Luku 6

Tilastollinen testaus

Tilastolliset testit ovat frekventistisen päättelyn käytetyimpiä (ja valitettavasti myös huonoiten ymmärrettyjä ja sen takia myös eniten väärinkäytettyjä) tilastollisen päättelyn menetelmiä. Niiden avulla pyritään ottamaan kantaa tilastollisia malleja koskeviin väitteisiin eli hypoteeseihin, kuten esim.

- Onko tutkittava lantti harhaton?
- Onko tietyllä käsittelyllä vaikutusta? (Käsittely voisi olla esimerkiksi uusi hoitomuoto jollekin sairaudelle tai uusi lannoite tai uusi opetusmenetelmä.)

Testaus sujuu käytännössä laskemalla tilanteeseen sopivan testisuureen arvo. Laskettua arvoa verrataan siihen, minkälaisia arvoja hypoteesin mukaisesta populaatiosta satunnaisvaihtelu huomioon ottaen pitäisi tulla. Mikäli laskettu testisuureen arvo poikkeaa riittävän paljon hypoteesin mukaisista tyypillisistä arvoista, hypoteesi hylätään. Tällöin meistä ei enää ole uskottavaa, että näin suuri poikkeama aiheutuisi satunnaisvaihtelusta, vaan pidämme uskottavampana selityksenä sitä, että asetettu hypoteesi ei pidä paikkaansa.

6.1 Testauksen peruskäsitteitä

Tarkastelemme testausta frekventistisessä viitekehyksessä parametrisessa tilastollisessa mallissa eli jakaumaperheessä

$$\{f(\mathbf{y};\theta),\ \theta \in \Theta\}.$$

Koska havaintoja vastaavan satunnaisvektorin $\mathbf{Y}$ jakauma tunnetaan täysin, mikäli parametrinarvo θ tunnetaan, niin vektorin $\mathbf{Y}$ jakaumaa koskevat väittämät eli (tilastolliset) hypoteesit voidaan pukea siihen muotoon, että parametri kuuluu johonkin tiettyyn parametrialueen osajoukkoon.

Esimerkiksi lantinheittoa tavallisesti mallinnetetaan binomikokeena, jossa onnistumistodennäköisyys θ on tuntematon ja toistojen lukumäärä n on tunnettu. Väite "lantti on harhaton" vastaa hypoteesia

$$\theta = \frac{1}{2}, \quad \text{eli} \quad \theta \in \{\frac{1}{2}\}.$$

Lantin harhattomuutta koskeva hypoteesi on *yksinkertainen* (engl. *simple*) eli *täysin määrätty*, sillä hypoteesia vastaava parametriavaruuden osajoukko sisältää vain yhden pisteen $\theta_0 = \frac{1}{2}$. Tällöin satunnaisvektorin $\mathbf{Y}$ jakaumalla on yptnf/ytf $f(\mathbf{y}; \theta_0)$, mikäli hypoteesi on tosi. Paljon tyypillisempää on, että hypoteesi on *yhdistetty* (engl. *composite*) eli *osittain määrätty*, mikä tarkoittaa sitä, että hypoteesia vastaava parametriavaruuden osajoukko koostuu useammasta kuin yhdestä pisteestä.

Jotkin hypoteesit saattavat ensinäkemältä vaikuttaa yksinkertaisilta, vaikka ne todellisuudessa ovat yhdistettyjä. Jos esimerkiksi normaalijakautuneessa populaatiossa $N(\mu, \sigma^2)$ molemmat parametrit ovat tuntemattomia, niin tällöin parametria μ koskeva tarkka hypoteesi

$$H : \mu = 0$$

on yhdistetty hypoteesi, sillä se vastaa parametriavaruuden osajoukkoa

$$\{(\mu, \sigma^2) : \mu = 0 \text{ ja } \sigma^2 > 0\}.$$

Testauksessa asetetaan ns. *nollahypoteesi* (engl. *null hypothesis*) H_0, joka on muotoa

$$H_0 : \theta \in \Theta_0, \tag{6.1}$$

missä $\Theta_0 \subset \Theta$ on ei-tyhjä parametriavaruuden osajoukko. Testauksen tavoitteena on arvioida havaintojen $\mathbf{y}$ avulla nollahypoteesin paikkansapitävyyttä.

Tavallisesti nollahypoteesi vastaa vakiintunutta teoriaa tai sitä pessimististä näkemystä, että käsittelyllä ei ole vaikutusta. Tällöin tutkija tahtoisi todellisuudessa löytää todisteita nollahypoteesin hylkäämiseksi, mutta koska vakiintunutta teoriaa ei voida hylätä löyhin perustein, sitä vastaan pitää saada vakuuttavia todisteita, ennen kuin yhteisö suostuu hylkäämään nollahypoteesin.

Monesti nollahypoteesin H_0 lisäksi muotoillaan myös *vaihtoehtoinen hypoteesi* eli *vastahypoteesi* (engl. *alternative hypothesis*, myös *study hypothesis*), jota tyypillisesti merkitään symbolilla H_1 (tai toisinaan H_A). Vastahypoteesin mukaan θ kuuluu parametriavaruuden osajoukkoon Θ_1, ts.

$$H_1 : \theta \in \Theta_1. \tag{6.2}$$

Tällöin vähintäänkin oletetaan, että

$$\Theta_0 \cap \Theta_1 = \emptyset,$$

ja usein (mutta ei aina) pätee $\Theta = \Theta_0 \cup \Theta_1$. Jos vastahypoteesi asetetaan, niin tämä tarkoittaa kannanottoa sen suhteen, mitä parametrin ajatellaan toteuttavan siinä tapauksessa, että nollahypoteesi osoittautuu epäilyksen alaiseksi.

Nollahypoteesia ja vastahypoteesia käsitellään testauksessa epäsymmetrisellä tavalla. Tämän asian ymmärtäminen on edellytys sille, että osaamme tulkita testin lopputuloksen järkevästi. Tilanteen voi ajatella olevan analoginen oikeudenkäynnin kanssa, jossa syytettynä on nollahypoteesi H_0. H_0 on oikeudenkäynnissä syytön, ellei sitä osoiteta syylliseksi.

Testaus sujuu siten, että aineistosta lasketaan tunnusluku $t(\mathbf{y})$, jota kutsutaan testisuureeksi (engl. *test statistic*). Testisuure asettaa mahdolliset havainnot järjestykseen jollakin seuraavista tavoista sen asian suhteen, miten tyypillisiä mahdollisia havaintoja pidämme nollahypoteesin vallitessa.

1. Pienet tunnusluvun arvot viittaavat siihen, että aineisto on sopusoinnussa H_0:n kanssa, ja suuret viittaavat ristiriitaan H_0:n kanssa.

2. Suuret tunnusluvun arvot viittavat sopusointuun H_0:n kanssa ja pienet arvot ristiriitaan sen kanssa.

3. Suuri poikkeama jostakin vertailuarvosta t_0 ylöspäin tai alaspäin viittaa ristiriitaan ja pieni poikkeama sopusointuun.

Lisäksi vaaditaan se, että hallitsemme testisuuretta vastaavan satunnaismuuttujan $t(\mathbf{Y})$ jakauman ainakin kaikilla nollahypoteesin mukaisilla parametrinarvoilla $\theta \in \Theta_0$. Usein testisuureena käytetään saranasuuretta, mikäli sellainen tunnetaan. Tällä kurssilla emme voi paneutua tämän syvällisemmin siihen, kuinka testisuure pitäisi valita.

Tilastollinen testi toimii sillä tavalla, että havaitusta aineistosta lasketaan testisuureen arvo, ja sitten tarkistetaan kuuluuko se *kriittiseen alueeseen* (engl. *critical region*) eli *hylkäysalueeseen* (engl. *rejection region*) C. Testi antaa yhden kahdesta vaihtoehtoisesta päätöksestä: se joko hylkää nollahypoteesin tai sitten ei sen mukaan, kuuluko testisuureen arvo kriittiseen alueeseen vai ei.

- Jos $t(\mathbf{y}) \in C$, niin testi *hylkää* (engl. *reject*) nollahypoteesin, eli testi on *merkitsevä* (engl. *significant*).

- Jos $t(\mathbf{y}) \notin C$, niin testi *hyväksyy* (engl. *accept*) nollahypoteesin (mikä voidaan ilmaista myös sanomalla, että *nollahypoteesi jää voimaan*), eli testi *ei ole merkitsevä* (engl. *not significant*).

Huomaa, että hylkääminen ja sen vastakohta, jota yksinkertaisuuden vuoksi tavallisimmin kutsutaan hyväksymiseksi, ovat testaukseen liittyviä teknisiä termejä. Se mitä käytännön johtopäätöksiä ja käytännön toimia testin lopputuloksen selvittyä tehdään, on eri asia kuin testin antama päätös. Varsinkin termi hyväksyä on harhaanjohtava. Mikäli H_0 hyväksytään, niin tutkija usein oikeasti edelleen epäilee nollahypoteesin paikkansapitävyyttä, mutta hän ei ole löytänyt aineistosta riittävän vakuttavaa todistetta sitä vastaan.

Kriittisen alueen muoto riippuu siitä, minkälaiset tunnusluvun arvot ovat nollahypoteesin kanssa yhteensopimattomia. Jos suuret tunnusluvun arvot ovat nollahypoteesin kannalta kriittisiä, niin kriittinen alue on muotoa

$$C = (u, \infty)$$

ts. testi hylkää nollahypoteesin, jos $t(\mathbf{y}) > u$. Tällöin kynnysarvoa u voidaan kutsua *kriittiseksi arvoksi* (engl. *critical value*). Tavallisesti kriittinen alue määräytyy testin merkitsevyystasosta.

Testien yhteydessä puhutaan niiden koosta tai merkitsevyystasosta. Käytämme näitä termejä synonyymeinä, mutta jotkut kirjoittajat tekevät näiden käsitteiden välille eron.

Määritelmä 6.1. Jos testin kriittinen alue on C, niin testin *koko* (engl. *size*) eli sen *merkitsevyystaso* (engl. *significance level*) on

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(t(\mathbf{Y}) \in C) = \sup_{\theta \in \Theta_0} P_\theta(H_0 \text{ hylätään}) \tag{6.3}$$

Tässä sup eli *supremum* tarkoittaa pienintä ylärajaa; testin koko α on pienin yläraja hylkäystodennäköisyydelle $P_\theta(t(\mathbf{Y}) \in C)$, kun satunnaisvektorilla $\mathbf{Y}$ on nollahypoteesin mukainen jakauma. Ts. $0 < \alpha < 1$, ja

$$P_\theta(t(\mathbf{Y}) \in C) \le \alpha, \qquad \text{kaikilla } \theta \in \Theta_0$$

ja kaikilla $\epsilon > 0$ on olemassa $\theta \in \Theta_0$ siten, että

$$P_\theta(t(\mathbf{Y}) \in C) > \alpha - \epsilon.$$

Usein hylkäystodennäköisyys $P_\theta(t(\mathbf{Y}) \in C)$ pysyy vakiona joukossa Θ_0. Näin käy automaattisesti, jos H_0 on yksinkertainen hypoteesi ja myös silloin, jos testisuure on saranasuure. Tässä tapauksessa testin merkitsevyystaso on yksinkertaisesti

$$\alpha = P_\theta(H_0 \text{ hylätään}), \qquad \text{millä tahansa } \theta \in \Theta_0. \tag{6.4}$$

Tyypillisesti testin merkitsevyystaso $0 < \alpha < 1$ asetetaan, ja sitten tämän informaation perusteella määritetään kriittinen alue C siten, että vaatimus (6.3) toteutuu. Ennen vanhaan ei ollut käytössä tilastollisia ohjelmia, ja merkitsevyystasolle α kiinnitettiin tavallisesti jokin seuraavista konventionaalisista arvoista

$$0.05, \quad 0.01, \quad \text{tai} \quad 0.001$$

sen takia, että näitä arvoja vastaavat kriittiset arvot löytyivät tilastollisista taulukoista. Nämä konventionaaliset tasot ovat mielivaltaisia, ja ne on valittu sillä perusteella, että vastaavat murtoluvut (yksi kahdestakymmenestä, yksi sadasta, yksi tuhannesta) ovat pyöreitä.

Testin tekemään päätökseen liittyy aina virheen mahdollisuus. Jos H_0 pitää paikkansa, mutta testi hylkää sen, tällöin tapahtuu *hylkäämisvirhe* eli *I lajin virhe* (engl. *type I error*). Jos H_1 pitää paikkansa, mutta testi hyväksyy H_0:n, tapahtuu *hyväksymisvirhe* eli *II lajin virhe* (engl. *type II error*).

Todellisuus	Päätös	
	H_0 hyväksytään	H_0 hylätään
H_0 tosi	oikea päätös	hylkäämisvirhe I lajin virhe
H_1 tosi	hyväksymisvirhe II lajin virhe	oikea päätös

Testissä nollahypoteesia ja vastahypoteesia kohdellaan epäsymmetrisellä tavalla. Merkitsevyystaso α on yläraja hylkäämisvirheen todennäköisyydelle. *Jos nollahypoteesi pitää paikkaansa*, niin testisuure saa hylkäämiseen johtavia arvoja niin harvoin, että hylkäämistodennäköisyys on enintään α. Tähän asti emme ole lainkaan miettineet sitä, mitä testissä tapahtuu jos H_1 on tosi.

Perinteinen tapa raportoida testin tulos on ollut kiinnittää testin koko α sekä kertoa testin päätös, eli hylkäsikö vai hyväksyikö testi nollahypoteesin, mutta nykyään ei välttämättä toimita näin yksioikoisesti.

6.2 Normaalijakautuneen populaation odotusarvon testaus, kun varianssi on tunnettu

Tarkastelemme satunnaisotosta $Y_1, \ldots, Y_n$ normaalijakaumasta $N(\mu, \sigma^2)$. Ts. satunnaismuuttujat Y_i ovat riippumattomia, ja niillä on kaikilla normaalijakauma $N(\mu, \sigma^2)$.

Tarkastelemme odotusarvon testausta, kun varianssi on tunnettu. Tällä tilanteella ei ole suurta käytännön arvoa. Tästä syystä esim. R-ohjelmistossa ei ole valmista funktiota, joka laskisi kätevästi tämän testin tulokset. Tätä testiä käsitellään sen vuoksi, että teoria on tässä tapauksessa helppoa.

