}, X_{Bn})$ ovat keskenään riippumattomat (kukin pari on “poimittu” mistään toisesta parista riippumatta) ja että erotukset D_i ovat normaalisti jakautuneet keskiarvona μ_D ja varianssina σ_D^2 . Keskiarvo μ_D kuvaa käsittelyjen vaikutusten keskimääräistä eroa.

Hypoteesit ovat muotoa

<i>kaksisuuntainen testaus</i>	<i>yksisuuntainen testaus</i>	
$H_0: \mu_D = \mu_0$	$H_0: \mu_D = \mu_0$	tai $H_0: \mu_D = \mu_0$
$H_1: \mu_D \neq \mu_0$	$H_1: \mu_D < \mu_0$	$H_1: \mu_D > \mu_0$

Erityisesti, jos nollahypoteesi on muotoa $\mu_D = 0$, testataan väittämää “käsittelyillä A ja B ei ole eroa”.

Mikäli otokset ovat peräisin nollahypoteesia $\mu_D = \mu_0$ noudattavista populaatioista, on testisuure

$$t = \frac{\bar{D} - \mu_0}{\frac{s_D}{\sqrt{n}}} \sim t(n-1),$$

missä \bar{D} on erotusten D_i keskiarvo ja s_D niiden keskihajonta,

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i, \quad s_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2}.$$

Käsittelyjen välisen keskimääräisen eron μ_D $100(1-\alpha)$ %:n luottamusväli on

$$\left(\bar{D} - t_{\alpha/2}(n-1) \frac{s_D}{\sqrt{n}}, \bar{D} + t_{\alpha/2}(n-1) \frac{s_D}{\sqrt{n}} \right),$$

missä $t_{\alpha/2}(n-1)$ toteuttaa ehdon $P(t(n-1) > t_{\alpha/2}(n-1)) = \alpha/2$. Piste-estimaatti on otoksesta laskettu erotusten keskiarvo \bar{D} .

Tehtävä 64: Vaatetusalan yritys harkitsee, onko uuden tuotteen värjäykseen käytettävä menetelmä A vai B. Erityisesti halutaan, että väri ei haalistu auringonvalossa. Kymmenen tuotenäytettä leikattiin kahtia, jonka jälkeen toiseen osaan sovellettiin menetelmää A ja toiseen osaan menetelmää B. Näin saadut 20 näytettä asetettiin tietyksi ajaksi auringonvaloon ja kokeen jälkeen havaittiin seuraavat värin voimakkuudet (pienet arvot tarkoittavat runsasta haalistumista):

Näyte	1	2	3	4	5	6	7	8	9	10
A	7.2	4.3	5.8	6.5	4.9	6.8	6.3	7.0	6.5	6.2
B	5.1	4.1	5.5	4.1	5.0	5.1	5.3	7.3	4.8	5.8

Tutki ongelmaa sopivan testin avulla.

2.3.4 Varianssin testaus

Jakaumien luonnehdinnassa on sijainnin (keskiarvo) ohella myös jakauman hajonta keskeinen ominaisuus. Jakauman hajonnan muodollisessa tarkastelussa käytetään keskihajonnan sijaan mieluummin varianssia.

Keskiarvotesteissä testisuurena käytettiin populaatiokeskiarvon estimaattorin, otoskeskiarvon, standardoitua muuttujaa. Varianssin testauksessa käytettävä tes-

tisuure muodostetaan vastaavasti populaatiovarianssin harhattoman estimaattorin, otosvarianssin pohjalta.

Seuraavassa tarkastellaan lyhyesti jakaumien hajontaa koskevaa päättelyä yhden jakauman varianssin ja kahden jakauman varianssien vertailun tapauksessa.

Yhden otoksen varianssitesti

Yhden otoksen varianssitestissä päättely koskee sitä, onko populaation varianssi σ^2 jonkin ennalta annetun luvun σ_0^2 suuruinen. Populaation tunnusluvut, keskiarvo ja varianssi, ovat nyt molemmat tuntemattomia, ja päättely nojautuu otoksesta laskettuun otosvarianssiin s^2 . Olkoon otoksen koko n . Testisuurena käytetään suuretta

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

jonka voidaan osoittaa olevan χ^2 -jakautunut vapausasteella $n-1$, mikäli otos on poimittu *normaalisti jakautuneesta* populaatiosta.

