7 Kahden populaation vertaaminen

- Vertaamme frekventistisen tilastotieteen keinoin 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}.
- Tarkastelemme nyt tällaisten toisistaan riippuvien otosten eli parittaisten (engl. *paired, related, matched*) otosten analysiointia.

Parittaisissa otoksissa tarkastellaan erotuksia

- Tarkastellaan 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 voidaan soveltaa t-luottamusväliä tai t-testiä erotuksiin d_i.

- Populaatioiden odotusarvojen erotus on

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

ja tyypillisesti σ^2 on tuntematon.

Parittainen t-luottamusväli tai t-testi

- Odotusarvoparametrien erotusta $\delta = \mu_1 - \mu_2$ voidaan nyt analysoida soveltamalla t-luottamusväliä tai t-testiä erotuksiin

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

- Saadaan parittainen t-luottamusväli tai parittainen t-testi.

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}.
- 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.

Milloin käytetään riippumattomien satunnaisotosten mallia

- 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.

Riippumattomien satunnaisotosten malli

- Oletamme havaintoja vastaavien satunnaismuuttujien Y_{ki} noudattavan jakaumia

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \sim N(\mu_1, \sigma_1^2) \quad (1)$$

$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \sim N(\mu_2, \sigma_2^2). \quad (2)$$

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

Estimaatteja ja estimaattoreita

- Populaatioiden parametreja (μ_1, σ_1^2) ja (μ_2, σ_2^2) voidaan estimoida otoskeskiarvolla ja otosvarianssilla siten, että populaation k parametrit estimoidaan populaatiosta k saadusta otoksesta.
- Alaindeksi kertoo, kummasta populaatiosta otossuureet on laskettu.

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

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

- Merkitsemme vastaavia estimaattoreita

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

Estimaattorien otantajakauma

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

$$\bar{Y}_1 \sim N(\mu_1, \frac{1}{n_1}\sigma_1^2) \qquad \bar{Y}_2 \sim N(\mu_2, \frac{1}{n_2}\sigma_2^2) \quad (3)$$

$$\frac{n_1 - 1}{\sigma_1^2} S_1^2 \sim \chi^2_{n_1 - 1} \qquad \frac{n_2 - 1}{\sigma_2^2} S_2^2 \sim \chi^2_{n_2 - 1}, \quad (4)$$

  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.

Kiinnostava parametri ja sen estimaatti

- Kinnostuksen kohteena on populaatioiden odotusarvojen erotus

$$\delta = \mu_1 - \mu_2.$$

- Sitä estimoidaan vastaavalla otoskeskiarvojen erotuksella,

$$\hat{\delta} = \bar{y}_1 - \bar{y}_2. \tag{5}$$

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

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2) \tag{6}$$

Erilaisia oletuksia variansseista

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

1. varianssit tunnettuja
2. varianssit yhtäsuuria, mutta tuntemattomia
3. varianssit erisuuria ja tuntemattomia.

Varianssit tunnettuja —- z-luottamusväli ja z-testi

- 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 $N(0,1)$.
- Nyt saadaan tuttuun tapaan johdettua luottamusväli odotusarvojen erotukselle $\delta = \mu_1 - \mu_2$ tai voidaan johtaa testit kaikille tapauksille

$$H_0 : \delta \leq \delta_0, \quad H_1 : \delta > \delta_0$$
$$H_0 : \delta \geq \delta_0, \quad H_1 : \delta < \delta_0$$
$$H_0 : \delta = \delta_0, \quad H_1 : \delta \neq \delta_0.$$

Tarkka hypoteesi on nyt yhdistetty

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:
$$\sigma_1^2 = \sigma_2^2 = \sigma^2.$$
Tässä σ^2 on tuntematon parametri.
- 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

- Muodostamme yhteiselle varianssille σ^2 sellaisen estimaattorin 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.

χ^2-jakauman ominaisuuksia

- Jos $X \sim \chi^2_\nu$, niin sen odotusarvo on

$$EX = \nu, \tag{7}$$

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

- χ^2-jakauman yhteenlaskuominaisuus: 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

$$X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2} \tag{8}$$

Tämä voidaan johtaa esim. gammajakauman yhteenlaskuominaisuudesta.

Yhdistetty varianssiestimaattori

Edellisten tietojen 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

$$\frac{n_1 + n_2 - 2}{\sigma^2} S_P^2 \sim \chi^2_{n_1+n_2-2}, \tag{9}$$

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

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{10}$$

$$= \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2 \right] \tag{11}$$

S_p^2 on harhaton

Yhdistetty varianssiestimaattori on harhaton, sillä jakaumatulosta (9) sekä χ^2-jakauman odotusarvon kaavaa käyttämällä

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

t-jakaumaa noudattava satunnaismuuttuja

- Otetaan huomioon, että

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

- Seuraus: seuraavalla satunnaismuuttujalla on t-jakauma vapausasteluvulla $n_1 + n_2 - 2$,

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

- Sieventämällä nähdään, että

$$t(\mathbf{Y}) = \frac{\bar{Y}_1 - \bar{Y}_2 - \delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \tag{12}$$

joten $t(\mathbf{Y})$ on saranasuure kiinnostavalle parametrille $\delta = \mu_1 - \mu_2$.

Luottamusvälin johto

Luottamustason $0 < 1 - \alpha < 1$ kaksisuuntainen luottamusväli johdetaan lähtemällä liikkeelle tuloksesta

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

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

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

t-luottamusväli

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

$$\hat{\delta} = \bar{y}_1 - \bar{y}_2 \tag{13}$$

sekä yhdistetty varianssiestimaatti

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}, \tag{14}$$

minkä jälkeen luottamusväli on muotoa

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

eli tarkemmin sanoen se on

$$\left[\hat{\delta} - t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \hat{\delta} + t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]. \tag{15}$$

t-testit

- Testisuuretta

$$t = t(\mathbf{y}) = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{16}$$

  verrataan t-jakaumaan vapausasteluvulla $n_1 + n_2 - 2$.

- Merkitsevyystason $0 < \alpha < 1$ kaksisuuntainen testi

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

  hylkää nollahypoteesin silloin, kun

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

Yksisuuntaiset t-testit

- Yksisuuntainen testi hypoteeseille

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

hylkää nollahypoteesin, jos

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

- Yksisuuntainen testi hypoteeseille

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

hylkää nollahypoteesin, jos

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

Huomautuksia testeistä

- Tavallisesti näissä testeissä $\delta_0 = 0$, sillä tavallisimmin 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.