Tilastotieteen johdantokurssi **5 op** Tentti 10.11.2020

Kaikki laskut voi tehdä ohjelmistolla tai käsin. Jos lisäät vastauksiisi R-ohjelmalla tehdessäsi käyttämäsi koodit ja käsin tehdessä välivaiheet, voit näiden perusteella saada osapisteitä vaikka lopullinen vastaus on väärin. Menestystä tenttiin!

1. (a) Tutkitaan ravinnon määrän vaikutusta vesiliskon lisääntymiseen. Muuttuja 'ruoka' kertoo vesiliskolle annettujen surviaisten toukkien lukumäärän ja muuttuja 'munat' kertoo vesiliskon munimien munien lukumäärän.

ruoka	1	2	3	4	5	6	7	8	9	10
munat	4	1	6	2	5	11	7	11	9	15

i. Laske muuttujan 'munat' keskiarvo (1.5 p) ja keskihajonta (1.5 p)

ii. Tarkastele muuttujien välistä riippuvuutta sopivan kuvion (1.5 p) ja korrelaatiokertoimen avulla (1.5 p)

(b) Tietoliikenneyhteyksissä esiintyvät viat ovat kiusallisia sekä asiakkaalle että yhteyksiä tarjoavalle operaattorille. Operaattori pyrkii korjaamaan vian mahdollisimman nopeasti saatuaan siitä tiedon asiakkaalta. Ohessa on operaattorin palvelukeskuksen 20 havainnon otos ajasta (minuutteina), joka operaattorilta meni vian korjaamiseen saatuaan siitä tiedon.

Korjausaika (min)	1.48	1.75	0.78	2.85	0.52	1.60	4.15	3.95	1.48	3.10
	1.02	0.53	0.93	1.60	0.80	1.05	6.32	3.93	5.45	0.97

Laadi korjausajan jakauman viiden numeron yhteenveto (4 p) ja piirrä sen perusteella jakaumaa kuvaava kuvio (1 p). Kuvaile sanallisesti minkä muotoinen jakauma on? (1 p)

2. Harhaisen nopan silmälukujen todennäköisyydet ovat seuraavat:

Silmäluku	1	2	3	4	5	6
Todennäköisyys	0.05	0.22	0.25	0.25	0.22	0.01

Määritellään satunnaismuuttuja $X =$ "nopan silmäluku"

(a) Piirrä X:n pistetodennäköisyysfunktio (2 p).

(b) Laske X:n odotusarvo (2 p).

(c) Laske X:n varianssi (2 p).

(d) Noppaa heitettiin kaksi kertaa. Laske todennäköisyys, että silmälukujen summa on pienempi tai yhtä suuri kuin 4 (2 p).

(e) Noppaa heitettiin sata kertaa. Laske arvio todennäköisyydelle, että silmälukujen summa on pienempi tai yhtä suuri kuin 320 (4 p)

3. Tarkastellaan poikkeavatko yli 75kg painavien ja alle 65 kg painavien äitien lasten syntymäpainot toisistaan. Alla ovat aineistossa olevien lasten painot kilogrammoissa (aineistossa ei ole kaksosia tai keskosia).

Äidin paino	Lapsen syntymäpaino						
yli 75kg	3.64	3.45	3.73	3.94	4.13	3.62	4.18
alle 65kg	3.11	4.19	2.77	2.88	3.11	3.55	2.85

Testaa ovatko lasten keskimääräiset syntymäpainot yhtä suuret eri painoisilla äideillä, $\alpha = 0.05$. Kirjoita näkyviin hypoteesit, testisuureen havaittu arvo, testisuureen jakauma nollahypoteesin ollessa totta ja tee päätelmät (12 p).

```
> leveneTest(lapsi.data$lapsen.paino, lapsi.data$paino.ryhma)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.5776 0.4619
      12
>
> t.test(lapsi.data$lapsen.paino[lapsi.data$paino.ryhma=="yli75"],
+  lapsi.data$lapsen.paino[lapsi.data$paino.ryhma=="alle65"],
+  paired = FALSE, var.equal=TRUE)

        Two Sample t-test

data:  lapsi.data$lapsen.paino[lapsi.data$paino.ryhma == "yli75"]
 and lapsi.data$lapsen.paino[lapsi.data$paino.ryhma == "alle65"]
t = 2.7795, df = 12, p-value = 0.01667
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1305888 1.0779826
sample estimates:
mean of x mean of y
 3.812857  3.208571

>
> t.test(lapsi.data$lapsen.paino[lapsi.data$paino.ryhma=="yli75"],
+  lapsi.data$lapsen.paino[lapsi.data$paino.ryhma=="alle65"],
```

```
+  paired = FALSE, var.equal=FALSE)

        Welch Two Sample t-test

data:  lapsi.data$lapsen.paino[lapsi.data$paino.ryhma == "yli75"]
 and lapsi.data$lapsen.paino[lapsi.data$paino.ryhma == "alle65"]
t = 2.7795, df = 9.295, p-value = 0.02079
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1148382 1.0937332
sample estimates:
mean of x mean of y
 3.812857  3.208571

>
> t.test(lapsi.data$lapsen.paino[lapsi.data$paino.ryhma=="yli75"],
+  lapsi.data$lapsen.paino[lapsi.data$paino.ryhma=="alle65"],
+  paired=TRUE)

        Paired t-test

data:  lapsi.data$lapsen.paino[lapsi.data$paino.ryhma == "yli75"]
 and lapsi.data$lapsen.paino[lapsi.data$paino.ryhma == "alle65"]
t = 2.213, df = 6, p-value = 0.06886
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06388696  1.27245838
sample estimates:
mean of the differences
              0.6042857
```