Hypoteesit ovat muotoa

<i>kaksisuuntainen testaus</i>	<i>yksisuuntainen testaus</i>	
$H_0: \sigma^2 = \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$	tai $H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 \neq \sigma_0^2$	$H_1: \sigma^2 < \sigma_0^2$	$H_1: \sigma^2 > \sigma_0^2$

Testisuureen jakauma on nyt epäsymmetrinen. Merkitään testisuureen havaittua arvoa χ^2_{hav} ja tarkastellaan ensin 2-suuntaista testiä. Merkitsevyystason α kriittistä arvoa merkitään $\chi^2_{\alpha/2}$ ja sille on voimassa $P(\chi^2 > \chi^2_{\alpha/2}) = \alpha/2$, eli kriittinen arvo erottaa jakauman tiheysfunktion ylemmästä hännästä $\alpha/2$ suuruisen alueen. Epäsymmetrisen jakauman tapauksessa se arvo, joka erottaa jakauman tiheysfunktion alemmasta hännästä vastaavansuuruisen alueen ei olekaan kriittisen arvon vastaluku, vaan se täytyy määrittää erikseen. Merkitään tätä arvoa $\chi^2_{1-\alpha/2}$ ja sille on siis voimassa $P(\chi^2 > \chi^2_{1-\alpha/2}) = 1-\alpha/2$, eli $P(\chi^2 < \chi^2_{1-\alpha/2}) = \alpha/2$. Kak-

sisuuntaisen testin päätössäännöt voidaan kriittisten arvojen $\chi^2_{\alpha/2}$ ja $\chi^2_{1-\alpha/2}$ avulla muotoilla seuraavasti:

Kaksisuuntaisen testin päätössäännöt (I):

$$\begin{aligned} \text{Jos } \chi^2_{\text{hav}} < \chi^2_{1-\alpha/2} \text{ tai } \chi^2_{\text{hav}} > \chi^2_{\alpha/2}, & H_0 \text{ hylätään,} \\ \text{jos } \chi^2_{1-\alpha/2} \leq \chi^2_{\text{hav}} \leq \chi^2_{\alpha/2}, & H_0 \text{ hyväksytään.} \end{aligned}$$

Yksisuuntaisessa testissä päätössäännöt ovat analogiset aiemmin esitettyjen kanssa. Merkitsevyystason α kriittinen arvo on χ^2_{α} . Testin $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 < \sigma_0^2$ tapauksessa sille on voimassa $P(\chi^2 < \chi^2_{\alpha}) = \alpha$ ja testin $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 > \sigma_0^2$ tapauksessa $P(\chi^2 > \chi^2_{\alpha}) = \alpha$. Päätössäännöt 1-suuntaiselle testille ovat seuraavat:

Yksisuuntaisen testin päätössäännöt (I):

Testi (1)	Testi (2)
$H_0: \sigma^2 = \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 < \sigma_0^2$	$H_1: \sigma^2 > \sigma_0^2$

$$\begin{aligned} \text{jos } \chi^2_{\text{hav}} < \chi^2_{\alpha}, H_0 \text{ hylätään} & \quad \text{jos } \chi^2_{\text{hav}} > \chi^2_{\alpha}, H_0 \text{ hylätään} \\ \text{jos } \chi^2_{\text{hav}} \geq \chi^2_{\alpha}, H_0 \text{ hyväksytään} & \quad \text{jos } \chi^2_{\text{hav}} \leq \chi^2_{\alpha}, H_0 \text{ hyväksytään} \end{aligned}$$

Havaittu merkitsevyystaso eli p-arvo eri vaihtoehtoisten hypoteesien tapauksessa määräytyy nyt seuraavasti:

kaksisuuntainen testi

$$H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 \neq \sigma_0^2: p = 2 \cdot \min\{P(\chi^2 \geq \chi^2_{\text{hav}}), \{P(\chi^2 \leq \chi^2_{\text{hav}})\}\}$$

yksisuuntaiset testit

$$H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 < \sigma_0^2: p = P(\chi^2 \leq \chi^2_{\text{hav}})$$

$$H_0: \sigma^2 = \sigma_0^2, H_1: \sigma^2 > \sigma_0^2: p = P(\chi^2 \geq \chi^2_{\text{hav}})$$

Päätössäännöt sekä yksi- että kaksisuuntaisessa testissä ovat samat kuin aiemmin:

Kaksi- ja yksisuuntaisen testin päätössäännöt (II):

jos $p < \alpha$, H_0 hylätään
 jos $p \geq \alpha$, H_0 hyväksytään.

