Luku 5

Tilastollinen testaus

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

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

5.1 Testauksen peruskäsitteitä

Tarkastelemme testausta frekventistisessä 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 parametrisessa mallissa aina muotoilla siten, että väitetään 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 θ_0. 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 kiinnostavaa 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

$$H_0 : \theta \in \Theta_0, \tag{5.1}$$

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 H_A). Vastahypoteesin mukaan θ kuuluu parametriavaruuden osajoukkoon Θ_1, ts.

$$H_1 : \theta \in \Theta_1. \tag{5.2}$$

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 täysin epäsymmetrisellä tavalla. 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*). Testisuureella täytyy olla jokin monotonisuusominaisuus, käytännössä yksi seuraavista:

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.

Edellisen lisäksi testisuureella $t(\mathbf{y})$ täytyy olla se ominaisuus, että hallitsemme 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 5.1. Jos testin kriittinen alue on C, niin testin *koko* (engl. *size*) eli sen *merkitsevyystaso* (engl. *significance level*) on

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(t(\mathbf{Y}) \in C) = \sup_{\theta \in \Theta_0} P_\theta(H_0 \text{ hylätään}) \tag{5.3}$$

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

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

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

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

Usein $P_\theta(t(\mathbf{Y}) \in C)$ pysyy vakiona joukossa Θ_0 — näin käy automaattisesti, jos H_0 on yksinkertainen tai jos testisuure on saranasuure — jolloin supremumin otosta ei tarvitse huolehtia.

Tyypillisesti testin merkitsevyystaso $0 < \alpha < 1$ asetetaan, ja sitten tämän informaation perusteella määritetään kriittinen alue C siten, että vaatimus (5.3) 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 täysin 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 *hylkäämissvirhe* 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.

5.2 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. Sitä käsitellään sen vuoksi, että teoria on tässä tapauksessa helpointa ymmärtää.

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

$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}. \tag{5.4}$$

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

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

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

$$z > z_\alpha. \tag{5.5}$$

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 (5.5) selvitetään tämän jälkimmäisen yhdistetyn nollahypoteesin $\mu \leq \mu_0$ paikkansapitävyttä. 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

$$z < -z_\alpha \tag{5.6}$$

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

$$|z| > z_{\alpha/2}. \tag{5.7}$$

Tämä perustuu siihen, että

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

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

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 3.3 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 tehdään aina z-testejä kiinteällä 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 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.

5.3 Testin voima

Määritelmä 5.2 (Testin voima)**.** Jos C on tarkasteltavan testin kriittinen alue, niin parametriavaruudella määriteltyä funktio

$$\pi(\theta) = P_\theta(t(\mathbf{Y}) \in C) = P_\theta(H_0 \text{ hylätään}) \tag{5.8}$$

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

Toisin sanoen, voimafunktio on testin hylkäystodennäköisyys parametrin funktiona. (Se mittaa *hylkäys*voimaa.) Ensimmäisen ja toisen lajin virheiden todennäköisyyden avulla lausuttuna

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

Jos testin koko on α, niin testin koon määritelmän (5.3) ja sen voiman määritelmän (5.8) 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.

5.4 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 siten, että *pieni p-arvo viittaa ristiriitaan* aineiston ja nollahypoteesin välillä.

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

$$p = p(\mathbf{y}) = \sup_{\theta \in \Theta_0} P_\theta[t(\mathbf{Y}) \geq t(\mathbf{y})] \tag{5.9}$$

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 määritelmässä oleva todennäköisyys kuitenkin riippuu parametrinarvosta $\theta \in \Theta_0$, niin tällöin oikea sanallinen määritelmä on edellisen sijasta

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

Testi saadaan suoritettua myös laskemalla aineistosta p-arvo, ja toimimalla seuraavasti

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

Voidaan osoittaa, 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.

5.5 z-testin p-arvo ja voima

Laskemme seuraavaksi jaksossa 5.2 käsitellyn z-testin p-arvon ja voimafunktion sekä yksi- että kaksisuuntaisessa tapauksessa.

