Johdatus tilastolliseen päättelyyn, kurssikoe 7.5.2012

Saat käyttää käsinkirjoitettua lunttilappua (yksi A4-arkki). Taulukkokirjan käyttö ei ole sallittua.

1. Käsillä on kaksi kolikkoa M_1 ja M_2, joista M_1:lle kruunan todennäköisyys on $p_1 = 0.4$ ja M_2:lle vastaavasti $p_2 = 0.7$. Kokeen johtaja valitsee kolikoista yhden ja antaa sen henkilölle, joka heittää kolikkoa kolme kertaa. Tulokseksi saatiin kruuna, klaava ja kruuna. Parametri $\theta = 1$, jos valittu kolikko oli M_1 ja toisessa tapauksessa $\theta = 2$.

a) Kirjoita uskottavuusfunktio.

b) Määritä θ:n suurimman uskottavuuden estimaatti.

Ratkaisu. a) Tehtävänannon mukaan parametriavaruus on kahden pisteen joukko $\{1, 2\}$. Uskottavuusfunktion L arvot pisteessä $\theta = 1$ ja $\theta = 2$ ovat

$$L(1) = 0.4 \times 0.6 \times 0.4 = 0.096, \qquad L(2) = 0.7 \times 0.3 \times 0.7 = 0.147$$

b) Koska $L(2) > L(1)$, niin SU-estimaatti on $\hat{\theta} = 2$.

Huomautuksia. Jotkut opiskelijat saivat tästä tehtävästä täydet pisteet kolmen rivin pituisella vastauksella. Uskottavuusfunktio saadaan havaintosatunnaisvektorin yhteispistetodennäköisyysfunktiosta, kun se lasketaan havaitulle jonolle $(1, 0, 1)$ (1 vastaa kruunaa ja 0 klaavaa):

$$\begin{aligned} f(\mathbf{y}; \theta) &= P(Y_1 = 1, Y_2 = 0, Y_3 = 1; \theta) = P(Y_1 = 1; \theta)\, P(Y_2 = 0; \theta)\, P(Y_3 = 1; \theta) \\ &= p_\theta\, (1 - p_\theta)\, p_\theta \end{aligned}$$

Tässä oletetaan (tavalliseen tapaan), että valitulla kolikolla tehdyt heitot ovat keskenään riippumattomia.

Yleisiä virheitä.

- Ymmärrettiin tehtävänannon vastaisesti, että uskottavuusfunktio olisi binomikokeen uskottavuusfunktio siten, että parametriavaruus olisi väli $(0, 1)$ tai $[0, 1]$.
- Kirjoitettiin uskottavuusfunktio käyttämällä vain ensimmäisen heiton tulosta (tai kahden ensimmäisen heiton tuloksia).
- Kirjoitettiin uskottavuusfunktiolle kaavoja, joiden mukaan molempia kolikkoja olisi heitetty, ja sitten nämä tulokset oltaisiin yhdistetty jollakin hämärällä tavalla.
- Ei kirjoitettu minkäänlaista kaavaa uskottavuusfunktiolle.
- Kirjoitettiin jokin sellainen luennoilla esiintynyt uskottavuusfunktio, jolla ei ole mitään yhteyttä tämän tehtävän tilanteen kanssa.
- Väitettiin, että uskottavuusfunktio saataisiin summaamalla todennäköisyydet

$$p_\theta + (1 - p_\theta) + p_\theta$$

- B-kohdassa ei maksimoitu a-kohdassa johdettua uskottavuusfunktiota.

2. Havainnot ovat luvut $y_1, \ldots, y_n$. Mallin mukaan satunnaismuuttujat $Y_1, \ldots, Y_n$ noudattavat riippumattomasti jakaumaa $N(\mu, \sigma^2)$, jossa sekä μ että σ^2 ovat tuntemattomia. Tahdomme testata hypoteesia $\mu = \mu_0$, kun vastahypoteesina on $H_1 : \mu > \mu_0$. Testi perustuu tunnuslukuun

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

a) Kuinka suureet $\bar{y}$ ja s määritellään?

b) Selitä, kuinka tämän testin p-arvo lasketaan (taulukoiden tai tietokoneen avulla).

c) Hylätäänkö vai hyväksytäänkö hypoteesi H_0, kun p-arvo on $p = 0.021$ ja merkitsevyystaso $\alpha = 0.05$?