4. Valitse oikea vaihtoehto. Kussakin kysymyksessä **täsmälleen yksi annetuista vaihtoehdosta** on oikein. Kysymyksen pisteytys on seuraava:
Oikea vastaus = 2 pistettä,
väärä vastaus = -2 pistettä,
ei vastausta = 0 pistettä.
Huomaa, että tehtävän kokonaispistemäärä voi olla myös negatiivinen. Jos kokonaispistemäärä on negatiivinen, se vähentää muista kysymyksistä saatuja pisteitä.

(a) Kun satunnaismuuttujat X_1 ja X_2 ovat riippumattomia ja niiden varianssit ovat yhtä suuria, niin

i. $Var(X_1 - X_2) = 0$
ii. $Var(X_1 - X_2) = Var(X_1) + Var(X_2)$
iii. Varianssia $Var(X_1 - X_2)$ ei voida laskea, koska $Cov(X_1, X_2)$ on tuntematon.

(b) Jos vastaat tämän tehtävän kysymyksiin arvaamalla, niin pistemääräsi odotusarvo on

i. 0

ii. suurempi kuin nolla

iii. pienempi kuin nolla

(c) Muuttuja 'Jääkiekkoilijan pelinumero' on

i. luokitteluasteikollinen muuttuja

ii. järjestysasteikollinen muuttuja

iii. välimatka-asteikollinen muuttuja.

(d) Kun koeasetelmana on kaltaistetut parit, niin sopiva analyysimenetelmä on

i. Parittainen t-testi

ii. Riippumattomien otosten t-testi

iii. Levenen testi

(e) Suomen presidentinvaalien ensimmäisellä kierroksella 1000 äänestäjän satunnaisotoksella saavutetaan n. 2% virhemarginaali.

i. Yhdysvalloissa tarvitaan paljon suurempi otos, koska siellä äänestäjien populaatio on suurempi.

ii. Yhdysvalloissa riittää pienempi otos, koska siellä ehdokkaita on vain kaksi (pienpuolueista ei olla kiinnostuneita).

iii. Myös Yhdysvalloissa 1000 äänestäjän otos johtaa ainakin suunnilleen yhtä tarkkaan arvioon kuin Suomessa.

(f) Yliopiston henkilöstö käyttää paljon vuokra-autoja matkustamiseen mm. Kuopion ja Joensuun kampusten välillä. Taloushallinnon järjestelmissä on tiedossa laskun loppusumma ja eräpäivä, mutta ei tietoa siitä minä päivänä autoa on käytetty. Kuhunkin laskuun on kuitenkin liitetty skannanttu paperilasku, josta ko. tiedot löytyvät. Vuoden 2019 vuokra-autojen käyttöön liittyvien kasvihuonekaasupäästöjen selvittämistä varten taloushallinnon järjestelmästä poimittiin kaikki sellaiset laskut, joiden eräpäivä on vuonna 2019, yhteensä 1231 laskua. Laskut asetettiin eurosumman mukaan suuruusjärjestykseen pienimmästä suurimpaan excel-tiedostoon, yksi lasku kullekin riville. Satunnaislukugeneraattorilla tuotettiin 200 eri lukua väliltä $1, \ldots, 1231$, ja exceltiedostosta poimittiin valittuja lukuja vastaavat rivit. Valittuja riviä vastaavien laskujen skannatut paperilaskut tarkistettiin ja niistä kirjattiin ylös ajetut kilometrit. Valittujen laskujen keskimääräinen kilometrimäärä oli 345 kilometriä, minkä perusteella arvioitiin että vuonna 2019 vuokra-autoilla oli ajettu yhteensä $1231 \times 345 = 424695$ kilometriä.

i. Otos on harhainen, koska laskut järjestettiin aluksi loppusumman mukaan.
ii. Otantakehikossa on sekä yli- että alipeittoa.
iii. Kyseessä on satunnaisotanta palauttaen.