Yksisuuntainen testi

Jos populaation varianssi σ^2 on tunnettu luku, niin

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

on saranasuure. Tarkastelemme nollahypoteesia

$$H_0 : \mu = \mu_0,$$

jossa μ_0 on tunnettu luku (esim. $\mu_0 = 0$). Tämä on yksinkertainen hypoteesi. Otamme ensin vastahypoteesiksi yksisuuntaisen hypoteesin

$$H_1 : \mu > \mu_0,$$

joka on yhdistetty hypoteesi. Tätä hypoteesiparia vastaavat parametriavaruuden osajoukot

$$\Theta_0 = \{\mu_0\}, \qquad \Theta_1 = (\mu_0, \infty)$$

Käytämme testisuureena tunnuslukua

$$z = t(\mathbf{y}) = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}. \tag{6.5}$$

Huomaa, että testisuure on tunnusluku (toisin kuin sitä vastaava saranasuure), sillä testisuureessa tuntemattoman parametrin μ tilalla on tunnettu arvo μ_0. Testisuuretta vastaavalla satunnaismuuttujalla

$$t(\mathbf{Y}) = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}}$$

on standardinormaalijakauma $N(0,1)$, kun nollahypoteesi pitää paikkansa. Nyt suuret testisuureen arvot ovat nollahypoteesin kannalta kriittisiä, sillä $\bar{Y}$ estimoi populaatioparametria μ, ja testisuure on kasvava funktio tästä estimaattorista.

Tason α testi saadaan aikaan käyttämällä kriittistä arvoa z_α, sillä

$$P_{\mu_0}(t(\mathbf{Y}) > z_\alpha) = P(Z > z_\alpha) = \alpha,$$

jossa $Z \sim N(0,1)$. Tästä nähdään, että luottamustason α testi hypoteesiparille

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0,$$

tekee päätöksen seuraavasti. Ensin lasketaan testisuureen arvo

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

ja sitten testi toimii seuraavasti

$$\begin{cases} \text{jos } z > z_\alpha, & H_0 \text{ hylätään} \\ \text{jos } z \le z_\alpha & H_0 \text{ hyväksytään.} \end{cases}$$

Ts. testin hylkää nollahypoteesin silloin (ja vain silloin), kun

$$z > z_\alpha. \tag{6.6}$$

Tämä on ns. *yksisuuntainen* eli *yksitahoinen z-testi* (engl. *one-sided* tai *one-tailed z-test).*

Tällä tavalla muotoiltuna yksisuuntainen testi on omituinen, sillä

$$\Theta_0 \cup \Theta_1 \neq \Theta,$$

vaan parametriavaruudesta jätetään kokonaan huomioimatta ne μ, joille $\mu < \mu_0$. On vaikea sanoa esim., tehdäänkö virhe vai toimitaanko oikein, jos todellisuudessa $\mu < \mu_0$, mutta nollahypoteesi hylätään.

Näemme myöhemmin, että sama yksisuuntainen testi on koon α testi myös hypoteesiparille

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0.$$

Tälle hypoteesiparille

$$\Theta_0 \cup \Theta_1 = (-\infty, \mu_0] \cup (\mu_0, \infty) = \mathbb{R} = \Theta.$$

On järkevää ajatella, että yksisuuntaisella testillä (6.6) selvitetään tämän jälkimmäisen yhdistetyn nollahypoteesin $\mu \leq \mu_0$ paikkansapitävyyttä. Tämän testin kriittinen alue sattuu olemaan paljon helpompi johtaa, jos nollahypoteesina käytetään yksinkertaista hypoteesia $\mu = \mu_0$. Tämä lienee se ainoa syy, miksi tätä tarkkaa nollahypoteesin muotoilua lainkaan käytetään yksisuuntaiselle z-testille.

Vastaavilla laskuilla nähdään, että sekä hypoteesiparille

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu < \mu_0$$

että hypoteesiparille

$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

luottamustason α testi tekee päätöksen seuraavasti. Ensin lasketaan testisuureen arvo

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

Tällä kertaa pienet arvot (ts. itseisarvoltaan suureet negatiiviset arvot) ovat nollahypoteesille kriittisiä. Testi hylkää nollahypoteesin silloin (ja vain silloin), kun

$$z < -z_\alpha \tag{6.7}$$

Kaksisuuntainen testi

Nyt nollahypoteesi ja vastahypoteesi ovat

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0.$$

Testisuure on edelleen

$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

ja sitä vastaavalla satunnaismuuttujalla $t(\mathbf{Y})$ on $N(0, 1)$-jakauma, kun H_0 pitää paikkansa. Nyt sekä suuret että pienet testisuureen arvot ovat nollahypoteesin kannalta kriittisiä. Kaksisuuntainen (engl. *two-sided*, *two-tailed*) z-testi luottamustasolla α hylkää nollahypoteesin (täsmälleen) silloin, kun

$$|z| > z_{\alpha/2}. \tag{6.8}$$

Tämä perustuu siihen, että

$$P(|Z| > z_{\alpha/2}) = \alpha,$$

kun $Z \sim N(0, 1)$.

Fiktiivinen numeerinen esimerkki

Planeetalla Z seurataan tiiviisti JTP-kurssin aineistoja, koska paikalliset tutkijat ovat huomanneet kosmisen yhteyden planeetan Z ilmaston tilan ja JTP-kurssin simuloitujen aineistojen parametrien, erityisesti kuvan 4.3 aineiston parametrien välillä. Valitettavasti planeetalla Z ei osata suomea, vaan koko väestö puhuu englantia (hassusti murtaen). Tämän takia kukaan tutkija ei ole saanut selville, että oikeasti $\mu = 0.2012$. Sen sijaan ollaan saatu selville, että kyseessä on satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, jossa $\sigma^2 = 1$. Planeetalla Z suositaan z-testejä merkitsevyystasolla 0.05.

Pessimistisimmät tutkijat ovat sitä mieltä, että $\mu \leq 0$, mikä tarkoittaa käytännössä koko planeetan pikaista tuhoa. Tämän takia teemme z-testin, jossa hypoteesit ovat

$$H_0 : \mu \leq 0, \qquad H_1 : \mu > 0.$$

Aineistossa

$$\bar{y} = 0.726, \quad n = 10,$$

Nollahypoteesia vastaava, aineistosta laskettu z-arvo on

$$z = \frac{\bar{y} - 0}{\sigma/\sqrt{n}} = 2.296$$

Koska $z > z_{0.05} = 1.645$, *nollahypoteesi* $\mu \leq 0$ *hylätään* merkitsevyystasolla 5 %. Planeetan sanomalehdet kirjoittavat etusivuillaan, että tutkijat ovat todistaneet, että maailmanloppua ei tule.

Suurin osa tutkijoista ei kuitenkaan hyväksy hypoteesin $\mu \leq 0$ tieteellistä relevanssia, vaan uskoo teoriaan, jonka mukaan $\mu = \frac{1}{2}$, joka tarkoittaa sitä että planeetan ilmasto säilyy ikuisesti yhtä suotuisana kuin nykyään. Tämän takia teemme z-testin, jossa hypoteesit ovat

$$H_0 : \mu = \frac{1}{2}, \qquad H_1 : \mu \neq \frac{1}{2}.$$

Tätä nollahypoteesia vastaava aineistosta laskettu z-arvo on

$$z = \frac{\bar{y} - \frac{1}{2}}{\sigma/\sqrt{n}} = 0.715,$$

Koska $|z| \leq z_{\alpha/2} = 1.960$, niin *nollahypoteesi* $\mu = \frac{1}{2}$ *hyväksytään* merkitsevyystasolla 5 %. Planeetan sanomalehdet kirjoittavat etusivullaan, että tutkijat ovat todistaneet, että ilmasto säilyy ikuisesti suotuisana.

Mikä tulosten uutisoinnissa oli vikana? Nollahypoteesin hylkääminen tarkoittaa sitä, että ollaan löydetty todisteita sitä vastaan. Nollahypoteesin hyväksyminen ei tarkoita sitä, että oltaisiin löydetty todisteita nollahypoteesin puolesta. Se tarkoittaa sitä, että ei olla löydetty painavia todisteita nollahypoteesia vastaan.

Mitä varten tässä esimerkissä hölmö ja epätosi nollahypoteesi $H_0 : \mu = \frac{1}{2}$ hyväksyttiin?

Tähän asiaan saamme lisävalaistusta sen jälkeen, kun olemme nähneet, kuinka z-testin voima saadaan laskettua.

6.3 Testin voima

Määritelmä 6.2 (Testin voima)**.** Jos C on tarkasteltavan testin kriittinen alue, niin parametriavaruudella määriteltyä funktio

$$\pi(\theta) = P_\theta(t(\mathbf{Y}) \in C) = P_\theta(H_0 \text{ hylätään}) \tag{6.9}$$

on nimeltään testin *voima* (engl. *power*) tai sen voimafunktio (engl. *power function*).

Toisin sanoen, voimafunktio on testin hylkäystodennäköisyys parametrin funktiona. (Se mittaa *hylkäys*voimaa.)

Testin voiman voi ilmaista monimutkaisemmalla tavalla ensimmäisen ja toisen lajin virheiden todennäköisyyden avulla, nimittäin

$$\pi(\theta) = \begin{cases} P_\theta(\text{I lajin virhe}), & \text{kun } \theta \in \Theta_0, \\ 1 - P_\theta(\text{II lajin virhe}), & \text{kun } \theta \in \Theta_1. \end{cases}$$

Jos testin koko on α, niin testin koon määritelmän (6.3) ja sen voiman määritelmän (6.9) mukaan testin koko (ts. sen merkitsevyystaso) on

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta).$$

Merkitsevyystaso asettaa siis ylärajan testin voimalle nollahypoteesin mukaisilla parametrinarvoilla. Luonnollisesti tahtoisimme, että testin voima olisi mahdollisimman suuri vaihtoehtohypoteesin mukaisilla parametrinarvoilla.

Testin voimaa on syytä tarkastella tutkimuksen suunnitteluvaiheessa. Yleisesti ottaen, *mitä suurempi on otoskoko, sitä suurempi on testin voima* (vaihtoehtohypoteesin mukaisilla parametrinarvoilla). Tavallisesti tutkijalla on käsitys siitä, miten suuret poikkeamat nollahypoteesin mukaisista parametrinarvoista ovat käytännössä merkittäviä. Tällöin otoskoko voidaan yrittää valita siten, että saavutetaan vähintään jokin annettu voima (esim. vähintään 80 %) aina, kun poikkeama nollahypoteesista on käytännön kannalta merkittävä.

Tietyt tieteelliset lehdet vaativat laskelmaa tilastollisen testin voimasta, mikäli sellaisia artikkelissa esitetään. Esim. *American Psychological Association* -yhdistyksen lehtien julkaisuohjeissa (6. laitos, v. 2010) ohjeistetaan

> routinely provide evidence that the study has sufficient power to detect effects of substantive interest.

6.4 Testin p-arvo

Vanhanaikainen tapa raportoida testin tulos on kertoa testin koko α sekä kertoa testin päätös, eli hylkäsikö vai hyväksyikö testi nollahypoteesin. Nykyään on tapana kertoa tämän lisäksi (tai sijasta) testin *p-arvo* (engl. *p-value*) eli *havaittu merkitsevyystaso* (engl. *observed significance level*). Testin p-arvo mittaa tietyllä tavalla nollahypoteesin ja aineiston yhteensopivuutta.

Testin p-arvon määrittely on hieman erilainen sen mukaan, mitkä testisuureen arvot ovat nollahypoteesille kriittisiä. Mikäli testisuureen $t(\mathbf{y})$ suuret arvot ovat nollahypoteesille kriittisiä, niin p-arvo määritellään kaavalla

$$p = p(\mathbf{y}) = \sup_{\theta \in \Theta_0} P_\theta[t(\mathbf{Y}) \geq t(\mathbf{y})] \tag{6.10}$$

Tässä $t(\mathbf{y})$ on havaitusta aineistosta $\mathbf{y}$ laskettu testisuureen arvo, ja $t(\mathbf{Y})$ on satunnaisvektorista $\mathbf{Y}$ laskettu testisuureen arvo, kun sillä on jakaumana mallin mukainen yptnf/ytf $f(\mathbf{y}; \theta)$. Useissa tilanteissa tässä merkitty todennäköisyys ei riipu siitä, mitä nollahypoteesin mukaista parametrinarvoa $\theta \in \Theta_0$ tarkastellaan, jolloin p-arvo voidaan määritellä sanallisesti seuraavasti.

> p-arvo on se todennäköisyys, jolla nollahypoteesin mukaisesta populaatiosta saadaan testisuureen arvo, joka on vähintään yhtä kummallinen kuin aineistosta laskettu testisuureen arvo.

Kummallisuutta mitataan testisuureen arvolla: kummallisempia ovat ne arvot jotka poikkeavat vielä enemmän nollahypoteesille kriittiseen suuntaan kuin havaittu arvo. Usein kummallisuuden sijasta sanotaan “vähintään yhtä äärevä” (engl. *at least as extreme as*).

Jos p-arvo on pieni (esim. 1 %), niin tällöin hyvin pienellä todennäköisyydellä nollahypoteesin mukaisesta populaatiosta saadaan vähintään yhtä kummallisia havaintoja kuin mitä todellisuudessa saatiin. Jos taas p-arvo on suuri (esim. 20 %), niin tällöin kohtuullisen usein nollahypoteesin mukaisesta populaatiosta saadaan havaintoja, jotka ovat vähintään yhtä kummallisia kuin mitä todellisuudessa havaittiin. Toisin sanoen

> pieni p-arvo viittaa ristiriitaan aineiston ja nollahypoteesin välillä.

Jos p-arvon määritelmässä (6.10) oleva todennäköisyys kuitenkin riippuu parametrinarvosta $\theta \in \Theta_0$, niin oikea sanallinen määritelmä on

- p-arvo on pienin yläraja sille todennäköisyydelle, jolla nollahypoteesin mukaisesta populaatiosta saadaan testisuureen arvo, joka on vähintään yhtä kummallinen kuin aineistosta laskettu testisuureen arvo.

Testin päätös saadaan luettua sen p-arvosta seuraavalla tavalla.

> Testi hylkää H_0:n, jos $p < \alpha$. Muussa tapauksessa H_0 jää voimaan.

Osoitamme esimerkeissä, että näin menetellen testin kooksi tulee α.

Nykyään on tapana ilmoittaa testin p-arvo (parilla desimaalilla), ja lisäksi varmuuden vuoksi kommentoida (asiantuntematonta lukijaa ajatellen), tulisko nollahypoteesi hylättyä vai hyväksyttyä konventionaalisilla merkitsevyystasoilla. Tämä on paljon informatiivisempaa kuin kertoa vain testin päätös jollakin kiinteällä merkitsevyystasolla.

6.5 z-testin p-arvo ja voima

Laskemme seuraavaksi jaksossa 6.2 käsitellyn z-testin p-arvon ja voimafunktion sekä yksi- että kaksisuuntaisessa tapauksessa.