Populaation varianssin σ^2 piste-estimaatiksi käy otoksesta laskettu otosvariانسси s^2 , ja sen $100(1-\alpha)$ %:n luottamusväli on

$$\left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} \right).$$

missä $P(\chi^2(n-1) > \chi_{\alpha/2}^2(n-1)) = \alpha/2$ ja $P(\chi^2(n-1) > \chi_{1-\alpha/2}^2(n-1)) = 1 - \alpha/2$.

Tehtävä 65 (*Helenius, s. 347*): Älykkyystesti on normeerattu siten, että laajassa populaatiossa pisteluvun keskihajonta on 4 pistettä. Halutaan selvittää, toimiiko testi tietyn ikäisten koululaisten populaatiossa vastaavalla tavalla siten, että keskihajonta on 4. Tätä varten poimittiin 10 oppilaan otos, johon valituille suoritettiin kyseinen testi. Otoksessa havaittiin keskihajonnaksi 5.2 pistettä. Oletetaan, että tarkasteltavassa koululaisten populaatiossa pisteluvun jakaumaa voidaan luonnehtia normaalijakauman $N(\mu, \sigma^2)$ avulla. Tutki ongelmaa soveltuvan testin avulla.

Kahden otoksen varianssitesti

Kahden otoksen varianssitestissä verrataan kahden jakauman varianssia. Testiä voidaan käyttää esimerkiksi haluttaessa selvittää keskiarvotestin yhteydessä varianssien yhtäsuuruusoletuksen paikkansapitävyyttä.

Tarkasteltavana on kaksi populaatiota, joiden variansseja merkitään σ_1^2 ja σ_2^2 . Molempien populaatioiden oletetaan olevan normaalisti jakautuneita, ja jakaumien keskihajonnat ja varianssit ovat tuntemattomia. Testaus perustuu jälleen otoksista laskettuihin otosvariansseihin, joten kummastakin populaatiosta poimitaan otos, joiden tulee olla toisistaan riippumattomat. Merkitään otoskokoja ja otosvariansseja

	otos- koko	otos- varianssi
populaatio 1	n_1	s_1^2
populaatio 2	n_2	s_2^2

Testissä pyritään päättämään, ovatko populaatioiden varianssit yhtä suuret, ts. onko $\sigma_1^2 = \sigma_2^2$ vai ei. Ehto voidaan lausua myös muodossa $\sigma_1^2/\sigma_2^2 = 1$. Hypoteesit ovat siis

<i>kaksisuuntainen testaus</i>	<i>yksisuuntainen testaus</i>		
$H_0: \sigma_1^2/\sigma_2^2 = 1$	$H_0: \sigma_1^2/\sigma_2^2 = 1$	tai	$H_0: \sigma_1^2/\sigma_2^2 = 1$
$H_1: \sigma_1^2/\sigma_2^2 \neq 1$	$H_1: \sigma_1^2/\sigma_2^2 < 1$		$H_1: \sigma_1^2/\sigma_2^2 > 1$

Mikäli otokset ovat peräisin nollahypoteesin mukaisista populaatioista, testisuure on

$$F = s_1^2/s_2^2 \sim F(n_1-1, n_2-1),$$

eli testisuure F on F-jakautunut vapausastein n_1-1, n_2-1 .

Luottamusvälin määrittäminen sivuutetaan.

Tehtävä 66 (*Helenius, s. 353*): Halutan verrata kahden eritasoisen kurssin kuulustelujen pistemäärien hajontoja. 21 peruskurssin opiskelijan pisteluvun varianssiksi havaittiin 478 ja 31 jatkokurssin opiskelijan varianssiksi 372. Eroavatko varianssit kyseisissä populaatioissa, kun oletetaan, että pistemäärien jakaumat ovat normaaleja?