Ratkaisu. a) $\bar{y}$ on otoskeskiarvo

$$\frac{1}{n} \sum_{i=1}^{n} y_i,$$

ja s on otoskeskihajonta (eli otosvarianssin neliöjuuri)

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Yleisiä virheitä a-kohdassa: otosvarianssin jakajana n eikä $n - 1$; unohdetaan ottaa neliöjuuri otosvarianssista. Jotkut opiskelijat luulevat, että on samantekevää, otetaanko neliöjuuri summasta vai summataanko yhteen juurretut termit.

b) P-arvo on se todennäköisyys, jolla nollahypoteesin mukaisesta populaatiosta saadaan testisuureen arvo, joka on vähintään yhtä kummallinen kuin aineistosta laskettu testisuureen arvo. Nollahypoteesin mukaiselle populaatiolle testisuureella on t_{n-1}-jakauma, ja nyt suuret testisuureen arvot ovat nollahypoteesille kriittisiä, joten

$$p = P(T > t) = 1 - F(t),$$

jossa $T \sim t_{n-1}$, ja F tarkoittaa t_{n-1}-jakauman kertymäfunktiota. Tämän arvon voi joko lukea sopivasta taulukosta tai laskea jollakin tietokoneohjelmalla.

Yleisiä virheitä b-kohdassa: ei lainkaan osata määritellä p-arvoa; sekoitetaan keskenään toisaalta testisuureen kriittinen arvo ja toisaalta testin p-arvo; käytetään määrittelyssä väärää jakaumaa (tavallisesti $N(0, 1)$) vertailujakaumana.

c) Koska p-arvo 0.021 on pienempi kuin merkitsevyystaso $\alpha = 0.05$, niin H_0 hylätään.

Yleisiä virheitä c-kohdassa: tehdään väärä päätös.

Luennoitsijan johtopäätös: Testin p-arvo on opiskelijoille vaikea käsite. Se pitää tulevina vuosina selittää seikkaperäisemmin ja havainnollisemmin, koska sitä käytetään lähes jokaisessa tilastotiedettä soveltavassa artikkelissa.

3. 96 nopanheitossa saatiin seuraavat määrät ykkösiä, kakkosia, jne.: 15, 7, 9, 20, 26, 19. Testaa merkitsevyystasolla $\alpha = 0.05$ hypoteesia H_0, jonka mukaan noppa on harhaton. Toisin sanoen H_0:n mukaan kunkin silmäluvun (1–6) todennäköisyys on sama.

Ratkaisu. Tässä tehtävässä voi oman valinnan mukaan käyttää joko Pearsonin χ^2-testisuuretta X^2 tai uskottavuusosamäärän testisuureetta W, joiden kaavat ovat

$$X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$
$$W = 2\sum_{i=1}^{6} O_i \log \frac{O_i}{E_i}$$

Tässä tehtävässä

$$E_i = 96 \times \frac{1}{6} = 16 \qquad \text{kaikilla } i = 1, \ldots, 6.$$

Testisuureet saavavat arvot

$$X^2 = 16, \qquad W = 16.8$$

Näitä verrataan $\chi^2(5)$-jakauman 0.05-yläkvantiiliin eli 0.95-kvantiiliin, joka on taulukon mukaan 11.07, ja suuret testisuureen arvot ovat nollahypoteesille kriittisiä. Nollahypoteesi hylätään kumpaa tahansa testisuuretta käyttämällä.