Yksisuuntainen z-testi

Yksisuuntaisessa testissä

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu > \mu_0,$$

jossa testisuure lasketaan kaavalla

$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

ovat suuret testisuureen arvot kriittisiä, ja $t(\mathbf{Y}) \sim N(0, 1)$, kun nollahypoteesi pitää paikkansa. Tämän takia testin p-arvo on

$$p = P_{\mu_0}(Z \geq z) = 1 - \Phi(z),$$

jossa Φ on $N(0, 1)$-jakauman kertymäfunktio, $Z \sim N(0, 1)$, ja z on aineistosta laskettu testisuureen arvo.

Tämä testi hylkää täsmälleen silloin, kun

$$\begin{aligned} & z > z_\alpha \\ \Leftrightarrow \quad & \Phi(z) > 1 - \alpha \\ \Leftrightarrow \quad & p = 1 - \Phi(z) < \alpha. \end{aligned}$$

Tällä tavalla olemme tarkistaneet yksisuuntaisen z-testin kohdalla, että testin päätös voidaan lukea sen p-arvosta.

Voimafunktio on

$$\pi(\mu) = P_\mu[t(\mathbf{Y}) > z_\alpha] = P_\mu\left[\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha\right],$$

ja tässä satunnaismuuttujan $(\bar{Y} - \mu_0)/(\sigma/\sqrt{n})$ jakauma on $N((\mu - \mu_0)/(\sigma/\sqrt{n}), 1)$, kun todellinen parametrinarvo on μ, vrt. kuva 6.1. Siis

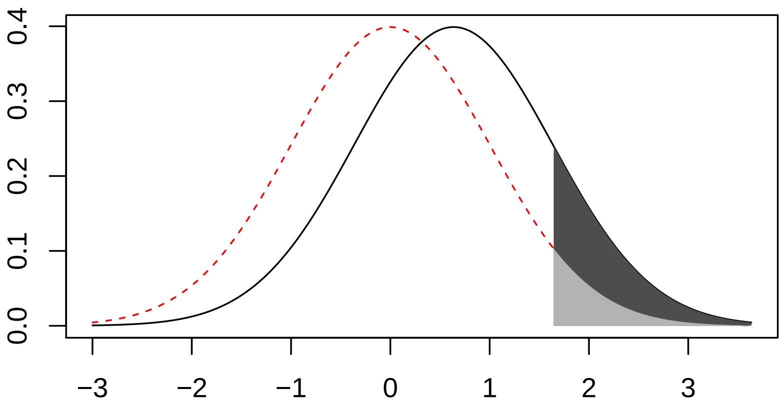
$$\begin{aligned} \pi(\mu) &= P_\mu\left[\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} > z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\ &= P\left[Z > z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\ &= 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}} - z_\alpha\right). \end{aligned}$$

Viimeinen yhtäsuuruus perustuu $N(0, 1)$-jakauman symmetrisyyteen, minkä takia sen kertymäfunktio toteuttaa identiteetin

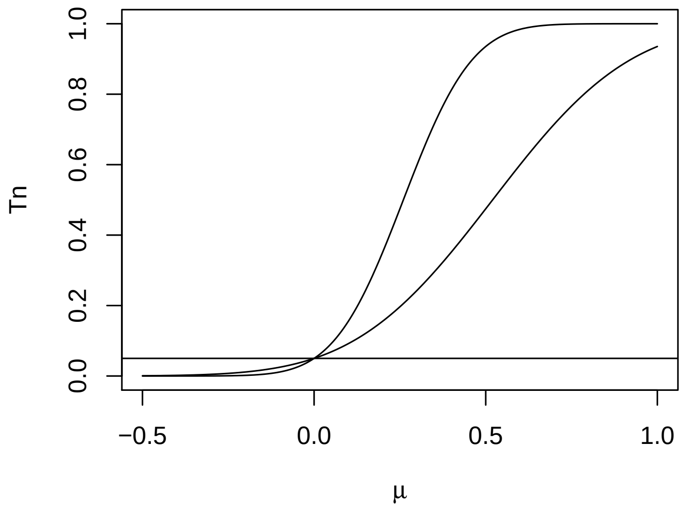
$$1 - \Phi(a) = \Phi(-a), \qquad \text{kaikilla } a.$$

Kuvassa 6.2 näytetään tämän testin voimafunktio, kun testin koko on $\alpha = 0.05$. Vaihtoehtohypoteesin mukaisilla parametrinarvoilla suuremmalla otoskoolla saavutetaan suurempi voima.

Kuva 6.1 Hypoteesin $H_0 : \mu = 0$ voiman laskeminen, kun todellisuudessa $\mu = 0.2012$. Normaalijakauman varianssi $\sigma^2 = 1$ on oletettu tunnetuksi. Normaalijakauman $N(0, 1)$ häntäalueen pinta-ala on $\alpha = 0.05$, ja testin voima on normaalijakauman $N((\mu - 0)/(\sigma/\sqrt{n}), 1)$ kuvaan merkityn häntäalueen pinta-ala.



Kuva 6.2 Yksisuuntaisen z-testin voimafunktio, kun $\alpha = 0.05$, $\mu_0 = 0$, $\sigma = 1$, ja otoskoko on $n = 10$ tai $n = 40$. Suuremmalla otoskoolla saavutetaan suurempi voima vastahypoteesin $\mu > \mu_0$ mukaisilla parametrinarvoilla. Testin koko on osoitettu vaakaviivalla.



Äsken johdetusta voimafunktion kaavasta (sekä kuvasta) nähdään, että se on aidosti kasvava funktio, minkä takia

$$\pi(\mu) \leq \pi(\mu_0) = \alpha, \quad \text{kaikilla } \mu \leq \mu_0.$$

Tämän takia käsittelemämme yksisuuntainen z-testi on tason α testi myös hypoteesiparille

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0.$$

Tälle hypoteesiparille

$$\Theta_0 \cup \Theta_1 = (-\infty, \mu_0] \cup (\mu_0, \infty) = \mathbb{R} = \Theta,$$

kuten aikaisemmin jo mainittiin.

Toisen mahdollisen yksisuuntaisen testin

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu < \mu_0$$

p-arvo on

$$p = P_{\mu_0}(Z \leq z) = \Phi(z),$$

jossa Φ on $N(0,1)$-jakauman kertymäfunktio, $Z \sim N(0,1)$, ja z on aineistosta laskettu testisuureen arvo, eli

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

jonka pienet arvot ovat nollahypoteesille kriittisiä.

Tämän testin voimafunktio on

$$\begin{aligned}
\pi(\mu) &= P_\mu[t(\mathbf{Y}) \leq -z_\alpha] \\
&= P\left[Z \leq -z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\
&= \Phi\left(-z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Tämä funktio on aidosti vähenevä, joten

$$\pi(\mu) \leq \pi(\mu_0) = \alpha, \quad \text{kaikilla } \mu \geq \mu_0.$$

Voimafunktion kuvaaja on peilikuva ensimmäiseksi käsitellyn yksisuuntaisen z-testin voimafunktion kuvaajasta. Tämä yksisuuntainen z-testi on tason α testi myös hypoteesiparille

$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

Kaksisuuntainen z-testi

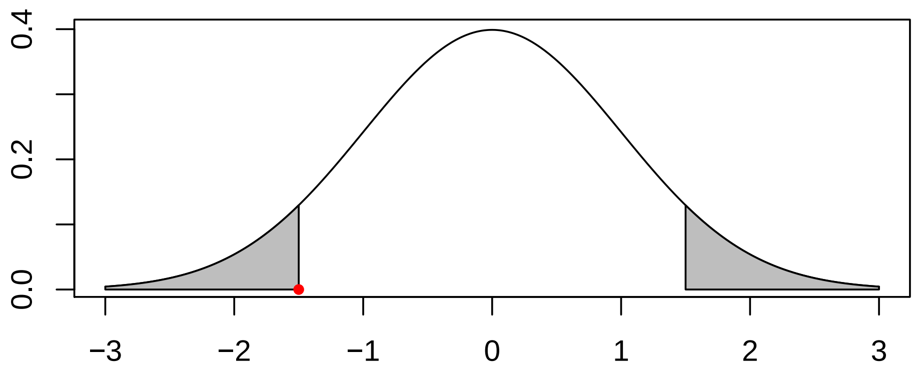
Kaksisuuntaisen testin

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

testaamisessa testisuureen

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

Kuva 6.3 Kaksisuuntaisen testin p-arvon määrittäminen, kun havainnosta saadaan $z = -1.5$.



sekä suuret että pienet (negativiset) arvot ovat nollahypoteesille kriittisiä.

P-arvo määritellään summaamalla häntätodennäköisyys molemmilta viitejakauman hänniltä, ts. kummallisia tai vielä kummallisempi arvoja ovat ne, joille

$$|Z| \geq |z|,$$

ks. kuva 6.3. Tällä perusteella

$$p = P[|Z| \geq |z|] = \Phi(-|z|) + 1 - \Phi(|z|) = 2(1 - \Phi(|z|))$$

Kaksisuuntainen z-testi hylkää täsmälleen silloin, kun

$$\begin{aligned} & |z| > z_{\alpha/2} \\ \Leftrightarrow \quad & \Phi(|z|) > 1 - \frac{\alpha}{2} \\ \Leftrightarrow \quad & p = 2(1 - \Phi(|z|)) < \alpha. \end{aligned}$$

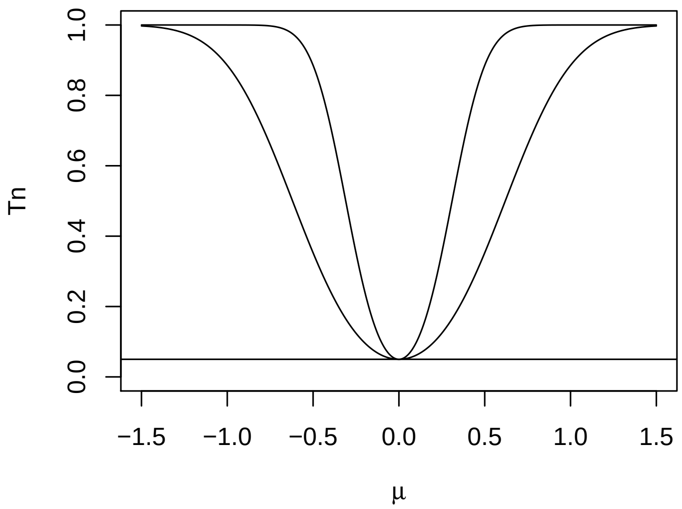
Nyt olemme tarkistaneet myös kaksisuuntiselle z-testille, että testin päätös voidaan lukea sen p-arvosta.

Kaksisuuntaisen testin voimafunktio on

$$\begin{aligned} \pi(\mu) &= P_\mu[|t(\mathbf{Y})| \geq z_{\alpha/2}] \\ &= P_\mu\left[\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}\right] + P_\mu\left[\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{\alpha/2}\right] \\ &= P\left[Z \geq z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] + P\left[Z \leq -z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right] \\ &= 1 - \Phi\left(z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Kuvassa 6.4 näytetään kaksisuuntaisen z-testin voimafunktio, kun testin koko on $\alpha = 0.05$. Suuremmalla otoskoolla saavutetaan suurempi voima.

Kuva 6.4 Kaksisuuntaisen z-testin voimafunktio, kun $\alpha = 0.05$, $\mu_0 = 0$, $\sigma = 1$, ja otoskoko on $n = 10$ tai $n = 40$. Suuremmalla otoskoolla saavutetaan suurempi voima vastahypoteesin $\mu \neq \mu_0$ mukaisilla parametrinarvoilla. Testin koko on osoitettu vaakaviivalla.



p-arvo ja voima numeerisessa esimerkissä

Palaamme planeetalle Z. Aineiston yhteenveto oli

$$\bar{y} = 0.726, \quad n = 10, \quad \sigma^2 = 1.$$

Aluksi testattiin yksisuuntaista hypoteesia

$$H_0 : \mu \leq 0, \qquad H_1 : \mu > 0,$$

joka hylättiin merkitsevyystasolla $\alpha = 0.05$. Tämän testin p-arvo ja testin voima todelliselle parametrinarvolle $\mu = 0.2012$ saadaan laskettua R:llä seuraavasti. Tässä funktio `pnorm` laskee normaalijakauman kertymäfunktion arvon.

```
mean.y <- 0.726
n <- 10
sigma <- 1
mu0 <- 0
alpha <- 0.05
mu.true <- 0.2012
sem <- sigma/sqrt(n)
z <- (mean.y - mu0)/sem
p <- 1 - pnorm(z)
zcrit <- qnorm(lower = FALSE, alpha)
pwr <- pnorm((mu.true - mu0)/sem - zcrit)
c(z, zcrit, p, pwr)

## [1] 2.29581 1.64485 0.01084 0.15658
```

Tulostuksista näemme, että testin p-arvo $p = 0.011$, joten tämä on yläraja sille todennäköisyydelle, että nollahypoteesin mukaisesta populaatiosta saadaan z-tunnusluvun arvo, joka on suurempi tai yhtä suuri kuin aineistosta laskettu z-tunnusluvun arvo. Testin voima todelliselle parametrinarvolle on vain 0.16. Ennen aineiston simulointia todennäköisyys, että testi tulee hylkäämään nollahypoteesin $\mu \leq 0$ oli (ainoastaan) 0.16, joten olimme aika onnekkaita koska pystyimme (todellisen tilanteen mukaisesti) hylkäämään tämän nollahypoteesin.

Toisen tutkijayhteisön mielipiteen mukaisesti seuraavaksi tehtiin testi

$$H_0 : \mu = \frac{1}{2}, \qquad H_1 : \mu \neq \frac{1}{2},$$

joka hyväksyttiin merkitsevyystasolla $\alpha = 0.05$. Tässä tapauksessa p-arvo ja testin voima voidaan laskea seuraavasti.

```
mu0 <- 0.5
z <- (mean.y - mu0)/sem
zcrit <- qnorm(alpha/2, lower = FALSE)
p <- 2 * (1 - pnorm(abs(z)))
pwr <- 1 - pnorm(zcrit - (mu.true - mu0)/sem) + pnorm(-zcrit -
    (mu.true - mu0)/sem)
c(z, zcrit, p, pwr)

## [1] 0.7147 1.9600 0.4748 0.1569
```

Nyt testin p-arvo on $p = 0.47$, joka on se todennäköisyys, että nollahypoteesin mukaisesta populaatiosta saadaan z-tunnusluvun arvo, joka on itseisarvoltaan vähintään yhtä suuri kuin aineistosta laskettu z-tunnusluvun itseisarvo. Testin voima todelliselle parametrinarvolle on (taas) noin 0.16. Ennen aineiston simulointia todennäköisyys, että testi tulee hylkäämään nollahypoteesin $\mu = \frac{1}{2}$ oli 0.16. Tässä tapauksessa testin voima on niin pieni, että hyväksymispäätöstä ei pitäisi tulkita todisteena H_0:n puolesta, mikäli $\mu = 0.2012$ on käytännön eli planeetan Z ilmaston kannalta merkittävästi erilainen kuin arvo $\mu_0 = \frac{1}{2}$.