Yksisuuntainen z-testi

Yksisuuntaisen testin

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

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, eli

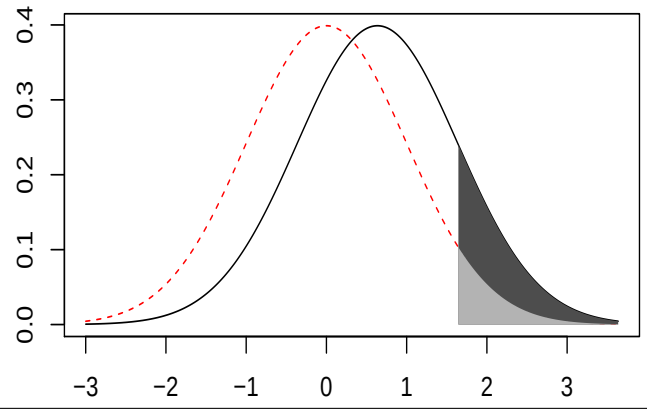
$$z = t(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}},$$

jonka suuret arvot ovat nollahypoteesille kriittisiä.

Voimafunktio on

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

Kuva 5.1 Hypoteesin $H_0 : \mu \leq 0$ 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.



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 5.1. 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 5.2 näytetään tämän testin voimafunktio, kun testin koko on $\alpha = 0.05$. Vaihtoehtohypoteesin mukaisilla parametrinarvoilla suuremmalla otoskoolla saavutetaan suurempi voima.

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

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

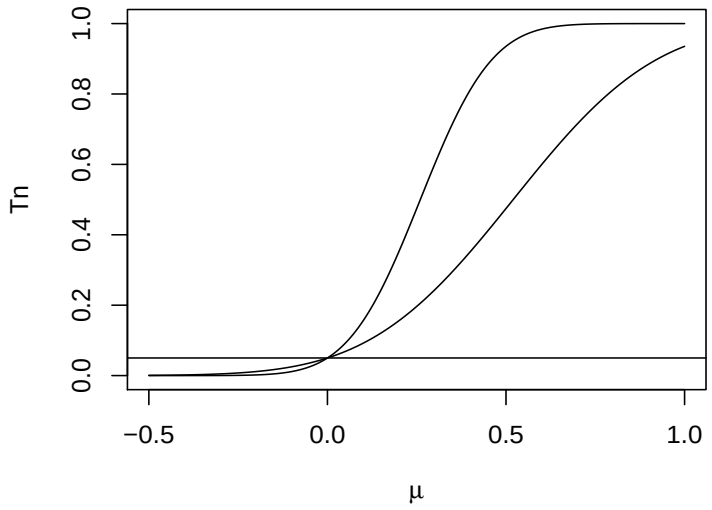
Tämän takia käsittelemämme yksisuuntainen z-testi on tason α testi myös hypoteesiparille

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

Tälle hypoteesiparille

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

Kuva 5.2 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.



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

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

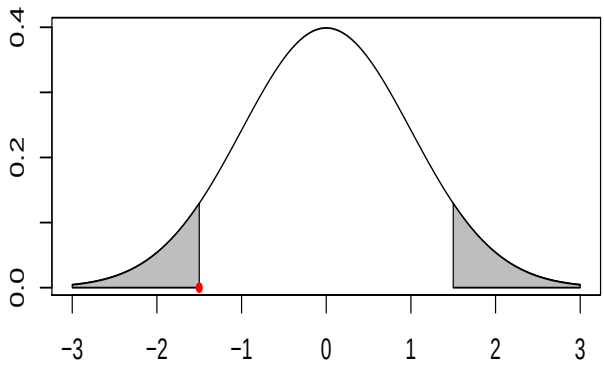
Tämä funktio on aidosti vähenevä, joten

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

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

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

Kuva 5.3 Kaksisuuntaisen testin p-arvon määrittäminen, kun havainnosta saadaan $z = -1.5$.