Yleisiä virheitä:

- Testisuureetta verrataan $\chi^2(5)$-jakauman 0.05-alakvantiiliin (tai muuten väärään kvantiiliin).
- Käytetään O_i:n tai E_i:n paikalla suhteellisia frekvenssejä eikä frekvenssejä
- Yritetään selvittää nopan harhattomuutta tekemällä jokin järjetön t-testi kuten esim.
  - yhtenä testattavista hypoteeseista on se, että jokin silmäluvuista 1–6 ylipäänsä tulee
  - testataan, onko nopan silmäluku keskimäärin 1/6.
- Tilanteen kannalta järkevästä t-testistä sai enintään 3 pistettä.

4. Tarkastelemme kahta riippumatonta satunnaisotosta

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \sim N(\mu_1, \sigma_1^2)$$
$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \sim N(\mu_2, \sigma_2^2)$$

Varianssit σ_1^2 ja σ_2^2 ovat tunnettuja, mutta odotusarvot μ_1 ja μ_2 ovat tuntemattomia. Kiinnostuksen kohteena on parametri $\tau = \mu_1 + 2\,\mu_2$.

a) Esitä parametrille τ harhaton estimaattori.

b) Kuinka parametrille τ lasketaan 95 %:n luottamusväli?

Ratkaisu. (Tämä tehtävä ratkeaa soveltamalla kappaleen 6 ideoita.)

a) (Tästä kohdasta max 2 pistettä) Parametria μ_1 voidaan estimoida harhattomalla estimaattorilla

$$\bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i}$$

ja parametria μ_2 harhattomalla estimaattorilla

$$\bar{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{2i}.$$

Kun näistä muodostetaan lineaarikombinaatio $\hat{\tau}(\mathbf{Y}) = \bar{Y}_1 + 2\,\bar{Y}_2$, niin saadaan harhaton estimaattori parametrille τ, sillä

$$E\hat{\tau}(\mathbf{Y}) = E[\bar{Y}_1 + 2\,\bar{Y}_2] = E\bar{Y}_1 + 2\,E\bar{Y}_2 = \mu_1 + 2\,\mu_2 = \tau.$$

Yleisiä virheitä a-kohdassa: esitetään harhainen estimaattori; ei anneta estimaattorille mitään kaavaa; annetaan kaava $\hat{\tau}(\mathbf{Y}) = \mu_1 + 2\,\mu_2$ joka ei kelpaa estimaattoriksi (koska μ_1 ja μ_2 ovat tuntemattomia parametreja).

b) (Tästä kohdasta max 4 pistettä) Estimaattori $\hat{\tau}(\mathbf{Y})$ on lineaarikombinaatio kahdesta riippumattomasta normaalijakaumaa noudattavasta satunnaismuuttujasta

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

minkä takia sillä on itsellään normaalijakauma, jonka parametrit ovat $E\hat{\tau}(\mathbf{Y}) = \tau$ ja

$$\operatorname{var} \hat{\tau}(\mathbf{Y}) = \operatorname{var} \bar{Y}_1 + 2^2 \operatorname{var} \bar{Y}_2 = \frac{1}{n_1}\,\sigma_1^2 + \frac{4}{n_2}\,\sigma_2^2.$$

Siispä estimaattorista $\hat{\tau}(\mathbf{Y})$ saadaan standardinormaalijakaumaa noudattava satunnaismuuttuja standardoimalla,

$$\frac{\hat{\tau}(\mathbf{Y}) - \tau}{\sqrt{\frac{1}{n_1}\,\sigma_1^2 + \frac{4}{n_2}\,\sigma_2^2}} \sim N(0, 1).$$

Kun ratkaistaan tuloksesta

$$P\left(\left| \frac{\hat{\tau}(\mathbf{Y}) - \tau}{\sqrt{\frac{1}{n_1}\,\sigma_1^2 + \frac{4}{n_2}\,\sigma_2^2}} \right| \le z_{0.025} \right) = 0.95$$

epäyhtälö muuttujan τ suhteen, saadaan luottamusväli

$$\hat{\tau} \pm z_{0.025} \times \sqrt{\frac{1}{n_1}\,\sigma_1^2 + \frac{4}{n_2}\,\sigma_2^2}.$$

Tässä $z_{0.025} = 1.96$.

Yleisiä virheitä b-kohdassa: estimaattorin varianssi arvattu/järkeilty väärin.