2.3.5 Suhteellisen osuuden testaus

Suhteellisen osuuden testauksessa tutkittavana on jonkin ominaisuuden suhteellinen osuus populaatiossa. Tyypillisiä esimerkkejä kiinnostavista suhteellisista osuuksista ovat jonkin poliittisen puolueen kannatusprosentti ja yrityksen tai tuotteen markkinaosuus. Kuten kaikessa testauksessa, tässäkin tapauksessa idea-

na on tehdä koko populaatiota koskevia johtopäätöksiä otoksesta lasketun suhteellisen osuuden perusteella.

Suhteellisen osuuden testauksia on kahta lajia: 1) yhden otoksen suhteellisen osuuden testaus, jossa arvioidaan, onko populaation suhteellinen osuus väitetyyn suuruinen ja 2) kahden otoksen suhteellisen osuuden testaus, jossa testataan sitä, ovatko kahden eri populaation suhteelliset osuudet samansuuruiset vai ei.

Suhteellisen osuuden otantajakauma

Populaation suhteellisen osuuden π estimaattorina käytetään otoksen suhteellista osuutta p . Kuten aiemminkin, estimaattorin otantajakauma on tarpeen tuntea ainakin likimäärin, jotta tilastollista päättelyä voitaisiin suorittaa.

Ominaisuuden A suhteellinen osuus populaatiossa, esimerkiksi vasenkätisten osuus on ominaisuuden A omaavien yksiköiden (siis esim. vasenkätisten) lukumäärä jaettuna populaation koolla. Tätä populaation suhteellista osuutta merkitään π llä ja se on siis suhdeluku (nollan ja ykkösen välillä), ei prosenttiluku. Populaation suhteellinen osuus voidaan tulkita myös ominaisuuden A esiintymistodennäköisyydeksi populaatiossa. Siis todennäköisyys, että satunnaisesti valitulla populaation jäsenellä on ominaisuus A , on π .

Jos nyt populaatiosta poimitaan palauttaen n yksilön otos, on satunnaismuuttuja K = ominaisuuden A esiintymisten lukumäärä n suuruudessa otoksessa binomijakautunut, $K \sim \text{Bin}(n, \pi)$. K :n jakauman odotusarvo on siten $EK = n\pi$ ja varianssi $\text{Var } K = n\pi(1-\pi)$.

Otoksen suhteellinen osuus p saadaan jakamalla ominaisuuden A otoksessa esiintymisten lukumäärä otoksen koolla, eli $p = K/n$. Koska K :n jakauma tunnetaan, on todennäköisyyden laskusääntöjen avulla helppo määrittää p :n jakauma: p on myös binomijakautunut ja sen odotusarvo ja varianssi ovat

$$Ep = E\left(\frac{K}{n}\right) = \frac{1}{n}EK = \frac{1}{n} \cdot n\pi = \pi$$

ja

$$\text{Var } p = \text{Var}\left(\frac{K}{n}\right) = \frac{1}{n^2} \text{Var } K = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}.$$

Estimaattorin p keskivirhe on täten

$$\sigma_p = \sqrt{\text{Var } p} = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Binomijakaumaa voidaan approksimoida normaalijakaumalla, mikäli otoskoko on riittävän suuri. Riittävä approksimaation tarkkuus katsotaan yleensä saavutettavan, jos $n\pi \geq 5$ ja $n(1-\pi) \geq 5$. Näiden ehtojen vallitessa siis p on likimain normaalisti jakautunut, $p \sim N(\pi, \frac{\pi(1-\pi)}{n})$.

Jos suoritetaan otanta palauttamatta, kuten useimmiten on laita, on K :n jakauma tarkkaan ottaen hypergeometrinen, ja edellisissä kaavoissa on suhteellisen osuuden p keskivirheen kaavaa korjattava kertoimella $\sqrt{\frac{N-n}{n-1}}$, jossa N on populaation koko. Kuitenkin, kuten aikaisemmin monisteessa on todettu, hyvin suuren populaation tapauksessa ($n/N \leq 0.05$) binomijakauma antaa riittävän tarkan approksimaation K :n jakaumasta, jolloin korjausta ei tarvitse tehdä.