Tämän tarinan eräänä opetuksena voidaan pitää sitä, että pienellä otoskoolla testin voima on helposti niin pieni, että emme voi sen avulla havaita käytännön kannalta tärkeitä eroja nollahypoteesista. Tällaisessa tilanteessa pitäisi yrittää hankkia suurempi otos.

6.6 Testien ja luottamusvälien duaalisuus

Tarkastelemme esimerkin vuoksi kaksisuuntaisen z-testin

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

hyväksymisaluetta, eli niitä μ_0 joita tason α testi (6.8) ei hylkää. Testi ei hylkää, mikäli $|z| \leq z_{\alpha/2}$ eli mikäli

$$\begin{aligned} & -z_{\alpha/2} \leq \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \\ \Leftrightarrow \quad & \bar{y} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu_0 \leq \bar{y} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \end{aligned}$$

Taulukko 6.1 Merkitsevyystason $0 < \alpha < 1$ testejä ja luottamusvälejä normaalijakautuneelle populaatiolle $N(\mu, \sigma^2)$, kun varianssi σ^2 on tunnettu. Tässä $z = (\bar{y} - \mu_0)/(\sigma/\sqrt{n})$, ja $Z \sim N(0,1)$ ja z_u on $N(0,1)$-jakauman u-yläkvantiili.

H_0	H_1	Hylkäysalue	p-arvo	Luottamusväli
$\mu \le \mu_0$	$\mu > \mu_0$	$z > z_\alpha$	$P(Z \ge z)$	$[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$
$\mu \ge \mu_0$	$\mu < \mu_0$	$z < -z_\alpha$	$P(Z \le z)$	$(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}]$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert z\rvert > z_{\alpha/2}$	$P(\lvert Z\rvert \ge \lvert z\rvert)$	$[\bar{y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$

mutta tunnistamme, että alempi epäyhtälöpari määrittelee tason $1 - \alpha$ kaksisuuntaisen z-luottamusvälin. Toisin sanoen, z-testi ei hylkää nollahypoteesia $\mu = \mu_0$ täsmälleen silloin, kun μ_0 kuuluu luottamustason $1 - \alpha$ luottamusväliin (5.10).

Voimme sanoa, että kaksisuuntainen z-luottamusväli saadaan kääntämällä kaksisuuntaisen z-testin hyväksymisalue. Tarkemmin sanoen, ratkaisemme kaikkien muotoa $H_0 : \mu = \mu_0$ olevien kaksisuuntaisten testien hyväksymisalueet. Voimme samaan tapaan kääntää myös yksisuuntaisten z-testien hyväksymisalueet, ja näin menetellen saadaan taulukossa 6.1 luetellut tapaukset.

Tässä valossa testin p-arvolle voidaan antaa uusi tulkinta. Mikäli $p > \alpha$, niin tällöin μ_0 kuuluu testiä vastaavaan luottamustason $1 - \alpha$ luottamusväliin. Jos taas $p < \alpha$, niin μ_0 ei kuulu ko. luottamusväliin.

Merkitsevyystason α testit ja luottamustason $1 - \alpha$ luottamusvälit yrittävät antaa vastauksen samantapaiseen kysymykseen, mutta erilaisista näkökulmista. Testissä kiinnitetään yksi parametrinarvo, ja tutkitaan ovatko havainnot sopusoinnussa tämän parametrinarvon kanssa. Luottamusväli yrittää kertoa suoraan, mitkä parametrinarvot ovat sopusoinnussa havaintojen kanssa. Tähän testien ja luottamusvälien vastaavuuteen voidaan viitata sanomalla, että ne ovat duaalisia käsitteitä.

Planeetan Z esimerkissä yksisuuntaista hypoteesia

$$H_0 : \mu \le \mu_0, \quad H_1 : \mu > \mu_0$$

vastaa parametrin μ yksisuuntainen 95 %:n luottamusväli $[0.20, \infty)$, joka ei sisällä esimerkissä kiinnostavaa arvoa 0, joten nollahypoteesi $H_0 : \mu \le 0$ hylätään merkitsevyystasolla 0.05. Testin p-arvo on $p = 0.011 < 0.05$, joten myös tästä nähdään, että $\mu_0 = 0$ ei kuulu ko. 95 %:n yksisuuntaiseen luottamusväliin ilman että luottamusväliä edes tarvitsee laskea.

Kaksisuuntaista hypoteesia

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \ne \mu_0$$

vastaa parametrin μ kaksisuuntainen 95 %:n luottamusväli $[0.10, 1.35]$, joka sisältää esimerkissä kiinnostavan arvon $\frac{1}{2}$, joten nollahypoteesi $\mu = \frac{1}{2}$ hyväksytään merkitsevyystasolla 0.05. Testin p-arvo on $p = 0.47$, mistä myöskin nähdään, että $\mu_0 = \frac{1}{2}$ kuuluu kaksisuuntaiseen 95 %:n luottamusväliin.

Testaaminen johtaa käyttäjän helposti mustavalkoiseen ajatteluun: nollahypoteesi joko hyväksytään tai hylätään. Tällöin käyttäjän huomio kiinnittyy pois siitä, kuinka epävarmaa aineiston antama informaatio parametrista on.

Taulukko 6.2 Merkitsevyystason $0 < \alpha < 1$ testejä ja luottamusvälejä normaalijakautuneen populaation $N(\mu, \sigma^2)$ odotusarvolle μ, kun myös varianssi σ^2 on tuntematon. Tässä $t = (\bar{y} - \mu_0)/(s/\sqrt{n})$, s on otoskeskihajonta, $T \sim t_{n-1}$ ja $t_{n-1}(u)$ on t_{n-1}-jakauman u-yläkvantiili.

H_0	H_1	Hylkäysalue	p-arvo	Luottamusväli
$\mu \le \mu_0$	$\mu > \mu_0$	$t > t_{n-1}(\alpha)$	$P(T \ge t)$	$[\bar{y} - t_{n-1}(\alpha)\frac{s}{\sqrt{n}}, \infty)$
$\mu \ge \mu_0$	$\mu < \mu_0$	$t < -t_{n-1}(\alpha)$	$P(T \le t)$	$(-\infty, \bar{y} + t_{n-1}(\alpha)\,\frac{s}{\sqrt{n}}]$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert t\rvert > t_{n-1}(\frac{\alpha}{2})$	$P(\lvert T\rvert \ge \lvert t\rvert)$	$[\bar{y} - t_{n-1}(\frac{\alpha}{2})\,\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\frac{\alpha}{2})\,\frac{s}{\sqrt{n}}]$

Sen sijaan luottamusväli kvantifioi epävarmuuden selkeällä tavalla. Kun piste-estimaatti sekä luottamusväli lasketaan käytännön kannalta kiinnostavalle parametrille, niin saadaan tuloksia, jotka voidaan tulkita suoraan. Sen sijaan asiantuntemattomat lukijat tulkitsevat testien tulokset toisinaan aivan nurinkurisella tavalla. Testauksessa tutkijalla pitäisi olla selkeä käsitys testin voimafunktiosta sellaisilla parametrinarvoilla, jotka ovat käytännön kannalta merkityksellisiä.

Vaikka testit ja luottamusvälit ovat duaalisia käsitteitä, niin sellaisissa yksinkertaisissa tilanteissa joissa molemmat lähestymistavat ovat mahdollisia *luottamusvälien laskeminen on parempi tapa analysoida aineistoa kuin testaaminen.*

6.7 Normaalijakautuneen populaation odotusarvon testaus, kun myös varianssi on tuntematon

Jos sekä odotusarvo että varianssi ovat tuntemattomia, niin testit perustetaan saranasuureelle

$$t(\mathbf{Y}) = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}},$$

jolla on t-jakauma vapausasteluvulla $n-1$, mikäli $\mu = \mu_0$. Tästä tiedosta saadaan johdettua t-testit eri tapauksille matkimalla z-testien johtoa. Tulokset on koottu taulukkoon 6.2. Käytännössä t-testi suoritetaan aina jollakin tarkoitukseen sopivalla tietokoneohjelmalla. Esim. R-ohjelmistosa kaikki taulukon 6.2 tulokset saadaan laskettua vaivattomasti funktiolla `t.test`.

Jos esimerkiksi vastahypoteesi on kaksisuuntainen

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \ne \mu_0,$$

ja μ_0 on kiinteä luku, niin testisuure lasketaan kaavalla

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}},$$

jossa s on otoskeskihajonta. Vastaavalla satunnaismuuttuja $t(\mathbf{Y})$ noudattaa t-jakaumaa $n-1$ vapausasteella, ja testisuureen sekä pienet että suuret arvot ovat kriittisiä. Testi hylkää nollahypoteesin, mikäli

$$|t| > t_{n-1}(\frac{\alpha}{2}),$$

ja sen p-arvo on

$$p = P(|T| \geq |t|) = 2(1 - F_{n-1}(|t|)),$$

jossa satunnaismuuttuja $T \sim t_{n-1}$, ja F_ν tarkoittaa t_ν-jakauman kertymäfunktiota. Kaksisuuntainen t-luottamusväli saadaan ratkaisemalla kaikki ne μ_0-arvot, joilla t-testi hyväksyisi nollahypoteesin $H_0 : \mu = \mu_0$ kaksisuuntaiselle vastahypoteesille:

$$\begin{aligned}
& \left| \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right| \leq t_{n-1}\left(\frac{\alpha}{2}\right) \\
\Leftrightarrow \quad & -t_{n-1}\left(\frac{\alpha}{2}\right) \leq \frac{\mu_0 - \bar{y}}{s/\sqrt{n}} \leq t_{n-1}\left(\frac{\alpha}{2}\right) \\
\Leftrightarrow \quad & \bar{y} - \frac{s}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right) \leq \mu_0 \leq \bar{y} + \frac{s}{\sqrt{n}} t_{n-1}\left(\frac{\alpha}{2}\right)
\end{aligned}$$

Muut taulukon 6.2 rivit saadaan järkeiltyä samaan tapaan.

Voimafunktion laskeminen t-testille on monimutkaisempaa kuin z-testille, mutta tämä kuitenkin onnistuu ns. epäkeskisen t-jakauman avulla. Tuloksena saadaan samantapaisia käyriä kuin z-testin voimafunktiolle.

6.8 Binomijakauman parametrin testaus

Jos tahdotaan testata lantin harhattomuutta, niin testi voidaan perustaa suoraan onnistumisten lukumäärälle $X \sim \mathrm{Bin}(n, p)$. (Onnistuminen voi nyt olla yhtä kuin kruunan saaminen yhdellä lantin heitolla.) Tässä tapauksessa hypoteesit ovat

$$H_0 : p = \frac{1}{2}, \qquad H_1 : p \neq \frac{1}{2}.$$

Testi voidaan perustaa sille idealle, että tutkitaan havainnon poikkeamaa nollahypoteesin mukaisen jakauman odotusarvosta $EX = n/2$. Kaksisuuntaisen testin p-arvo voidaan laskea kaavalla

$$p = P_{1/2}\left(\left| X - \frac{n}{2} \right| \geq \left| x - \frac{n}{2} \right| \right).$$

Tässä alaindeksi $1/2$ tarkoittaa sitä, että oletamme nollahypoteesin mukaisesti, että $X \sim \mathrm{Bin}(n, 1/2)$, ja x on havaittu onnistumisten lukumäärä.

Tämän testin p-arvo saadaan laskettua karkeasti käyttämällä normaaliapproksimaatiota, jonka mukaan nollahypoteesin vallitessa satunnaismuuttujalla

$$\frac{X - EX}{\sqrt{\operatorname{var} X}} = \frac{X - \frac{n}{2}}{\sqrt{n \frac{1}{2} \frac{1}{2}}}$$

on osapuilleen standardinormaalijakauma $N(0, 1)$. Tämän takia p-arvo saadaan osapuilleen kaavalla

$$2\left(1 - \Phi\left(\frac{2}{\sqrt{n}} \left| x - \frac{n}{2} \right| \right)\right),$$

tai kaavalla

$$2\left(1 - \Phi\left(\frac{2}{\sqrt{n}} \left(\left| x - \frac{n}{2} \right| - \frac{1}{2} \right)\right)\right),$$

Jälkimmäisessä kaavassa tehtiin jatkuvuuskorjaus, ja Φ on $N(0, 1)$-jakauman kertymäfunktio.

Jos $n = 1000$ ja onnistumisia on tullut $x = 460$, niin nämä p-arvon normaaliapproksimaatiot ovat seuraavat.

```
n <- 1000
x <- 460
2 * pnorm(lower = FALSE, 2/sqrt(n) * abs(x - n/2))

## [1] 0.01141

2 * pnorm(lower = FALSE, 2/sqrt(n) * (abs(x - n/2) - 1/2))

## [1] 0.01248
```

Binomijakauman häntätodennäköisyydet saadaan laskettua tarkasti tietokoneohjelmilla. Esimerkiksi, jos $n = 1000$ ja onnistumisia on tullut $x = 460$, niin testin p-arvo saadaan selvitettyä ilman approksimaatioita komennolla

```
binom.test(460, 1000, p = 0.5)

##
##  Exact binomial test
##
## data:  460 and 1000
## number of successes = 460, number of trials = 1000, p-value =
## 0.01244
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4288 0.4915
## sample estimates:
## probability of success
##                   0.46
```

Voimme samaan tapaan käsitellä muutkin kaksisuuntaiset testit

$$H_0 : p = p_0, \qquad H_1 : p \neq p_0,$$

ja kääntämällä näiden testien hyväksymisalueet saadaan jaksossa 5.8 mainittu Clopperin ja Pearsonin tarkka luottamusväli parametrille p.

Yksisuuntaiset testit saadaan käsiteltyä samaan tapaan.

6.9 p-arvo ei ole todennäköisyys sille, että nollahypoteesi pitää paikkansa

P-arvoa voidaan ajatella mittana sille, miten hyvin havainto on sopusoinnussa nollahypoteesin kanssa. Koska tämä käsite on vaikeatajuinen, tilastotieteen soveltajilla on kaikenlaisia harhakäsityksiä siitä, mitä p-arvo tarkoittaa.

Eräs yleinen harhakäsitys on se, että p-arvo on todennäköisyys sille, että nollahypoteesi pitää paikkansa. Jotta tällaiseen täysin virheelliseen käsitykseen voisi päätyä, pitää tehdä monta räikeää käsitteellistä virhettä, kuten esimerkiksi seuraavat.