Kaksisuuntainen z-testi

Kaksisuuntaisen testin

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

testaamisessa testisuureen

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

sekä suuret että pienet (negatiiviset) 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 5.3. Tällä perusteella

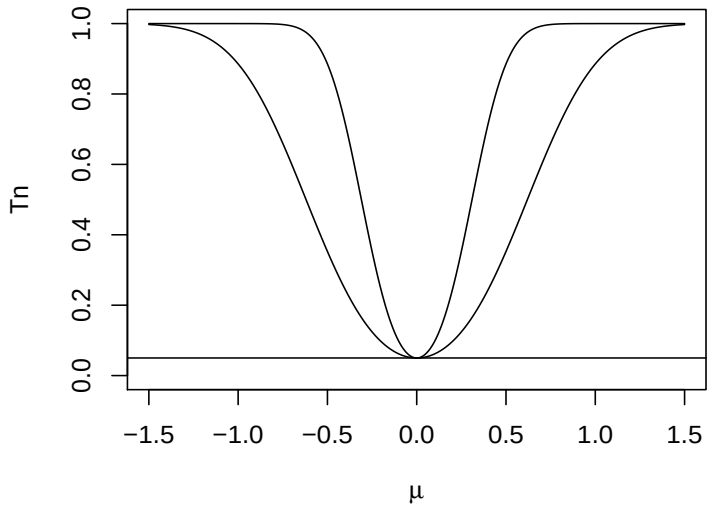
$$p = P[|Z| \geq |z|] = \Phi(-|z|) + 1 - \Phi(|z|) = 2(1 - \Phi(|z|))$$

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

Kuva 5.4 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] 2.29581358 1.64485363 0.01084327 0.15658245
```

Taulukko 5.1 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
$\mu \leq \mu_0$	$\mu > \mu_0$	$z > z_\alpha$	$P(Z \geq z)$	$[\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$z < -z_\alpha$	$P(Z \leq z)$	$(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}]$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z\rvert > z_{\alpha/2}$	$P(\lvert Z\rvert \geq \lvert z\rvert)$	$[\bar{y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$

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.

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] 0.7146748 1.9599640 0.4748100 0.1568721
```

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 $\mu_0 = \frac{1}{2}$.

5.6 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 (5.7) ei hylkää. Testi ei hylkää, mikäli $|z| \leq z_{\alpha/2}$ eli mikäli

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

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 (4.10).

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 5.1 luetellut tapaukset.

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

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

vastaa parametrin μ yksisuuntainen 95 %:n luottamusväli $[0.20, \infty)$, joka ei sisällä esimerkissä kiinnostavaa arvoa 0, joten nollahypoteesi $H_0 : \mu \leq 0$ hylätään merkitsevyystasolla 0.05. Kaksisuuntaista hypoteesia

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

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.

Testaaminen johtaa käyttäjän helposti mustavalkoiseen ajatteluun: nollahypoteesi joko hyväksytään tai hylätään. Tällöin käyttäjän huomio kiinnittyy pois siitä, kuinka epävarmaa aineiston antama informaatio parametrista on. Sen sijaan luottamusväli kvantifioi epävarmuuden selkeällä tavalla. Kun pisteestimaatti 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 *luottamusvälien laskeminen on parempi tapa analysoida aineistoa kuin testien suorittaminen* sellaisissa yksinkertaisissa tilanteissa, joissa molemmat lähestymistavat ovat mahdollisia.

Suositus: Laske mieluummin piste-estimaatteja ja luottamusvälejä. Älä testaa ellei sinun ole pakko.

Taulukko 5.2 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
$\mu \leq \mu_0$	$\mu > \mu_0$	$t > t_{n-1}(\alpha)$	$P(T \geq t)$	$[\bar{y} - t_{n-1}(\alpha)\frac{s}{\sqrt{n}}, \infty)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$t < -t_{n-1}(\alpha)$	$P(T \leq t)$	$(-\infty, \bar{y} + t_{n-1}(\alpha)\frac{s}{\sqrt{n}}]$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert > t_{n-1}(\frac{\alpha}{2})$	$P(\lvert T\rvert \geq \lvert t\rvert)$	$[\bar{y} - t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}]$

5.7 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 5.2. Käytännössä t-testi suoritetaan aina jollakin tarkoitukseen sopivalla tietokoneohjelmalla. Esim. R-ohjelmistosa kaikki taulukon 5.2 tulokset saadaan laskettua vaivattomasti funktiolla `t.test`.

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.

5.8 Binomijakauman parametrin testaus

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

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

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 $X \sim \text{Bin}(n, 1/2)$, ja x on havaittu onnistumisten lukumäärä.