Seuraavaksi esiteltävässä testissä testisuure saadaan standardoimalla p , jolloin testisuure noudattaa standardoitua normaalijakaumaa. Huom! Testi antaa luotettavia tuloksia vain, jos otoskoko on riittävän suuri, ts. jos edellä mainitut ehdot $n\pi \geq 5$ ja $n(1-\pi) \geq 5$ ovat voimassa.

Suhteellisen osuuden testaus - yksi otos

Yhden otoksen suhteellisen osuuden testauksessa ollaan kiinnostuneita kohdepopulaation suhteellisesta osuudesta π . Testauksessa pyritään päättämään otoksesta lasketun estimaatin p avulla, eroaako π tietystä väitetystä suhteellisesta osuudesta π_0 vai ei. Hypoteesit ovat muotoa

<i>kaksisuuntainen</i>	<i>yksisuuntainen</i>	
<i>testaus</i>	<i>testaus</i>	
$H_0: \pi = \pi_0$	$H_0: \pi = \pi_0$	tai $H_0: \pi = \pi_0$
$H_1: \pi \neq \pi_0$	$H_1: \pi < \pi_0$	$H_1: \pi > \pi_0$

Mikäli otos on peräisin nollahypoteesin mukaisesta populaatiosta, eli $\pi = \pi_0$, on testisuure

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1),$$

eli tehdään z-testi.

Populaation suhteellisen osuuden π piste-estimaatti on p ja sen $100(1-\alpha) \%$:n luottamusväli on

$$(p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}),$$

missä $z_{\alpha/2}$ toteuttaa ehdon $\Phi(z_{\alpha/2}) = 1-\alpha/2$.

Tehtävä 67 (*Karjalainen ja Ruuskanen, s.160*): Kurkunsiemementen itävyydeksi on todettu 82%. Vuoden varastoinnin jälkeen kylvettiin 80 siementä, joista iti 70. Voidaanko katsoa, että varastoinnin aikana tapahtui muutos siementen itävyydessä?

Suhteellisen osuuden testaus - kaksi riippumatonta otosta

Kahden otoksen suhteellisen osuuden testauksessa tarkastellaan kahden populaation suhteellisten osuuksien π_1 ja π_2 eroa, esimerkiksi eroavatko TV-uutisten katsojien suhteelliset osuudet alle 20-vuotiaiden ja vähintään 20-vuotiaiden populaatioissa. Testausta varten poimitaan kummastakin populaatiosta toisistaan riippumattomat otokset, joista laskettujen suhteellisten osuuksien erotusta $p_1 - p_2$ käytetään populaatioiden suhteellisten osuuksien erotuksen $\pi_1 - \pi_2$ estimaattina. Tässä tarkastellaan vain tapausta, jossa testataan, ovatko populaatioiden suhteelliset osuudet samat vai eivät. Hypoteesit ovat tällöin muotoa

<i>kaksisuuntainen</i>	<i>yksisuuntainen</i>	
<i>testaus</i>	<i>testaus</i>	
$H_0: \pi_1 - \pi_2 = 0$	$H_0: \pi_1 - \pi_2 = 0$	tai $H_0: \pi_1 - \pi_2 = 0$
$H_1: \pi_1 - \pi_2 \neq 0$	$H_1: \pi_1 - \pi_2 < 0$	$H_1: \pi_1 - \pi_2 > 0$

Estimaattorin $p_1 - p_2$ otantajakauma voidaan johtaa vastaavasti kuin edellä, jolloin odotusarvoksi saadaan

$$E(p_1 - p_2) = \pi_1 - \pi_2$$

ja keskivirheeksi

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

Otantajakaumaa voidaan pitää likimain normaalina, jos otoskoot ovat riittävän suuria ($n_1\pi_1 \geq 5$, $n_1(1-\pi_1) \geq 5$ ja $n_2\pi_2 \geq 5$, $n_2(1-\pi_2) \geq 5$).

Mikäli nollahypoteesi $\pi_1 - \pi_2 = 0$ on voimassa, populaatioiden suhteelliset osuudet ovat samat. Tätä populaatioiden yhteistä suhteellista osuutta merkitään π llä. Tällöin estimaattorin $p_1 - p_2$ odotusarvoksi tulee nolla ja keskivirheeksi

$$\sqrt{\pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Nyt π on tuntematon, ja sen estimaattorina voidaan käyttää otoksista saatujen suhteellisten osuuksien painotettua keskiarvoa

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

Kun estimaattori $p_1 - p_2$ vielä standardoidaan, saadaan standardoitua normaali-jakaumaa noudattava testisuure

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1).$$

Tehdään siis z-testi.