1. Unohdetaan, että toimitaan frekventistisen tilastotieteen puitteissa. Frekventistisessä tilastotieteessä parametrinarvoihin tai tilastollisiin hypoteeseihin ei liitetä todennäköisyyksiä.

2. Tämän sekaannuksen lisäksi ajatellaan, että $P(H_0 \mid Y = y)$ on sama asia kuin $P(Y = y \mid H_0)$.

3. Lopuksi vielä ajatellaan, että p-arvo on sama asia kuin $P(Y = y \mid H_0)$, mitä se ei ole. P-arvo on häntätodennäköisyys, tarkemmin sanoen todennäköisyys sille, että nollahypoteesin mukaisesta populaatiosta saadaan arvo, joka on yhtä kummallinen tai vielä kummallisempi kuin havaittu arvo.

Testin p-arvo ei ole todennäköisyys sille, että nollahypoteesi pitää paikkansa, vaikka näin lukijalle kerrotaan lukuisissa tilastotieteen soveltajille tarkoitetussa oppikirjoissa.

6.10 Tilastollisten testien väärinkäyttöä

Monille tilastotieteen soveltajille on syntynyt sellainen mielikuva, että kokeellisen tutkimuksen päämääränä on laskea p-arvo jollekin testille. Jos p-arvo on riittävän pieni (esim. $p < 0.05$), niin tuloksen saa julkaistua jossakin alan lehdessä. Jos p-arvo ei ole riittävän pieni, tutkimusta ei kannata lähettää arvioitavaksi, koska sitä ei kuitenkaan tulla julkaisemaan. Valitettavasti tämä harha ei koske yksinomaan yksittäisiä tutkijoita, vaan tällainen käsitys on ollut yleinen myös vaikutusvaltaisten lehtien arvioijien ja toimittajien parissa. Tällainen käytäntö johtaa *julkaisuharhaan* (engl. *publication bias*): kirjallisuudessa julkaistaan enimmäkseen nollahypoteesin hylkääviä tutkimuksia riippumatta siitä, mikä todellisuudessa on asian laita. Tällainen käytäntö perustuu väärinkäsityksiin, rituaaleihin ja taikauskoon eikä sillä ole mitään tekemistä kunnollisen tieteellisen tutkimuksen kanssa eikä kunnollisen tilastotieteen soveltamisen kanssa (ks. esim. [3] tai [2]).

Lisäksi useimmiten julkaisuissa testataan nollahypoteeseja, joista jo ennen tutkimuksen tekoa tiedetään, että ne eivät voi pitää paikkaansa. Nämä ovat ns. hölmöjä nollahypoteeseja (engl. *silly null*). Käsittelyllä on todellisuudessa kuitenkin aina jokin vaikutus, joten nollahypoteesi $\mu = 0$ (ei vaikutusta) ei voi pitää kirjaimellisesti paikkaansa, eikä kukaan oikeasti usko tätä tarkkaa, pistemäistä (engl. *sharp null*, *point null*) nollahypoteesia. Jos paikkansa pitämätöntä pistemäistä nollahypoteesia ei saada testillä hylättyä, niin syynä on se, että testillä ei ollut riittävästi voimaa, eli otoskoko oli liian pieni.

Jos taas otoskokoa kasvatetaan, ja vihdoin pystytään nollahypoteesi hylkäämään, niin voi olla, että aineiston nojalla arvioitu vaikutuksen suuruus on niin pieni, että sillä ei ole mitään käytännön merkitystä.

> *Kysymys:* Jos nollahypoteesin mukaan $\mu = 0$, ja paras estimaattimme vaikutuksen suuruudelle on $\hat{\mu} = 0.1$, niin onko tällä erolla käytännössä merkitystä?

Tämä on kysymys, johon tilastotiede ei pysty antamaan vastausta. Vastauksen pitää tulla substanssialan asiantuntijalta. (Mitä asiaa mitattiin? Mitä yksikköjä käytettiin? jne.) Tilastollinen merkitsevyys (engl. *statistical significance*) ja käytännön merkittävyys (engl. *practical significance*) ovat aivan eri asioita.

On hedelmällisempää ja informatiivisempaa yrittää estimoida vaikutuksen suuruutta ja yrittää kvantifioida estimaattiin liittyvää epävarmuutta (keskivirhe, luottamusväli!) kuin yrittää testata, onko vaikutus nolla.

Tilastotieteen soveltajien intoon testata kaikkea mahdollista viitataan usein lyhenteellä NHST (*null hypothesis significance testing*). Hakusanan NHST avulla on helppo löytää tämän käytännön kritiikkiä.

Joillakin tilastotieteeseen vahvasti nojaavilla aloilla (esim. lääketiede, terveystieteet ja psykologia) on käynnissä uudistusliike, jossa pyritään pois epätarkoituksenmukaisesta hypoteesien testauksesta. Tämän sijasta

- lasketaan piste-estimaatteja ja väliestimaatteja vaikutuksen suuruudelle,
- yhdistetään aikaisempien tutkimusten tuloksia, eli harrastetaan meta-analyysia.

Jotkut kirjoittajat käyttävät tästä uudistusliikkeestä nimitystä uusi tilastotiede (engl. *new statistics*) [1] — tilastotieteen näkökulmasta uudessa tilastotieteessä ei ole paljoa uutta.

Kerrataan lopuksi vielä seuraavat asiat:

- Nollahypoteesin hyväksyminen testissä ei tarkoita sitä, että oltaisiin löydetty todisteita nollahypoteesin puolesta. Se tarkoittaa sitä, että ei olla löydetty riittävän painavia todisteita nollahypoteesia vastaan.
- Testin tekemän päätöksen voi lukea p-arvosta, joka mittaa sitä kuinka hyvin aineisto on sopusoinnussa nollahypoteesin kanssa.
- Testin p-arvo ei ole todennäköisyys sille, että nollahypoteesi pitää paikkansa.
- Testaamisen sijasta kannattaa laskea piste-estimaatteja ja luottamusvälejä, mikäli tämä on mahdollista.

Kirjallisuutta

[1] Geoff Cumming. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* Routledge, 2012.

[2] Esa Läärä. Statistics: reasoning on uncertainty, and the insignificance of test null. *Annales Zoologici Fennica*, 46:138–157, 2009.

[3] John A. Nelder. Statistics for the millennium: From statistics to statistical science. *The Statistician*, 48:257–269, 1999.

Luku 7

Kahden populaation vertaaminen

Tässä luvussa tarkastelemme frekventistisen tilastotieteen keinoin tilanteita, joissa vertaillaan kahden populaation odotusarvoparametrien suuruuksia populaatiosta saatujen otosten perusteella. Rajoitumme tapaukseen, jossa populaatiot oletetaan normaalijakautuneiksi.

7.1 Kahden populaation vertailu, kun otosten välillä on yhteyttä

Usein koeasetelma tuottaa mittauspareja (y_{1i}, y_{2i}), jonka komponentit ovat keskenään samankaltaisia. Esimerkiksi, jos mittaukset y_{1i} ja y_{2i} tehdään kaikilla i samasta otosyksiköstä (esim. samasta henkilöstä) ennen ja jälkeen käsittelyn, niin silloin vastaavia satunnaismuuttujia Y_{1i} ja Y_{2i} ei voida pitää riippumattomina, vaan samaan otosyksikköön i liittyvät muuttujat Y_{1i} ja Y_{2i} ovat samankaltaisempia kuin eri yksikköihin i ja j liittyvät muuttujat Y_{1i} ja Y_{2j}. Samanlainen tilanne syntyy, jos y_{1i} ja y_{2i} saadaan eri otosyksiköistä, jotka on kuitenkin valittu sillä tavalla, että ne ovat jonkin attribuutin mukaan samankaltaisia. Tarkastelemme nyt tällaisten toisistaan riippuvien otosten eli *parittaisten* (engl. *paired, related, matched*) otosten analysiointia.

Tällainen tilanne voidaan käsitellä tarkastelemalla erotuksia

$$d_i = y_{1i} - y_{2i}, \qquad i = 1, \ldots, n$$

Mikäli vastaavia satunnaismuuttujia

$$D_i = Y_{1i} - Y_{2i}, \qquad i = 1, \ldots, n$$

voidaan pitää riippumattomina, samoinjakautuneina ja (ainakin likimäärin) normaalijakautuneina,

$$D_i \sim N(\delta, \sigma^2), \qquad i = 1, \ldots, n,$$

niin tällöin populaatioiden odotusarvojen erotus on

$$\delta = \mu_1 - \mu_2,$$

ja tyypillisesti σ^2 on tuntematon. Odotusarvoparametrien erotusta $\delta = \mu_1 - \mu_2$ voidaan nyt analysoida soveltamalla t-luottamusväliä tai t-testiä erotuksiin d_i.

Tässä ns. *parittaisessa t-luottamusvälissä* tai *parittaisessa t-testissä* ei tarvitse esim. olettaa, että populaatioilla Y_{1i} ja Y_{2i} olisi sama varianssi (kuten seuraavassa jaksossa tehdään), vaan jakaumaoletukset tehdään erotuksille D_i. Jakaumaoletus on voimassa esim. silloin, kun kaksikomponenttiset vektorit (Y_{1i}, Y_{2i}) ovat satunnaisotos jostakin kaksiulotteiseta normaalijakaumasta.

7.2 Kaksi riippumatonta otosta normaalijakautuneista populaatioista

Tarkastelemme tilannetta, joka mallinnetaan kahdella riippumattomalla satunnaisotoksella normaalijakaumista $N(\mu_1, \sigma_1^2)$ ja $N(\mu_2, \sigma_2^2)$. Populaatiosta 1 saadaan n_1 havaintoa y_{1i} ja populaatiosta 2 saadaan n_2 havaintoa y_{2i}. Tavoitteena on verrata populaatioiden odotusarvoja μ_1 ja μ_2, jotka ovat tuntemattomia parametreja. Kehitämme tätä varten sekä luottamusvälejä että testejä.

Tällä tavalla voitaisiin mallintaa koetilanne, jossa tehdään mittauksia populaatiosta 1, jonka yksilöihin kohdistetaan käsittely 1 sekä populaatiosta 2, jonka yksilöihin kohdistetaan käsittely 2, mikäli kaikki yksilöt ovat toisistaan erillisiä. Mikäli tämä on käytännössä mahdollista, populaatiot mielellään muodostetaan satunnaistamalla, eli jakamalla havaintoyksiköt satunnaisesti kahteen ryhmään.

Oletus kahdesta riippumattomasta satunnaisotoksesta tarkoittaa sitä, että oletamme havaintoja vastaavien satunnaismuuttujien Y_{ki} noudattavan jakaumia

$$Y_{11}, Y_{12}, \dots, Y_{1n_1} \sim N(\mu_1, \sigma_1^2) \tag{7.1}$$

$$Y_{21}, Y_{22}, \dots, Y_{2n_2} \sim N(\mu_2, \sigma_2^2), \tag{7.2}$$

ja että kaikki satunnaismuuttujat Y_{ki} ovat riippumattomia. Otoskoot n_1 ja n_2 voivat olla erisuuria.

Populaatioiden parametreja (μ_1, σ_1^2) ja (μ_2, σ_2^2) voidaan estimoida tuttuun tapaan otoskeskiarvolla ja otosvarianssilla siten, että populaation k parametrit estimoidaan populaatiosta k saadusta otoksesta. Osoitamme alaindeksillä, kummasta populaatiosta otossuureet on laskettu. Käytämme merkintöjä

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}, \qquad \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2, \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2$$

Merkitsemme näitä estimaatteja vastaavia satunnaismuuttujia (ts. estimaattoreita) estimaatteja vastaavilla suurilla kirjaimilla

$$\bar{Y}_1, \quad \bar{Y}_2, \quad S_1^2 \quad \text{ja} \quad S_2^2.$$

Tiedämme jakson 4.4.2 kaavojen (4.14)–(4.16) perusteella, että

$$\bar{Y}_1 \sim N(\mu_1, \frac{1}{n_1}\sigma_1^2) \qquad \bar{Y}_2 \sim N(\mu_2, \frac{1}{n_2}\sigma_2^2) \tag{7.3}$$

$$\frac{n_1 - 1}{\sigma_1^2} S_1^2 \sim \chi^2_{n_1 - 1} \qquad \frac{n_2 - 1}{\sigma_2^2} S_2^2 \sim \chi^2_{n_2 - 1}, \tag{7.4}$$

ja että lisäksi toisaalta $\bar{Y}_1$ ja S_1^2 ovat riippumattomia ja että $\bar{Y}_2$ ja S_2^2 ovat riippumattomia satunnaismuuttujia. Nyt itseasiassa kaikki neljä satunnaismuuttujaa ovat riippumattomia sillä perusteella, että riippumattomista otoksista johdettavat estimaattorit ovat keskenään riippumattomia.

Kinnostuksen kohteena on populaatioiden odotusarvojen erotus

$$\delta = \mu_1 - \mu_2,$$

ja sitä estimoidaan vastaavalla otoskeskiarvojen erotuksella,

$$\hat{\delta} = \bar{y}_1 - \bar{y}_2. \tag{7.5}$$

Vastaavalla estimaattorilla on normaalijakauma sen takia, että riippumattomien normaalijakaumaa noudattavien satunnaismuuttujien lineaarikombinaatio tunnetusti noudattaa aina normaalijakaumaa. Laskemalla erotuksen odotusarvo ja varianssi saadaan johdettua kyseisen normaalijakauman parametrit, ja tällä tavalla nähdään, että

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{1}{n_1}\,\sigma_1^2 + \frac{1}{n_2}\,\sigma_2^2) \tag{7.6}$$

Tässä vaiheessa joudutaan erilaisiin tarkasteluihin sen mukaan, mitä populaatioiden variansseista oletetaan.