Edellä tarvittavat binomijakauman häntätodennäköisyydet saadaan laskettua suoraan tietokoneohjelmilla. Esimerkiksi, jos $n = 1000$ ja onnistumisia on tullut $x = 475$, niin tämä testi saadaan tehtyä 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:
 0.4287663 0.4914698
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 4.8 mainittu Clopperin ja Pearsonin tarkka luottamusväli parametrille p.

Yksisuuntaiset testit saadaan käsiteltyä samaan tapaan.

5.9 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. Useissa tilastotieteen soveltajien kirjoittamissa oppikirjoissa testaukseen liittyvät keskeiset käsitteet selitetään joko väärin tai vähintään harhaanjohtavasti. Yritän seuraavalla yliyksinkertaistetulla esimerkillä ehkäistä eräiden harhakäsitysten muodostumista.

Tarkastelemme yksinkertaista esimerkkiä, jossa meillä on yksi havainto $N(\mu, 1)$-populaatiosta. Populaation varianssia pidetään tunnettuna, ja nollahypoteesi ja vastahypoteesi ovat seuraavat yksinkertaiset hypoteesit,

$$H_0 : \mu = 0, \qquad H_1 : \mu = 10.$$

Havaittu arvo on $y = 3.1$.

Jos testisuure on $Z = (Y - 0)/\sigma$, ja käytetään yksisuuntaista z-testiä (5.6), niin p-arvo on

$$1 - \Phi(y) = 0.00097.$$

Nollahypoteesi hylätään vielä merkitsevyystasolla 0.001. Saatu havainto tukee kuitenkin paljon enemmän vastahypoteesia kuin nollahypoteesia, sillä $y = 3.1$ olisi erittäin kummallinen havainto vastahypoteesin mukaisesta populaatiosta.

Astutaan nyt hetkeksi frekventistisen kehikon ulkopuolelle, ja ajatellaan tilannetta, jossa luontoäiti arpoo satunnaisesti joko nollahypoteesin tai vastahypoteesin mukaisen populaation todennäköisyydellä $\frac{1}{2}$, ja sen jälkeen arpoo valitusta populaatiosta mallin mukaisen havainnon. Tällaisessa tilanteessa voimme puhua nollahypoteesin todennäköisyydestä, ja lähdemme sitä nyt laskemaan. Ennen havaintoa todennäköisyydet ovat

$$P(H_0) = \frac{1}{2}, \qquad P(H_1) = \frac{1}{2}.$$

Havainnon jälkeen nollahypoteesin todennäköisyys saadaan Bayesin kaavan eräällä versiolla, nimittäin kaavalla

$$\begin{aligned} P(H_0 \mid Y = y) &= \frac{P(H_0)\,P(Y = y \mid H_0)}{P(H_0)\,P(Y = y \mid H_0) + P(H_1)\,P(Y = y \mid H_1)} \\ &= \frac{P(H_0)\,f_Y(y \mid H_0)}{P(H_0)\,f_Y(y \mid H_0) + P(H_1)\,f_Y(y \mid H_1)} \end{aligned} \tag{5.10}$$

Tässä kaavassa $f_Y(y \mid H_0)$ tarkoittaa normaalijakauman $N(0,1)$ tiheysfunktion arvoa pisteessä y, ja $f_Y(y \mid H_1)$ normaalijakauman $N(10,1)$ arvoa pisteessä y.

Kaava (5.10) voidaan perustella soveltamalla Bayesin kaavaa (tai ehdollisen todennäköisyyden määritelmää) sekä ottamalla raja-arvo, seuraavalla tavalla. Jos $\epsilon > 0$, niin

