30A02000 Tilastotieteen perusteet
Tieto- ja palvelujohtamisen laitos
Aalto-yliopisto

L Viitasaari
Syksy 2020

2. välikokeen vastaukset

1 a) **Ei totta.** (Huom. luottamusvälin tulkinta aiheuttaa klassisesti hankaluuksia, vrt. tehtävä 4c)

b) **Totta.**

c) **Ei totta.**

d) **Ei totta.**

e) **Totta.**

f) **Totta.**

2 Sanallinen tehtävä. Tyypillisimmät pistemenetykset kun listattuja käsitteitä puuttuu tai selitetty puolivillaisesti. Erityisesti p-arvon tulkinta todennäköisyytenä saada havaittuja tai vieläkin poikkeuksellisempia arvoja nollahypoteesin pätiessä.

3 a) Estimaattori on ”paras arvaus” tuntemattomalle parametrille perustuen satunnaismuuttujiin (kuvitteelliset havainnot jotka kenties joskus saadaan) ja siten satunnaismuuttuja myös. Estimaatti on lukuarvo joka saadaan kun havainnot joskus oikeasti kerätään.

b) Harhaton jos estimaattorin odotusarvo on todellinen parametri (eli keskimäärin saadaan oikeita tuloksia). Tarkentuva jos otoskoon kasvaessa estimaattori ”on lähempänä” todellista arvoa (missä mielessä satunnaismuuttuja on ”lähellä” annettua ei satunnaista numeroa on haastavampi käsitellä eikä kuulu tälle kurssille)

c) $\hat{\mu} = 4.8$ ja $\hat{\sigma} = 3.259$. Tyypilliset pistemenetykset jos käytetty vain excelin komentoa hajonnalle perustelematta harhattomuutta, tai käytetty esim. excelin väärää komentoa `hajonta.p`.

4 a) lineaarisessa regressiossa selitetään/ennustetaan muuttujaa y toisen muuttujan x avulla (voi olla toki useita) lineaarisesti. Toimii sitä paremmin mitä voimakkaampaa lineaarinen riippuvuus muuttujien välillä on.

b) Luottamusväli on muotoa $\overline{x} \pm z_\alpha \frac{s}{\sqrt{n}}$ missä nyt otoskoko $n = 10$, keskiarvo $\overline{x} = \hat{\mu} = 4.8$ ja hajonta $s = \hat{\sigma} = 3.259$ (tehtävä 3c), ja ehdosta $P(-z_\alpha < Z < z_\alpha) = 0.9$ rajaksi $z_\alpha = 1.645$. Saadaan 4.8 ± 1.690

c) Virheitä on artikkelissa useita ja tekstiä voi myös tulkita monella tavalla. Erilaiset virheet ja tulkinnat kelpasivat kunhan perustelut ovat oikean suuntaiset. Tyypillisimmät pistemenetykset luottamusvälin väärästä tulkinnasta, esim. perustelusta ”viiden prosentin todennäköisyys, että oikea kannatus on välin 8-12 ulkopuolella” (vrt. tehtävä 1a)

5 a) Kyseessä testi odotusarvolle kaksisuuntaisella $H_1 : \mu \neq 10$. Testisuure -2.38 ja p-arvo $0.017 > 0.01 \Rightarrow H_0$ jää voimaan eli linjassa ei vikaa. Tyypilliset pistemenetykset väärästä testausasetelmasta ja virheellisistä johtopäätöksistä

b) Kyseessä odotusarvojen parivertailu yksisuuntaisella $H_1 : \mu_1 < \mu_2$. Testisuure -1.88 ja p-arvo $0.03 < 0.05 \Rightarrow H_0$ hylätään eli linja 2 tuottaa pidempiä ruuveja. Tyypilliset pistemenetykset väärästä testausasetelmasta ja virheellisistä johtopäätöksistä

6 a) Kyseessä testi suhteelliselle osuudelle yksisuuntaisella $H_1 : \pi_0 < 0.12$. Testisuure -1.46 ja p-arvo $0.07 > 0.05 \Rightarrow H_0$ jää voimaan eli kannatus ei ole laskenut. Tyypilliset pistemenetykset väärästä testausasetelmasta ja virheellisistä johtopäätöksistä

b) Kyseessä parivertailu suhteellisille osuuksille yksisuuntaisella $H_1 : \pi_1 > \pi_2$ (ryhmä 1 on pääkaupunkiseudulla asuvat). Testisuure 0.75 ja p-arvo $0.226 > 0.01 \Rightarrow H_0$ jää voimaan eli kannatus ei ole suurempaa pääkaupunkiseudulla. Tyypilliset pistemenetykset väärästä testausasetelmasta ja virheellisistä johtopäätöksistä

7 a) Keskeisen raja-arvolauseen mukaan (oikein skaalattu) summa riippumattomia satunnaismuuttujia on likimain normaalinen, joten testisuureissa esiintyvät keskiarvot ja/tai suhteelliset osuudet ovat likimain normaalisia. Siten esimerkiksi odotusarvon testissä testisuure $t = \frac{\overline{X}}{\frac{\sigma}{\sqrt{n}}}$ on likimain normaalinen (suora seuraus keskeisestä raja-arvolauseesta). Toisaalta jos hajontaa σ ei tunneta, käytetään $t = \frac{\overline{X}}{\frac{s}{\sqrt{n}}}$ missä nyt myös s on satunnainen, ja siksi keskeisen raja-arvolauseen soveltaminen ei suoraan onnistu. Vaikka tarkka matemaattinen tarkastelu ei kuulu kurssin aihepiiriin, vaadittiin täysiin pisteisiin tämän seikan edes ylimalkainen huomioiminen. Toisin sanoen, tyypilliset pistemenetykset tulivat suppeista perusteluista ”koska oletetaan niin” tai lyhyt ”koska keskeinen raja-arvolause”.

b) Kummallisuuksia löytyy useita ja moni oli huomannut vaikkapa sen, että testisuureen arvo 3.03 otoskoolla 50 on hyvin erikoinen, sillä se tarkoittaisi otokseen osuneen 2.63 vikaantunutta! Lisäksi erittäin pieni p-arvo johtaa nollahypoteesin (herkkään, tavanomaisilla merkitsevyystasoilla) hylkäämiseen. Tällöin vaihtoehtoinen hypoteesi astuu voimaan ja todetaan, että vikaantuneiden osuus on jotain muuta kuin väitetty 0.01. Yritysjohtaja tekee siis johtopäätökset nurinpäin! Erityisenä piirteenä huomaamme myös, että nollahypoteesin pätiessä todellinen osuus $\pi_0 = 0.01$ on niin pieni, että binomijakauman normaaliapproximaatioon otoskoko $n = 50$ ei riitä alkuunkaan, ja normaalijakauman käyttö on erittäin epäilyttävää! Yleisesti pieni otoskoko oli huomioitu epäilyttävänä yksityiskohtana, mutta ilman sen yhdistämistä normaaliapproximaation virheeseen ei täysiä pisteitä voinut saada. Tästä syntyi tyypilliset vähennykset pisteisiin.