Varianssit tunnettuja

Jos molemmat varianssiparametrit σ_1^2 ja σ_2^2 ovat tunnettuja vakioita, niin satunnaismuuttuja

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{\sqrt{\frac{1}{n_1}\,\sigma_1^2 + \frac{1}{n_2}\,\sigma_2^2}}$$

noudattaa standardinormaalijakaumaa $N(0, 1)$. Tällä tavalla saadaan tuttuun tapaan johdettua luottamusväli odotusarvojen erotukselle $\delta = \mu_1 - \mu_2$ tai voidaan johtaa testit yksisuuntaiselle hypoteeseille

$$H_0 : \delta \le \delta_0, \quad H_1 : \delta > \delta_0$$

tai yksisuuntaiselle hypoteesille

$$H_0 : \delta \ge \delta_0, \quad H_1 : \delta < \delta_0$$

tai kaksisuuntaiselle hypoteesille

$$H_0 : \delta = \delta_0, \qquad H_1 : \delta \ne \delta_0.$$

Huomaa, että tässä tarkka hypoteesi $H_0 : \delta = \delta_0$ on oikeasti yhdistetty hypoteesi, koska se vastaa parametriavaruuden osajoukkoa

$$\Theta_0 = \{(\mu_1, \mu_2) : \mu_1 - \mu_2 = \delta_0\}.$$

Varianssit yhtäsuuria, mutta tuntemattomia

Edellistä käyttökelpoisempi tilanne on se, jossa populaatioiden varianssit ovat tuntemattomia, mutta ne oletetaan yhtäsuuriksi. Ts. oletetaan, että

$$\sigma_1^2 = \sigma_2^2 = \sigma^2,$$

jossa σ^2 on tuntematon parametri. Tällöin mallin parametrivektori on

$$(\mu_1, \mu_2, \sigma^2).$$

Kiinnostuksen kohteena on odotusarvojen erotus $\delta = \mu_1 - \mu_2$. Tämä tilanne on erikoistapaus ns. *varianssianalyysistä* (engl. *analysis of variance, ANOVA*), johon voi tutustua tarkemmin lineaaristen mallien kursseilla tai oppikirjoista.

Avainajatus analyysissä on muodostaa yhteiselle varianssille σ^2 sellainen estimaattori S_p^2, jonka jakauman hallitsemme, ja joka käyttää hyväksi molempien otoksien sisältämän informaation varianssista σ^2. Tämän jälkeen osaamme laskea parametrin δ estimaatin keskivirheen, ja loppu on tuttujen ideoiden soveltamista.

Käytämme hyväksi χ^2-jakauman ominaisuuksia. Jos $X \sim \chi^2_\nu$, niin sen odotusarvo on

$$EX = \nu, \tag{7.7}$$

mikä voidaan päätellä esim. gammajakauman odotusarvon kaavan avulla, sillä χ^2_ν jakauma on gammajakauma Gamma$(\frac{1}{2}\nu, \frac{1}{2})$. Tarvitsemme myös χ^2-jakauman yhteenlaskuominaisuutta. Jos

$$X_1 \sim \chi^2_{\nu_1}, \quad X_2 \sim \chi^2_{\nu_2}, \quad X_1 \perp\!\!\!\perp X_2,$$

niin tällöin

$$X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2} \tag{7.8}$$

Tämä voidaan johtaa esim. gammajakauman yhteenlaskuominaisuudesta.

Tietojen (7.3) sekä khiin neliön jakauman yhteenlaskuominaisuuden nojalla

$$\frac{n_1 - 1}{\sigma^2} S_1^2 + \frac{n_2 - 1}{\sigma^2} S_2^2 \sim \chi^2_{n_1+n_2-2},$$

ts. on voimassa

$$\frac{n_1 + n_2 - 2}{\sigma^2} S_P^2 \sim \chi^2_{n_1+n_2-2}, \tag{7.9}$$

kun määrittelemme *yhdistetyn* varianssiestimaattorin S_p^2 (engl. *pooled variance estimator*) seuraavana estimaattorien S_1^2 ja S_2^2 lineaarikombinaationa,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{7.10}$$

$$= \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2 \right] \tag{7.11}$$

Yhdistetty varianssiestimaattori on harhaton, sillä jakaumatulosta (7.9) sekä χ^2-jakauman odotusarvon kaavaa käyttämällä

$$ES_p^2 = E\left[\frac{\sigma^2}{n_1 + n_2 - 2} \frac{n_1 + n_2 - 2}{\sigma^2} S_p^2 \right] = \frac{\sigma^2}{n_1 + n_2 - 2} (n_1 + n_2 - 2) = \sigma^2.$$

Kun otetaan tuloksen (7.9) lisäksi huomioon se seikka, että

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2\right)$$

niin nähdään, että seuraavalla satunnaismuuttujalla on t-jakauma vapausasteluvulla $n_1 + n_2 - 2$,

$$t(\mathbf{Y}) = \frac{(\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)) \,/\, \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,\sigma\right)}{S_p/\sigma} \sim t_{n_1+n_2-2}.$$

Sieventämällä nähdään, että

$$t(\mathbf{Y}) = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \tag{7.12}$$

joten $t(\mathbf{Y})$ on saranasuure kiinnostavalle parametrille $\delta = \mu_1 - \mu_2$. Luottamusvälit ja testit voidaan perustaa tälle tulokselle.

Luottamustason $0 < 1 - \alpha < 1$ kaksisuuntainen luottamusväli saadaan johdettua tuttuun tapaan lähtemällä liikkeelle tuloksesta

$$P_{(\mu_1,\mu_2,\sigma^2)}\left(\left|\frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| \le t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Ratkaisemalla tämä epäyhtälö tuntemattoman δ suhteen nähdään, että jokaisessa parametriavaruuden pisteessä pätee todennäköisydellä $1 - \alpha$ paikkansa kaksoisepäyhtälö

$$\begin{aligned}\bar{Y}_1 - \bar{Y}_2 - t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \delta \le \\ \bar{Y}_1 - \bar{Y}_2 + t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) S_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\end{aligned}$$

Ts. luottamustason $1 - \alpha$ kaksisuuntainen luottamusväli saadaan laskemalla aineistosta parametrin $\delta = \mu_1 - \mu_2$ estimaatti

$$\hat{\delta} = \bar{y}_1 - \bar{y}_2 \tag{7.13}$$

sekä yhdistetty varianssiestimaatti

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \tag{7.14}$$

minkä jälkeen luottamusväli on muotoa

estimaatti $\pm$ (t-jakauman kriittinen piste) $\times$ estimaatin keskivirhe

eli tarkemmin sanoen se on

$$\left[\hat{\delta} - t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \hat{\delta} + t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{7.15}$$

Merkitsevyystason $0 < \alpha < 1$ kaksisuuntainen testi tarkalle hypoteesille

$$H_0 : \delta = \delta_0, \qquad H_1 : \delta \neq \delta_0$$

saadaan suoritettua laskemalla aineistosta testisuure

$$t = t(\mathbf{y}) = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{7.16}$$

jota verrataan t-jakaumaan vapausasteluvulla $n_1 + n_2 - 2$. Testi hylkää nollahypoteesin täsmälleen silloin, kun

$$|t| > t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right).$$

Yksisuuntainen testi hypoteeseille

$$H_0 : \delta \leq \delta_0, \qquad H_1 : \delta > \delta_0$$

hylkää nollahypoteesin, jos

$$t > t_{n_1+n_2-2}(\alpha),$$

ja hypoteeseille

$$H_0 : \delta \geq \delta_0, \qquad H_1 : \delta < \delta_0$$

hylkää nollahypoteesin, jos

$$t < -t_{n_1+n_2-2}(\alpha),$$

Tavallisesti näissä testeissä $\delta_0 = 0$. Usein testataan tarkkaa nollahypoteesia $\mu_1 - \mu_2 = 0$, jonka mukaan populaatiolla on sama odotusarvo. Huomaa, että tämä tarkka hypoteesi on yhdistetty, sillä se vastaa parametriavaruuden osajoukkoa

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma^2) : \mu_1 = \mu_2, \sigma^2 > 0\}$$

Varianssit erisuuria ja tuntemattomia

Jos varianssit ovat tuntemattomia, ja ne eivät ole yhtäsuuria, niin ratkaistavana on ns. Behrensin–Fisherin ongelma, jolle ei löydy tarkkaa ratkaisua. Sen sijaan on löydetty likimääräisiä ratkaisuja.

Esim. R:n funktio `t.test` käyttää tässä tilanteessa ns. Welchin testiä, joka taas perustuu ns. Satterthwaiten approksimaatioon. R käyttää kahden populaation vertailuun Welchin testiä, ellei sitä pyydetä erikseen olettamaan, että varianssit ovat yhtäsuuret.

Luku 8

Yhteensopivuuden ja riippumattomuuden testaaminen

Tässä kappaleessa tarkastelemme eräitä kuuluisia frekventistisen tilastotieteen testejä, joilla voidaan tutkia diskreettien havaintojen sopivuutta erilaisten tilastollisten mallien kanssa. Kaikki tämän kappaleen testit ovat likimääräisiä, ja niiden käyttö vaatii suurta otoskokoa.

8.1 Pearsonin testisuure

Tilastollinen malli koostuu diskreeteistä satunnaismuuttujia $Y_1, \ldots, Y_n$, joista kukin voi saada minkä tahansa arvoista $1, 2, \ldots, k$. Oletamme, että satunnaismuuttujat Y_h ovat riippumattomia ja samoin jakautuneita, jolloin niiden yhteisjakauma tiedetään, mikäli tunnetaan eri vaihtoehtojen $1, \ldots, k$ eli eri luokkien todennäköisyydet

$$p_i = P(Y_h = i), \qquad i = 1, \ldots, k \tag{8.1}$$

Luokkien todennäköisyyksien summa on yksi, joten mallissa on $k - 1$ vapaata parametria, joiksi voidaan valita $k - 1$ ensimmäisen luokan todennäköisyydet $p_1, \ldots, p_{k-1}$. Tämän jälkeen p_k voidaan laskea muiden p_i funktiona kaavalla

$$p_k = 1 - p_1 - p_2 - \cdots - p_{k-1}. \tag{8.2}$$

Binomikoe on tämän mallin erikoistapaus, jossa vaihtoehtoja on vain kaksi, ja joista yhtä pidetään onnistumisena ja toista epäonnistumisena. Tämän jakson mallia voidaan kutsua multinomikokeeksi.

Olkoon n_i luokan i havaittu frekvenssi, eli n_i on niiden indeksien h lukumäärä, joille $y_h = i$. Merkitsemme vastaavia satunnaismuuttujia symboleilla N_i. Binomijakauman määritelmän perusteella

$$N_i \sim \mathrm{Bin}(n, p_i), \qquad i = 1, \ldots, k,$$

joten $EN_i = np_i$.

Jos i ja j ovat eri luokkia, niin N_i ja N_j ovat riippuvia satunnaismuuttujia. Esim. jos binomikokeessa onnistumista kutsutaan luokaksi yksi ja epäonnistumista luokaksi kaksi, niin onnistumisten lukumäärän N_1 jakauma ja epäonnistumisten lukumäärän N_2 jakauma on

$$N_1 \sim \mathrm{Bin}(n, p), \qquad N_2 \sim \mathrm{Bin}(n, 1-p),$$

mutta koska $N_2 = n - N_1$, niin satunnaismuuttujan N_2 arvo tiedetään täysin, jos satunnaismuuttujan N_1 arvo tiedetään. Jos luokkia on enemmän kuin kaksi, niin frekvenssien keskinäinen riippuvuus ei ole enää yhtä äärimmäistä, mutta riippuvuus säilyy kuitenkin myös tässä tapauksessa. Sen sijaan alla olevat yksittäisten kokeiden lopputulokset Y_h toki ovat riippumattomia. Kun multinomikokeen tuloksia analysoidaan, niin analyysissä pitää ottaa huomioon frekvenssien keskinäinen riippuvuus.

Pearsonin testisuure vertaa luokkien havaittuja frekvenssejä n_i niiden odotettuihin frekvensseihin np_i seuraavalla tavalla:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i}. \tag{8.3}$$

Tavanomaisempi tapa mitata vektorin $(n_1, \ldots, n_k)$ ja vektorin $(np_1, \ldots, np_k)$ välistä etäisyyttä olisi esim. laskea vektoreiden euklidinen etäisyys, jolloin ensin summataan komponenttien neliöidyt erotukset $(n_i - np_i)^2$, ja lopuksi tuloksesta lasketaan neliöjuuri. Tällainen vertailu ei tässä tapauksessa ole hyvä ajatus, sillä todennäköisemmissä luokissa on odotettavissa enemmän satunnaisvaihtelua kuin harvinaisemmissa luokissa. Tätä varten Personin testisuureessa erotuksen neliö jaetaan luokan odotetulla frekvenssillä.

Pearsonin testisuure voidaan ilmaista myös kaavalla

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{8.4}$$

jossa $O_i = n_i$ on havaittu frekvenssi (engl. *observed frequency*), ja $E_i = np_i$ on odotettu frekvenssi (engl. *expected frequency*).

Karl Pearson esitti v. 1900 perustelun sille, miksi hänen nimeään kantavalla testisuureella on suuressa otoksessa likimain χ^2-jakauma. Pearsonin esittämä testi on ensimmäinen tunnettu tilastollinen testi, ja se avasi uuden aikakauden tilastollisen päättelyn historiassa.

Yksinkertaisimmassa yhteensopivuustestissä (engl. *goodness of fit test*) testataan yksinkertaista nollahypoteesia

$$H_0 : p_1 = \pi_1, \ldots, p_{k-1} = \pi_{k-1}, p_k = \pi_k, \tag{8.5}$$

jossa luvut (π_i) ovat jonkin teorian mukaisia luokkien todennäköisyyksiä, jotka oletetaan tunnetuiksi. Aineistosta lasketaan havaitut frekvenssit $O_i = n_i$. Nollahypoteesin vallitessa odotetut frekvenssit ovat

$$E_i = n\,\pi_i, \qquad i = 1, \ldots, k.$$

Suuret Pearsonin testisuureen X^2 arvot ovat nollahypoteesille kriittisiä. Nollahypoteesi hylätään merkitsevyystason $0 < \alpha < 1$ testissä, mikäli lasketun testisuureen arvo on suurempi kuin χ^2_{k-1} jakauman α-yläkvantiili. Huomaa, että

χ^2_ν-jakauman vapausasteluku ν on tässä täysin määrätyn nollahypoteesin tapauksessa

$$\nu = k - 1, \tag{8.6}$$

eli se on yhtä kuin mallin vapaiden parametrien lukumäärä $k-1$. Testin p-arvo on

$$1 - F(t),$$

jossa F on χ^2_{k-1}-jakauman kertymäfunktio, ja t on testisuureen laskettu arvo.