$$\begin{aligned} P[H_0 \mid Y \in (y-\epsilon, y+\epsilon)] &= \frac{P[H_0 \text{ ja } Y \in (y-\epsilon, y+\epsilon)]}{P[Y \in (y-\epsilon, y+\epsilon)]} \\ &= \frac{P(H_0)\,P[Y \in (y-\epsilon, y+\epsilon) \mid H_0]}{P(H_0)\,P[Y \in (y-\epsilon, y+\epsilon) \mid H_0] + P(H_1)\,P[Y \in (y-\epsilon, y+\epsilon) \mid H_1]} \\ &= \frac{P(H_0)\,\int_{y-\epsilon}^{y+\epsilon} f_Y(u \mid H_0)\,\mathrm{d}u}{P(H_0)\,\int_{y-\epsilon}^{y+\epsilon} f_Y(u \mid H_0)\,\mathrm{d}u + P(H_1)\,\int_{y-\epsilon}^{y+\epsilon} f_Y(u \mid H_1)\,\mathrm{d}u} \\ &\approx \frac{p(H_0)\,2\,\epsilon f_Y(y \mid H_0)}{p(H_0)\,2\,\epsilon f_Y(y \mid H_0) + p(H_1)\,2\,\epsilon f_Y(y \mid H_1)} \\ &= \frac{P(H_0)\,f_Y(y \mid H_0)}{P(H_0)\,f_Y(y \mid H_0) + P(H_1)\,f_Y(y \mid H_1)} \end{aligned}$$

Yllä approksimaatio tulee aina vain paremmaksi (esim. integraalilaskennan väliarvolauseen nojalla), kun $\epsilon > 0$ lähestyy nollaa, joten rajalla saadaan väitetty kaava.

Seuraavaksi sijoitamme annetut numerot kaavaan (5.10), jolloin saamme tulokset

$$P(H_0 \mid Y = 3) = 0.9999999944, \qquad P(H_1 \mid Y = 3) = 0.0000000056.$$

Huomaa, että testin p-arvolla $p = 0.00097$ ei ole mitään tekemistä näistä kummankaan luvun kanssa. Tässä esimerkissä testin p-arvo oli pieni, ja nollahypoteesi hylättäisiin tilastollisella testillä kaikilla konventionaalisilla merkitsevyystasoilla, mutta nollahypoteesi on aineiston perusteella erittäin todennäköinen.

Tämän esimerkin jälkeen astumme takaisin frekventistisen tilastollisen päättelyn viitekehykseen. Tilastotieteen soveltajat usein kuvittelevat naiivisti, että p-arvo on todennäköisyys sille, että nollahypoteesi pitää paikkansa. Jotta tällaiseen täysin virheelliseen käsitykseen voisi päätyä, pitää tehdä monta vakavaa virhettä:

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

2. Ajatellaan, että $P(H_0 \mid Y = y)$ on sama asia kuin $P(Y = y \mid H_0)$.

3. Ajatelaan, 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. Tässä yhteydessä oikea tulkinta kaavalle $P(Y = y \mid H_0)$ olisi todennäköisyystiheys $f_Y(y \mid H_0)$.

5.10 Tilastollisten testien väärinkäyttöä

Monille tilastotieteen soveltajille on syntynyt sellainen mielikuva, että kokeellisen tutkimuksen päämäränä on laskea p-arvo jollekin testille. Jos p-arvo on riittävän pieni, 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 ilmiötä tässä mitattiin? Mitä yksikköjä käytettiin? jne.) Tilastollinen merkitsevyys (engl. *statistical significance*) ja käytännön merkittävyys (engl. *practical significance*) ovat aivan eri asioita.

On paljon 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 ja intoon tulkita virheellisesti testien tuloksia 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 juuri mitään uutta.

Tilastotieteilijät saavat suurelta osalta syyttää itseään siitä, että tätä yhtä tilastotieteen välinettä on niin paljon käytetty väärin. Eräs tärkeä tekijä on ollut se, että tilastotieteen teknisiä käsitteitä kuvaamaan on valittu yleiskielestä sellaisia termejä, kuten merkitsevä (engl. *significant*), hyväksyä (engl. *accept*), hylätä (engl. *reject*), virhe (engl. *error*), voima (engl. *power*). Näillä termeillä on erittäin vahva merkitys yleiskielessä, eikä tilastotieteen soveltaja välttämättä ymmärrä, milloin sanoja käytetään yleiskielen merkityksessä ja milloin teknisinä termeinä. Tekniset termit saavat naiivin soveltajan mielessä yleiskielestä tutun ja testauksen yhteydessä turhan juhlallisen merkityksen.

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 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.