Populaatioiden suhteellisten osuuksien erotuksen $\pi_1 - \pi_2$ piste-estimaatti on otoksista laskettu $p_1 - p_2$ ja sen $100(1-\alpha)$ %:n luottamusväli on

$$\left(p - z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, p + z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right),$$

missä $z_{\alpha/2}$ toteuttaa ehdon $\Phi(z_{\alpha/2}) = 1-\alpha/2$.

Tehtävä 68 (*Karjalainen-Ruuskanen, s.161*): Tutkittaessa väestön koulutustasoa saatiin seuraavat havainnot: Etelä-Suomen läänin 200 henkilön otoksessa 104 oli saanut ammatillisen koulutuksen ja Länsi-Suomen läänin 100 henkilön otoksessa ammatillinen koulutus oli 43 henkilöllä. Testaa 1 % merkitsevyystasolla, oliko ammatillisen koulutuksen saaneiden suhteellinen osuus suurempi Etelä-Suomen kuin Länsi-Suomen lääneissä.

2.3.6 Korrelaatiokertoimen testaus

Kahden muuttujan X ja Y välistä lineaarista riippuvuutta voidaan tutkia Pearsonin korrelaatiokertoimen avulla, mikäli molemmat muuttujat ovat vähintään välimatka-asteikollisia. Esitetään tässä vielä kertauksen vuoksi monisteen alkuosasta tuttu korrelaatiokertoimen kaava

$$r_{XY} = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}}$$

missä n = havaintoparien lukumäärä (otoksen koko)

x_i = muuttujan X havaittu arvo tilastoyksikön i kohdalla

y_i = muuttujan Y havaittu arvo tilastoyksikön i kohdalla.

Muuttujat X ja Y ovat keskenään lineaarisesti riippumattomat silloin, kun korrelaatiokertoimen arvo on nolla. Jos X ja Y ovat populaatiossa toisistaan riippumattomat, niin silloin otoksesta lasketun korrelaatiokertoimenkin pitäisi olla lähellä nollaa. Kuinka suurta otoskorrelaation poikkeamaa nollasta voidaan vielä pitää uskottavana sattuman aiheuttamana poikkeamana, jos X ja Y tosiaan ovat toistaan riippumattomat populaatiossa?

Tämä voidaan testata samaan tapaan kuin edellä. Populaation korrelaatiokertoimen ρ estimaattorina käytetään otoskorrelaatiota r , joka lasketaan yllä olevalla kaavalla. Korrelaation yhteydessä useimmin esiintyvä kysymys on, ovatko X ja Y populaatiossa lineaarisesti riippuvia vai eivät, jolloin nollahypoteesina esiintyy väittämä: ”ei lineaarista riippuvuutta”, eli $H_0: \rho = 0$. Testaus voi olla yksi- tai kaksisuuntainen. Kaksisuuntaisen testin vastahypoteesi on muotoa ”on lineaarista riippuvuutta”. Yksisuuntaisessa testissä vastahypoteesi on joko muotoa ”on posi-

tiivista lineaarista riippuvuutta" ($\rho > 0$) tai "on negatiivista lineaarista riippuvuutta" ($\rho < 0$). Hypoteesit ovat siis

<i>kaksisuuntainen</i>	<i>yksisuuntainen</i>	
<i>testaus</i>	<i>testaus</i>	
$H_0: \rho = 0$	$H_0: \rho = 0$	<i>tai</i> $H_0: \rho = 0$
$H_1: \rho \neq 0$	$H_1: \rho < 0$	$H_1: \rho > 0$

Testausta varten on taas tunnettava estimaattorin r jakauma, ainakin likimäärin. Jos otos on peräisin nollahypoteesin $\rho = 0$ mukaisesta populaatiosta ja sekä X että Y ovat populaatiossa normaalisti jakautuneita, voidaan osoittaa, että testisuure

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

noudattaa t-jakaumaa vapausasteella $n-2$. Suoritetaan siis t-testi.