Esimerkki 8.1 (Nopan harhattomuuden testaaminen) Simuloidaan ensin $n = 2000$ nopanheittoa harhaisesta nopasta, jolle silmälukujen todennäköisyydet ovat

$$P(Y=1) = P(Y=2) = P(Y=3) = P(Y=4) = P(Y=5) = \frac{4}{25},$$
$$P(Y=6) = \frac{5}{25}.$$

R:llä nopanheittojen simulointi ja havaittujen frekvenssien laskeminen sujuvat seuraavasti.

```
n <- 2000
p <- c(4, 4, 4, 4, 4, 5)
p <- p/sum(p)
y <- sample(1:6, size = n, replace = TRUE, prob = p)
obs <- table(y)
```

Testaan nyt eräässä tällaisessa simuloinnissa saatuja frekvenssejä, kun (epätoden) nollahypoteesin mukaan kaikilla silmäluvuilla on sama todennäköisyys 1/6. Havaitut frekvenssit, odotetut frekvenssit sekä suureet $(O_i - E_i)^2/E_i$ ovat

silmäluku	1	2	3	4	5	6	summa
havaittu O_i	311	318	306	342	316	407	2000
odotettu E_i	333.3	333.3	333.3	333.3	333.3	333.3	2000
$(O_i - E_i)^2/E_i$	1.496	0.705	2.241	0.225	0.901	16.280	21.85

Allaolevassa koodissa lasketaan Pearsonin X^2-testisuureen arvo, χ^2_{k-1}-jakauman kriittinen arvo sekä testisuureita vastaavat p-arvot.

```
obs <- c(`1` = 311, `2` = 318, `3` = 306, `4` = 342, `5` = 316,
    `6` = 407)
n <- sum(obs)
alpha <- 0.05
p0 <- c(1, 1, 1, 1, 1, 1)/6
expected <- n * p0
print(x2 <- sum((obs - expected)^2/expected))

## [1] 21.85

nu <- length(p0) - 1
print(crit <- qchisq(alpha, df = nu, lower = FALSE))

## [1] 11.07

print(p.value.x2 <- pchisq(x2, df = nu, lower = FALSE))

## [1] 0.0005591
```

Nollahypoteesi hylätään merkitsevyystasolla $\alpha = 0.05$. Sama testi saadaan suoritettua myös soveltamalla R:n funktiota `chisq.test`.

```
chisq.test(obs)

##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 21.85, df = 5, p-value = 0.0005591
```

△

Esimerkki 8.2 Tarkistetaan seuraavaksi, miten käy kun Pearsonin testisuure lasketaan binomikokeessa, jossa onnistumisten lukumäärä K noudattaa binomijakaumaa $\mathrm{Bin}(n, p)$. Tällöin saadaan taulukko

luokka	onnistuminen	epäonnistuminen	yhteensä
havaittu	K	$n - K$	n
odotettu	np	$n(1-p)$	n

Tällöin

$$\begin{aligned} X^2 &= \frac{(K-np)^2}{np} + \frac{[(n-K) - n(1-p)]^2}{n(1-p)} \\ &= \frac{(K-np)^2}{np(1-p)}. \end{aligned}$$

Pearsonin testisuureen teorian mukaan tämän satunnaismuuttujan pitäisi likimain noudattaa jakaumaa χ^2_1. Toisaalta binomijakauman normaalijakauma-approksimaation mukaan satunnaismuuttuja

$$\frac{K - np}{\sqrt{np(1-p)}}$$

noudattaa suurella otoskoolla n osapuilleen standardinormaalijakaumaa $N(0, 1)$, ja Pearsonin testisuure X^2 on yhtä kuin tämän satunnaismuuttujan neliö. Todennäköisyyslaskennasta tiedetään, että mikäli $Z \sim N(0, 1)$, niin

$$Z^2 \sim \chi^2_1.$$

Nämä tulokset antavat tuntumaa siihen, milloin khiin neliön jakauma-approksimaatio toimii hyvin ja milloin taas huonosti. △

Pearsonin X^2-testisuureeseen perustuva testi on likimääräinen, sillä se perustuu likimääräiseen suuren otoskoon jakaumatulokseen. Milloin tämä approksimaatio on tarpeeksi hyvä? Kirjallisuudessa löytyy tähän tilanteeseen erilaisia suosituksia. Testiä sovelletaan huolta vailla esim. silloin, jos kaikille luokille niiden odotetut frekvenssit ovat vähintään viisi. Jos joidenkin luokkien odotetut frekvenssit ovat liian pieniä, niin sitten kyseisiä luokkia voidaan yhdistää keskenään ennen testin soveltamista.

Yhteensopivuustestien voima kasvaa otoskoon kasvaessa. Jos nollahypoteesi ei pidä tarkasti paikkaansa, niin kyllin suurella otoskoolla se jossakin vaiheessa hylätään, vaikka poikkeama olisi niin pieni, että sillä ei ole käytännön kannalta merkitystä.

Usein yhteensopivuustestissä nollahypoteesi on yhdistetty, ja se voidaan esittää kaavalla

$$H_0 : p_1 = p_1(\theta), \dots, p_k = p_k(\theta), \qquad \text{jollekin } \theta \in \Theta_0. \tag{8.7}$$

Tällöin odotettuja frekvenssejä ei saada laskettua ennen kuin parametrin θ arvo on estimoitu. Mikäli θ estimoidaan havaituista frekvensseistä suurimman uskottavuuden menetelmällä, ja estimaatti on $\hat{\theta}$, niin tällöin testisuureena voidaan käyttää tuttua Pearsonin testisuuretta,

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Uusi asia on se, että odotetut frekvenssit pitää laskea käyttämällä parametrin tilalla sen SU-estimaattia

$$E_i = n\, p_i(\hat{\theta}), \qquad i = 1, \dots, k. \tag{8.8}$$

Jos parametrilla θ on d vapaata komponenttia (ts. komponenttia, joita mikään sidosehto ei kytke toisiinsa) niin tällöin testisuureella (satunnaismuuttujaksi ymmärrettynä) on suurella otoskoolla osapuilleen χ^2-jakauma vapausasteluvulla

$$\nu = k - 1 - d. \tag{8.9}$$

Tämän asian perusteli ensimmäisenä Fisher 1920-luvulla. Hän osoitti samalla, että K. Pearson oli tehnyt omissa laskelmissaan virheen vapausasteluvun kohdalla.

Ennenkuin SU-menetelmää voidaan käyttää, pitää tietenkin pystyä kirjoittamaan uskottavuusfunktio. Yhden havaintosatunnaismuuttujan Y_h pistetodennäköisyysfunktio voidaan ilmaista kaavalla

$$\begin{aligned} P(Y_h = y_h) &= \begin{cases} p_1 & \text{jos } y_h = 1, \\ p_2 & \text{jos } y_h = 2, \\ \vdots & \\ p_k & \text{jos } y_h = k. \end{cases} \\ &= p_1^{1(y_h=1)}\, p_2^{1(y_h=2)} \cdots p_k^{1(y_h=k)} \end{aligned}$$

Tässä $1(y_h = i)$ on osoitinmuuttuja sille, että y_h:n arvo on i, eli

$$1(y_h = i) = \begin{cases} 1 & \text{mikäli } y_h = i, \\ 0 & \text{muuten.} \end{cases}$$

Havaintoja $y_1, \dots, y_n$ vastaava uskottavuusfunktio on

$$L(p_1, \dots, p_{k-1}) = \prod_{h=1}^{n} p_1^{1(y_h=1)}\, p_2^{1(y_h=2)} \cdots p_k^{1(y_h=k)}$$

Kun luokkien todennäköisyyksien p_i potenssit yhdistetään, uskottavuusfunktiolle saadaan lauseke

$$L(p_1, \dots, p_{k-1}) = p_1^{n_1}\, p_2^{n_2} \, \dots \, p_k^{n_k}, \tag{8.10}$$

jossa n_i on niiden indeksien h lukumäärä, joille $y_h = i$, eli n_i on luokan i havaittu frekvenssi.

Esimerkki 8.3 Kalbfleish [2, Esimerkki 11.2.2] esittää aineiston, jossa on etsitty hiukkaslaskurilla alfahiukkasia. Kokeessa on toistettu 2608 kertaa tietyn pituinen mittaus, jossa on laskettu laskuriin osuvien alfahiukkasten lukumäärä. Sitten on taulukoitu niiden mittausten lukumäärä, joissa laskuri on havainnut i kappaletta alfahiukkasia, kun $i = 0, 1, \dots, 9$, mutta kaikki tapaukset joissa on havaittu $i \geq 10$ alfahiukkasta on yhdistetty yhdeksi luokaksi. Havaitut frekvenssit ovat seuraavassa taulukossa.

luokka	0	1	2	3	4	5	6	7	8	9	≥ 10	yht.
havaittu	57	203	383	525	532	408	273	139	45	27	16	2608

Kokeessa tutkitaan, noudattaako alfahiukkasten lukumäärä Poissonin jakaumaa. Poissonin jakauma on diskreetti todennäköisyysjakauma, jonka pistetodennäköisyysfunktio on

$$g(x;\theta) = \mathrm{e}^{-\theta}\,\frac{\theta^x}{x!}, \qquad x = 0, 1, 2, \dots,$$

jossa parametri $\theta > 0$ on jakauman odotusarvo. Tällöin luokat $0, \dots, 9$ vastaavat havaittujen alfahiukkasten lukumääriä, ja niiden todennäköisyydet ovat

$$p_i = g(i;\theta), \qquad i = 0, 1, \dots, 9.$$

Viimeinen luokka vastaa sitä tapahtumaa, että laskuri havaitsee vähintään 10 alfahiukkasta, ja sen todennäköisyys on

$$p_{10} = 1 - \sum_{i=0}^{9} g(i;\theta).$$

Luokkien todennäköisyyksiä ei saada laskettua ennen kuin Poissonin jakauman parametri on estimoitu. SU-estimaatiksi saadaan (tietokoneen avulla) $\hat{\theta} = 3.8703$. Kun tämä sijoitetaan Poissonin jakauman pistetodennäköisyysfunktioon, saadaan Pearsonin testisuure X^2 laskettua.

luokka	havaittu	odotettu	$(O_i - E_i)^2/E_i$
0	57	54.38	0.13
1	203	210.47	0.27
2	383	407.30	1.45
3	525	525.46	0.00
4	532	508.42	1.09
5	408	393.55	0.53
6	273	253.86	1.44
7	139	140.36	0.01
8	45	67.90	7.73
9	27	29.20	0.17
≥ 10	16	17.08	0.07
summa	2608	2608	12.88

Arvoa $X^2 = 12.88$ verrataan χ^2_ν-jakaumaan, kun vapausasteluku on

$$\nu = k - 1 - d = 11 - 1 - 1 = 9.$$

Taulukko 8.1 Kontingenssitaulukko, jolla on r riviä ja c saraketta.

$$\begin{array}{cccc|c}
n_{11} & n_{12} & \cdots & n_{1c} & n_{1\bullet} \\
n_{21} & n_{22} & \cdots & n_{2c} & n_{2\bullet} \\
\vdots & \vdots & & \vdots & \vdots \\
n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r\bullet} \\
\hline
n_{\bullet 1} & n_{\bullet 2} & \cdots & n_{\bullet c} & n
\end{array}$$

Testin p-arvoksi saadaan $p = 0.17$. Tämä ei ole erityisen pieni: jos Poisson-malli pitää paikkaansa, niin 17 %:n todennäköisyydellä saadaan testisuureelle arvoja joiden mukaan havaitut ja odotetut frekvenssit poikkeavat ainakin näin paljon toisistaan. Nollahypoteesi jää voimaan, jos käytetään esim. 5 %:n merkitsevyystasoa. △

8.2 Riippumattomuuden testaaminen kontingenssitaulukossa

Tarkastellaan n otosyksikköä, joista kustakin mitataan kaksi diskreettiä ominaisuutta (x_h, y_h), jossa $h = 1, \dots, n$. Ominaisuus x_h saa yhden arvoista $1, \dots, r$ ja ominaisuus y_h yhden arvoista $1, \dots, c$.

Tilastollinen malli koostuu n riippumattomasta ja samoin jakautuneesta satunnaismuuttujaparista (X_h, Y_h), joille

$$p_{ij} = P(X_h = i, Y_h = j), \qquad i = 1, \dots, r, \quad j = 1, \dots, c.$$

Tilanne on muuten samanlainen kuin edellisessä jaksossa, mutta nyt luokkia indeksoidaan kahdella indeksillä i ja j eikä enää yhdellä indeksillä. Luokkia on rc, joten vapaita parametreja p_{ij} on $rc - 1$, mikäli niitä ei rajoiteta lisäämällä malliin oletuksia.

Havaitut frekvenssit ovat

$$n_{ij}, \qquad i = 1, \dots, r, \quad j = 1, \dots, c.$$

ja ne voidaan esittää taulukkona, jossa i indeksoi vaakarivejä ja j sarakkeita. Tälläistä taulukkoa kutsutaan *kontingenssitaulukoksi* (engl. *contingency table*). Taulukon rivin i summaa merkitään $n_{i\bullet}$ ja sarakkeen j summaa $n_{\bullet j}$, ts. alaindeksi piste tarkoittaa summaamista kyseisen indeksin yli, eli

$$n_{i\bullet} = \sum_{j=1}^{c} n_{ij}, \qquad i = 1, \dots, r$$
$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}, \qquad j = 1, \dots, c.$$

Testattavan nollahypoteesin mukaan satunnaismuuttujat X_h ja Y_h ovat riippumattomat, mikä tarkoittaa sitä, että kaikilla i ja j niiden yhteisjakauman pistetodennäköisyysfunktio saadaan kertomalla keskenään vastaavat reunajakaumien pistetodennäköisyydet, eli

$$p_{ij} = P(X_h = i, Y_h = j) = P(X_h = i)\, P(Y_h = j) = \gamma_i\, \delta_j, \quad \text{missä}$$
$$\gamma_i = P(X_h = i), \qquad \delta_j = P(Y_h = j).$$

Reunajakaumien pistetodennäköisyydet $\gamma_1, \dots, \gamma_r$ ja $\delta_1, \dots, \delta_c$ ovat tuntemattomia parametreja. Nollahypoteesi voidaan ilmaista myös sanomalla, että rivi- ja sarakeluokittelut ovat riippumattomia.

Jaksossa 8.5 osoitetaan, että nollahypoteesin vallitessa suurimman uskottavuuden estimaatit ovat

$$\hat{\gamma}_i = \frac{n_{i\bullet}}{n}, \qquad i = 1, \dots, r \tag{8.11}$$

$$\hat{\delta}_j = \frac{n_{\bullet j}}{n}, \qquad j = 1, \dots, c. \tag{8.12}$$

Jälleen kerran todennäköisyysparametrien suurimman uskottavuuden estimaatit ovat vastaavat suhteelliset frekvenssit.

Kun nollahypoteesi pitää paikkansa, niin tuntemattomia vapaita parametreja on

$$d = r - 1 + c - 1$$

kappaletta, sillä molemmat reunatodennäköisyysfunktiot summautuvat ykköseksi. Khiin neliön testissä vapausasteiden lukumääräksi saadaan kaavan (8.9) mukaan

$$\nu = rc - 1 - (r - 1 + c - 1) = (r - 1)(c - 1).$$

Havaitut frekvenssit ovat $O_{ij} = n_{ij}$, ja nollahypoteesin vallitessa odotetut frekvenssit ovat

$$E_{ij} = n\, \hat{\gamma}_i\, \hat{\delta}_j = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$$

Pearsonin χ^2-testisuure on

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Koon $0 < \alpha < 1$ testissä testisuuretta verrataan χ^2_ν jakauman α-yläkvantiiliin, jossa vapausasteluku

$$\nu = (r - 1)(c - 1).$$

Esimerkki 8.4 (ABO-veriryhmien geneettinen perusta [1, Jakso 4.5]) Ihmisiä luokitellaan eri veriryhmiin veren ominaisuuksien perusteella. Tunnetuin veriryhmäjärjestelmä on ns. ABO-veriryhmäjärjestelmä, jossa kunkin yksilön veriryhmä on joko A, B, AB tai O. Veriryhmä määritetään tarkistamalla, onko henkilön veressä antigeeniä A tai antigeeniä B. Veriryhmät nimetään tuloksen perusteella seuraavan taulukon mukaisesti

	Ei B	On B
Ei A	O	B
On A	A	AB

Vielä 1920-luvulla oli epäselvää, minkälaisesta geneettisestä mekanismista ABO-veriryhmät määräytyvät. Yksi mahdollinen selitys oli kahden riippumattoman lokuksen malli, jossa lokuksen yksi alleeli määrää, onko veressä antigeeniä A vai ei, ja lokuksen kaksi alleeli määrää, onko veressä antigeeniä B vai ei. Jos kahden riippumattoman lokuksen malli on tosi ja jos otamme populaatiosta satunnaisotoksen, niin tällöin veriryhmistä muodostetussa kontingenssitaulukossa rivi- ja sarakeluokittelut ovat riippumattomia. Englannissa 1930-luvulla tehdystä otoksesta saatiin seuraava veriryhmien kontingenssitaulukko

	Ei B	On B
Ei A	202	35
On A	179	6

Tästä taulukosta laskettuna Pearsonin χ^2 testisuureen arvo on $X^2 = 15.73$. Tätä verrataan χ^2_1-jakauman α-yläkvantiiliin. Esim. jos merkitsevyystaso on $\alpha = 0.05$, niin tämä kriittinen arvo on 3.84, joten kahden lokuksen malli tulee testissä hylättyä. R:llä testisuureiden arvot voidaan laskea seuraavasti.

```
print(bloodgroups <- matrix(c(202, 179, 35, 6), 2, 2))

##      [,1] [,2]
## [1,]  202   35
## [2,]  179    6

print(n <- sum(bloodgroups))

## [1] 422

print(gamma.hat <- rowSums(bloodgroups)/n)

## [1] 0.5616 0.4384

print(delta.hat <- colSums(bloodgroups)/n)

## [1] 0.90284 0.09716

observed <- bloodgroups
print(expected <- n * (gamma.hat %o% delta.hat))

##      [,1]  [,2]
## [1,]  214 23.03
## [2,]  167 17.97

print(x2 <- sum((observed - expected)^2/expected))

## [1] 15.73

nu <- (2 - 1) * (2 - 1)
print(crit <- qchisq(0.05, df = nu, lower = FALSE))

## [1] 3.841

print(p.x2 <- pchisq(x2, df = nu, lower = FALSE))

## [1] 7.298e-05
```

Testin voi suorittaa myös funktiolla `chisq.test`. Tämä funktio käyttää oletusarvoisesti ns. jatkuvuuskorjausta. Alla olevassa koodissa pyydetään erikseen olemaan käyttämättä jatkuvuuskorjausta, jotta tuloksia voitaisiin suoraan verrata itse tehtyjen laskujen kanssa.

```
chisq.test(bloodgroups, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  bloodgroups
## X-squared = 15.73, df = 1, p-value = 7.298e-05
```

Nykyään tiedetään, että ABO-veriryhmä määräytyy yhdestä lokuksesta, jolla voi olla kolme alleelia A, B tai O, joista A ja B ovat dominoivia ja O resessiivinen. Nollahypoteesin hylkääminen on oikea päätös. $\triangle$

8.3 Homogeenisuuden testaaminen

Kuten edellisessä jaksossa, nytkin oletetaan, että otosyksiköillä on kaksi diskreettiä ominaisuutta x ja y, ja x:n mahdolliset arvot ovat $1, \dots, r$ ja y:n mahdolliset arvot $1, \dots, c$. Populaatio jaetaan ominaisuuden x arvojen määräämiin ositteisiin siten, että ositteessa i ominaisuuden x arvo on i. Kustakin osasta tehdään riippumaton kokoa n_i oleva otos

$$Y_{ih}, \qquad h = 1, \dots, n_i.$$

Tavoitteena on testata, ovatko ositteiden jakaumat samat eli ovatko ositteet homogeenisia. Nollahypoteesi on

$$H_0 : p_{ij} = \pi_j, \qquad \text{kaikilla } i = 1, \dots, r \text{ ja } j = 1, \dots, c, \tag{8.13}$$

missä

$$p_{ij} = P(Y_{ih} = j),$$

ja todennäköisyydet $(\pi_1, \dots, \pi_c)$ ovat tuntemattomia.

Todennäköisyyksien (π_i) suurimman uskottavuuden estimaateiksi saadaan

$$\hat{\pi}_j = \frac{n_{\bullet j}}{n}, \qquad j = 1, \dots, c. \tag{8.14}$$

Testissä muodostetaan taulukko, jossa vaakariville i tulee ositteesta i lasketut frekvenssit, ja vaakarivin i frekvenssien summa $n_{i\bullet}$ on ositteesta i tehdyn otoksen koko n_i. Osoittautuu, että tämän jälkeen testaus voidaan tehdä aivan samalla tavalla kuin edellisessä jaksossa. Pearsonin testisuureetta verrataan χ^2_ν-jakaumaan, jossa vapausasteluku on

$$\nu = (r-1)(c-1).$$

8.4 Uskottavuusosamäärän testisuure

Tämän luvun tilanteissa käytetään usein myös muita testisuureita kuin Pearsonin testisuureetta. Merkitään nyt tilastollisen mallin uskottavuusfunktiota $L(\boldsymbol{\theta}; \mathbf{Y})$, sen logaritmista uskottavuusfunktiota $\ell(\boldsymbol{\theta}; \mathbf{Y})$ ja suurimman uskottavuuden estimaattoria symbolilla $\hat{\boldsymbol{\theta}}(\mathbf{Y})$. Suurimman uskottavuuden estimaattorien asymptoottisen teorian perusteella tiedetään, että jos havaintosatunnaisvektorilla $\mathbf{Y}$ on parametria $\boldsymbol{\theta}$ vastaava jakauma, niin tällöin *uskottavuusosamäärän testisuureella* (engl. *likelihood ratio statistic*)

$$W = 2[\ell(\hat{\boldsymbol{\theta}}(\mathbf{Y}); \mathbf{Y}) - \ell(\boldsymbol{\theta}; \mathbf{Y})] = 2 \log \frac{L(\hat{\boldsymbol{\theta}}(\mathbf{Y}); \mathbf{Y})}{L(\boldsymbol{\theta}; \mathbf{Y})}$$

on (tiettyjen oletusten vallitessa) χ^2-jakauma, jossa vapausasteluku on yhtä kuin mallin vapaiden parametrien lukumäärä.

Tarkastelemme ensin jakson 8.1 tilastollista mallia silloin, kun luokkien todennäköisyydet $p_1, \dots, p_{k-1}$ ovat vapaita parametreja. Uskottavuusosamäärän testisuureessa tarvittavat log-uskottavuusarvot ovat kaavan (8.10) mukaan

$$\ell(\hat{p}_1, \dots, \hat{p}_{k-1}; \mathbf{y}) = \sum_{i=1}^{k} n_i \log \hat{p}_i$$

$$\ell(p_1, \dots, p_{k-1}; \mathbf{y}) = \sum_{i=1}^{k} n_i \log p_i$$

Näytämme seuraavassa jaksossa, että SU-estimaatit ovat vastaavat suhteelliset frekvenssit

$$\hat{p}_1 = \frac{n_1}{n}, \dots, \hat{p}_{k-1} = \frac{n_{k-1}}{n}, \hat{p}_k = \frac{n_k}{n},$$

joten uskottavuusosamäärän testisuure on

$$W = 2 \sum_{i=1}^{k} n_i \log \frac{\hat{p}_i}{p_i} = 2 \sum_{i=1}^{k} n_i \log \frac{n_i}{n\, p_i} \tag{8.15}$$

$$= 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i}, \tag{8.16}$$

Jos jokin n_i tai O_i on nolla, niin tässä kaavassa pitää tulkita $0 \log 0 = 0$. Suurella otoskoolla testisuuretta vastaavalla satunnaismuuttujalla on osapuilleen χ^2_{k-1}-jakauma.

Myös muissa tämän kappaleen tilanteissa voidaan myös käyttää Pearsonin testisuureen sijasta uskottavuusosamäärän testisuuretta, mutta tällöin tarvitaan sisäkkäisten mallien vertailuun tarkoitettua uskottavuusosamäärän testisuureetta [1, kaava (4.38)] Se on kaikissa tämän kappaleen tilanteissa muotoa

$$W = 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i},$$

tai kaksiulotteisille taulukoille muotoa

$$W = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \log \frac{O_{ij}}{E_{ij}},$$

jossa odotetut frekvenssit lasketaan samoin kuin Pearsonin testisuureelle. Uskottavuusosamäärän testisuureetta verrataan täsmälleen samaan χ^2-jakaumaan kuin Pearsonin testisuuretta.

8.5 Suurimman uskottavuuden estimaatit

Tässä jaksossa johdamme tässä kappaleessa ilmoitetut suurimman uskottavuuden estimaattien kaavat. Kaavat olisi mahdollista johtaa monella menetelmällä. Voisimme eliminoida yhden todennäköisyysparametereista käyttämällä sitä tietoa, että ne summautuvat ykköseksi. Toinen mahdollisuus olisi käyttää Lagrangen keinoa rajoitteellisten optimointitehtävien ratkaisemiseksi. Tässä jaksossa tulokset kuitenkin johdetaan käyttämällä juuri tähän tilanteeseen sopivaa epäyhtälöä. Tällä konstilla johdoista tulee paljon yksinkertaisempia kuin yleisemmillä keinoilla.

Johdamme ensin parametreille $p_1, \ldots, p_k$ eli luokkien todennäköisyyksille suurimman uskottavuuden estimaatit jakson 8.1 tilastollisessa mallissa siinä tapauksessa, kun $k-1$ luokan todennäköisyydet $p_1, \ldots, p_{k-1}$ ovat vapaita parametreja. Otamme johdossa huomioon, että parametreilla on lineaarinen rajoite

$$p_1 + \cdots + p_k = 1$$

ja että havainnoilla n_i on rajoite

$$n_1 + \cdots + n_k = n.$$

Käytämme hyväksi aputulosta, joka sanoo, että luonnollisen logaritmin $\log(x)$ kuvaaja jää pisteeseen $x = 1$ piirretyn tangenttinsa alapuolelle.

$$\log(x) \le x - 1, \qquad \text{kaikilla } x > 0. \tag{8.17}$$

Yhtäsuuruus saavutetaan ainoastaan pisteessä $x = 1$. Tämän väitteen voi tarkistaa helposti analyysin keinoilla.

Estimaatien kaavat nähdään helposti seuraavan lauseen avulla.

Lause 8.1. *Jos k ei-negatiivista lukua $n_i \ge 0$ summautuvat luvuksi $n > 0$, eli*

$$\sum_{i=1}^{k} n_i = n,$$

niin tällöin mille tahansa luvuille $p_i \ge 0$ jotka summautuvat ykköseksi pätee

$$\sum_{i=1}^{k} n_i \log p_i \le \sum_{i=1}^{k} n_i \log \frac{n_i}{n},$$

missä käytämme tarvittaessa sopimusta $0 \log 0 = 0$.

Todistus. Esitän todistuksen vain siinä tapauksessa, jossa kaikki $n_i > 0$. Yleisen tapauksen saa todistettua helposti samaan tapaan. Väitetyn epäyhtälön vasemman ja oikean puolen erotus on

$$\begin{aligned} \sum \left(n_i \log p_i - n_i \log \frac{n_i}{n} \right) &= \sum n_i \log \frac{p_i}{n_i/n} \le \sum n_i \left(\frac{p_i}{n_i/n} - 1 \right) \\ &= n \sum p_i - \sum n_i = n - n = 0, \end{aligned}$$

missä sovelsimme epäyhtälöä (8.17). □

Suurimman uskottavuuden estimaatit ovat siis vastaavat suhteelliset frekvenssit, eli

$$\hat{p}_i = \frac{n_i}{n}, \qquad i = 1, \ldots, k. \tag{8.18}$$

Jos testataan riippumattomuutta kontingenssitaulukossa, niin logaritminen uskottavuusfunktio on nollahypoteesin $p_{ij} = \gamma_i \, \delta_j$ vallitessa

$$\begin{aligned}
\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, \log p_{ij} &= \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, \log(\gamma_i \, \delta_j) = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \, [\log \gamma_i + \log \delta_j] \\
&= \sum_{i=1}^{r} \log \gamma_i \sum_{j=1}^{r} n_{ij} + \sum_{j=1}^{c} \log \delta_j \sum_{i=1}^{r} n_{ij} \\
&= \sum_{i=1}^{r} n_{i\bullet} \log \gamma_i + \sum_{j=1}^{c} n_{\bullet j} \log \delta_j \\
&\leq \sum_{i=1}^{r} n_{i\bullet} \log \frac{n_{i\bullet}}{n} + \sum_{j=1}^{c} n_{\bullet j} \log \frac{n_{\bullet j}}{n},
\end{aligned}$$

mistä nähdään, että suurimman uskottavuuden estimaateilla on kaavat (8.11) ja (8.12).

Homogeenisuuden testauksessa uskottavuusfunktio on nollahypoteesin $p_{ij} = \pi_j$ vallitessa

$$L(\pi_1, \ldots, \pi_{c-1}) = \prod_{i=1}^{r} \pi_1^{n_{i1}} \, \pi_2^{n_{i2}} \ldots \pi_c^{n_{ic}} = \pi_1^{n_{\bullet 1}} \, \pi_2^{n_{\bullet 2}} \ldots \pi_c^{n_{\bullet c}},$$

joten suurimman uskottavuuden estimaattien kaavat (8.14) saadaan suoraan lauseesta 8.1, kun ensin siirrytään tarkastelemaan uskottavuusfunktion logaritmia.

Kirjallisuutta

[1] A. C. Davison. *Statistical Models.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2003.

[2] J. G. Kalbfleisch. *Probability and Statistical Inference II.* Springer, 1979.