Tehtävä 69: 27 henkilön otoksella tutkittiin kahdessa eri testissä menestymisen riippuvuutta. Merkitään X = testissä 1 saatu pistemäärä ja Y = testissä 2 saatu pistemäärä. Testipistemäärien väliseksi korrelaatiokertoimeksi saatiin $r = 0.389$. Onko perusjoukossa positiivista lineaarista riippuvuutta muuttujien X ja Y välillä?

2.3.7 Eräitä ei-parametrisia testejä

Seuraavaksi tarkastellaan kahta ei-parametrista testiä, χ^2 -yhteensopivuus- ja χ^2 -riippumattomuustestiä. Kumpikin testi soveltuu jo luokitteluasteikon tasoisille muuttujille ja niiden käyttäminen edellyttää ainoastaan, että käytettävät havainnot ovat riippumaton otos jostakin populaatiosta. χ^2 -yhteensopivuustesti koskee yksiulotteista ja χ^2 -riippumattomuustesti kaksiulotteista frekvenssijakaumaa.

Näitä testejä käytettäessä on erityisesti muistettava, että testeissä tarkastellaan aina absoluuttisia frekvenssejä, ei siis koskaan suhteellisia tai prosenttisia frekvenssejä.

χ^2 -yhteensopivuustesti

χ^2 -yhteensopivuustestillä tutkitaan, noudattaako populaation frekvenssijakauma jotain tunnettua jakaumaa. Tämä oletettu jakauma edustaa testissä sitä teoreettista mallia, joka ilmaistaan nollahypoteesin avulla. Esimerkiksi tutkittaessa eri tuotemerkkien menekkiä, voidaan teoreettisena mallina pitää tasaista jakaumaa, eli jokaisen tuotemerkin menekki on yhtä suuri. Testissä verrataan otoksesta mitattuja eli *havaittuja frekvenssejä* oletetun jakauman mukaisiin eli *odotettuihin frekvensseihin*. Jos näiden välinen ero on pieni, voidaan havaintojen katsoa olevan peräisin oletetusta jakaumasta.

Hypoteesit muotoillaan tapauskohtaisesti seuraavaan muotoon:

- H_0 : havaittujen ja odotettujen frekvenssien välillä ei ole eroa, eli havaintoaineisto on peräisin oletettua jakaumaa $f(X)$ noudattavasta populaatiosta.
 H_1 : havaittujen ja odotettujen frekvenssien välillä on eroa eli havaintoaineisto on peräisin populaatiosta, joka ei noudata jakaumaa $f(X)$.

Havaituista ja odotetuista frekvensseistä muodostetaan seuraava taulukko:

Luokka	L_1	L_2	\dots	L_k	yhteensä
Havaittu frekvenssi	f_1	f_2	\dots	f_k	n
Odotettu frekvenssi	e_1	e_2	\dots	e_k	n

missä L_i = tarkasteltavan muuttujan arvojen luokka i

f_i = luokan i havaittu frekvenssi

e_i = luokan i odotettu frekvenssi

k = luokkien lukumäärä

n = havaintojen lukumäärä eli otoskoko.

Testisuurena on luvusta 5.2.1 tuttu χ^2 -suure sovellettuna 1-riviseen eli $1 \times k$ -tyyppiseen kontingenssitauluun:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

Jos odotetut frekvenssit ovat hyvin pieniä, ei testiä voi soveltaa sellaisenaan. Mikäli ehdot

1. korkeintaan 20% odotetuista frekvensseistä saa olla alle 5

2. jokaisen odotetun frekvenssin on oltava suurempi kuin 1

eivät toteudu, on vierekkäisiä luokkia yhdisteltävä.

Jos edelliset kaksi vaatimusta täyttyvät, testisuure noudattaa χ^2 -jakaumaa vapausasteena $k-m-1$, eli $\chi^2 \sim \chi^2(k-m-1)$. Luku m on niiden oletetun jakauman parametrien määrä, joka mahdollisesti joudutaan testauksen yhteydessä estimoimaan. Tasaisen jakauman tapauksessa estimoitavia parametreja ei ole, joten tällöin $m=0$. Jos oletettu jakauma olisi esim. Poisson-jakauma, pitäisi jakauman parametri λ estimoida, eli m olisi 1.

Tehtävä 58 (*Vasama-Vartia (II) s. 537*): Pelaaja epäili rahapelissä käytetyn arpanopan olevan harhainen. Tämän epäilynsä tueksi hän suoritti 120 heiton koesarjan, jonka tulokset ilmenevät alla olevasta taulukosta.

silmluku	1	2	3	4	5	6
havaitut						
frekvenssit	12	16	20	17	22	33

Testaa, oliko pelaaja epäilyissään oikeassa.

χ^2 -riippumattomuustesti

Riippumattomuustesteillä selvitetään, onko tarkasteltavien muuttujien välillä riippuvuutta populaatiossa vai ei. χ^2 -riippumattomuustestissä verrataan ristiintaulukoitujen muuttujien havaittuja solufrekvenssejä riippumattomuushypoteesin mukaisiin odotettuihin frekvensseihin. Jos nämä poikkeavat toisistaan merkittävästi, voidaan pitää uskottavana, että riippumattomuusoletus oli virheellinen. Jos merkitään tarkasteltavia muuttujia X :llä ja Y :llä, voidaan hypoteesit muotoilla seuraavasti:

H_0 : havaittujen ja odotettujen frekvenssien välillä ei ole eroa, eli muuttujat X ja Y ovat toisistaan riippumattomia

H_1 : havaittujen ja odotettujen frekvenssien välillä on eroa eli muuttujat X ja Y ovat toisistaan riippuvat.

Havainnoista muodostetaan ensin kontingenssitaulukko, jonka perusteella lasketaan odotetut frekvenssit (ks. monisteen luku 5.2.1). χ^2 -testisuure noudattaa χ^2 -jakaumaa vapausasteena $(r-1)(s-1)$,

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2((r-1)(s-1))$$

missä f_{ij} = solun ij havaittu frekvenssi

e_{ij} = solun ij odotettu frekvenssi

r = taulukon rivien lukumäärä

s = taulukon sarakkeiden lukumäärä.

Odotettuja frekvenssejä koskevat samat vaatimukset kuin χ^2 -yhteensopivuustestissä.

Tehtävä 59 (*Karjalainen-Ruuskanen s. 151*): Kysyttäessä suomalaisilta nuorilta heidän käsitystään siitä, miksi työttömällä ei ole työtä, saatiin seuraava havainto-aineisto:

	Ei saa työtä	Ei halua työtä
Pääkaupunkiseudun nuori	14	78
Muun kaupungin nuori	31	64
Maaseudun nuori	32	63

Testaa, ovatko nuoren asuinpaikka ja hänen mielipiteensä työttömyyden syystä riippumattomia.

3 KIRJALLISUUTTA

Anderson, D.R., Sweeney D.J. ja Williams T.A.: Statistics for Business and Economics, 5th ed., West Publishing Company.

Grönroos, M.: Johdatus tilastotieteeseen, 2003, Finn Lectura.

Heikkilä, J.: Tilastotieteen ABC-kirja 1, 1993, Gummerus.

Heikkilä, T.: Tilastollinen tutkimus, 1998, Edita.

Helenius, H.: Tilastollisten menetelmien perustiedot, 1989, Statcon.

Karjalainen, L. ja Ruuskanen, A.: Tilastomatematiikka, 1994, Gummerus.

Liski, E. ja Puntanen, S.: Tilastotieteen peruskurssi I, 1987, Tampereen Yliopisto.

Liski, E. ja Puntanen, S.: Tilastotieteen peruskurssi II, 1987, Tampereen Yliopisto.

Manninen, P.: Tilastotiedettä yhteiskuntatieteilijöille, 1976, Gaudeamus.

Neter, J. , Wasserman W. ja Whitmore G.: Applied Statistics, 1992, Allyn ha Bacon.

Niiniluoto, I.: Tieteellinen päättely ja selittäminen, 1983, Otava.

Nummenmaa, T., Konttinen, R., Kuusinen, J., Leskinen, E.: Tutkimusaineiston analyysi, 1997, WSOY.

Vasama, P-M ja Vartia, Y.: Johdatus tilastotieteeseen I, 1973, Gaudeamus.

Vasama, P-M ja Vartia, Y.: Johdatus tilastotieteeseen II, 1973, Gaudeamus.