Tilastotieteen johdantokurssi **5 op** Tentti 14.12.2020

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. Tehtäväpaperin R-koodeista poimittuja arvoja ei tarvitse laskea uudestaan mutta niiden käyttö tulee perustella. Menestystä tenttiin!

1. Tutkitaan ravinnon määrän vaikutusta rupiliskon lisääntymiseen. Muuttuja 'ruoka' kertoo rupiliskolle annettujen surviaisten toukkien lukumäärän ja muuttuja 'munat' kertoo rupiliskon munimien munien lukumäärän.

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

(a) Laske muuttujan 'munat' keskiarvo (3 p) ja keskihajonta (3 p)

(b) Tarkastele muuttujien välistä riippuvuutta sopivan kuvion (3 p) ja korrelaatiokertoimen avulla (3 p)

2. Opiskelijalle tarjotaan osa-aikatyötä laboratoriossa. Viikoittain työaika vaihtelee 10 ja 25 tunnin välillä tasaisesti. Eri viikkojen työajat ovat riippumattomia. Viikoittaista työaikaa kuvaavan satunnaismuuttujan X tiheysfunktio on

$$f_X(x) = \begin{cases} \frac{1}{15}, & 10 \leq x \leq 25 \\ 0, & \text{muutoin} \end{cases}$$

Alla olevassa R-koodissa on laskettu joitakin tähän tiheysfunktioon liittyviä integraaleja, joita voit hyödyntää vastauksissa.

```
> f <- function(x) rep(1/15,length(x))
> integrate(f, 10, 25)$value
[1] 1
> f <- function(x) rep(1/15,length(x))
> xf <- function(x) x*f(x)
> integrate(xf, 10, 25)$value
17.5
> x2f <- function(x) x^2*f(x)
> integrate(x2f, 10, 25)$value
325
> xEX2f <- function(x) (x-17.5)^2*f(x)
> integrate(xEX2f, 10, 25)$value
[1] 18.75
```

(a) Piirrä X:n tiheysfunktion kuvaaja (2 p).

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

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

(d) Laske todennäköisyys, että kahdella peräkkäisellä viikolla kummankin viikon työaika jää alle 20 tuntiin. (2 p).

(e) Opiskelija saa palkaa 8.8 euroa/tunti. Opintotuen tuloraja on 8352 euroa vuodessa. Esitä perusteltu arvio todennäköisyydelle, että henkilön vuotuiset tulot ylittävät tulorajan 8352 euroa. Ajatellaan, että henkilö käy töissä vuoden jokaisella viikolla ja vuodessa on 52 viikkoa. (4 p)

3. Tarkastellaan poikkeavatko yli 75kg painavien ja alle 65 kg painavien äitien lasten syntymäpainot toisistaan. R:ssä vektoriin `yli75` on tallennettu yli 75kg painavien äitien lasten syntymäpainot ja vektoriin `alle65` alle 65kg painavien äitien lasten syntymäpainot. Testaa ovatko lasten keskimääräiset syntymäpainot yhtä suuret eri painoisilla äideillä, $\alpha = 0.05$. Kirjoita näkyviin hypoteesit. Selitä päättelyketju ja tee päätelmät. Mikä on testisuureen jakauma nollahypoteesin ollessa totta?

```
> lapsi.data <- data.frame(
+  lapsen.paino = c(yli75, alle65),
+ paino.ryhma = rep(c("yli75", "alle65"), c(20,20))
+ )
>
> 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  8.8513 0.005069 **
      38
>
> t.test(yli75, alle65, var.equal=T, paired=F)

        Two Sample t-test

data:  yli75 and alle65
t = 6.5566, df = 38, p-value = 9.85e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4953134 0.9377991
sample estimates:
mean of x mean of y
 3.915262  3.198706

> t.test(yli75, alle65, var.equal=F, paired=F)

        Welch Two Sample t-test

data:  yli75 and alle65
t = 6.5566, df = 24.42, p-value = 8.094e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4912012 0.9419113
sample estimates:
mean of x mean of y
 3.915262  3.198706

> t.test(yli75, alle65, paired=T)

        Paired t-test

data:  yli75 and alle65
t = 6.1276, df = 19, p-value = 6.856e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4717979 0.9613146
sample estimates:
mean of the differences
              0.7165562
```

4. Valitse oikea vaihtoehto. Kussakin kysymyksessä **täsmälleen yksi annetuista vaihtoehdosta** on oikein. Kysymyksen pisteytys on seuraava:
Oikea vastaus = 2 pistettä,
väärä vastaus = -1 pistettä,
ei vastausta = 0 pistettä.
Huomaa, että tehtävän kokonaispistemäärä voi olla myös negatiivinen. Jos kokonaispistemäärä on negatiivinen, se vähentää muista kysymyksistä saatuja pisteitä.

(a) Satunnaismuuttujat X_1 ja X_2 ovat riippumattomia ja $Var(X_1) = 2Var(X_2)$. Tarkastellaan satunnaismuuttujan $X_2 - X_1$ varianssia.

i. Varianssia $Var(X_2 - X_1)$ ei voida laskea, koska $Var(X_2) - Var(X_1)$ on negatiivinen.

ii. Varianssia $Var(X_2 - X_1)$ ei voida laskea, koska $Cov(X_1, X_2)$ on tuntematon.

iii. $Var(X_2 - X_1) = 3Var(X_2)$.

(b) Jos tiedät tämän kysymyksen jokaisessa kuudessa kohdassa varmasti yhden väittämän vääräksi, ja valitset vastauksesi kahdesta jäljelle jäävästä väittämistä arvaamallla, niin pistemääräsi odotusarvo koko tehtävästä on

i. 6

ii. 3

iii. 0

(c) Johdantokurssin kurssipalautteessa kysyttiin opiskelijan mielipidettä kurssin työmäärästä asteikolla 'liian pieni', 'sopiva' ja 'liian suuri'. Opiskelijoiden vastaukset sisältävä muuttuja on

i. luokitteluasteikollinen muuttuja

ii. järjestysasteikollinen muuttuja

iii. välimatka-asteikollinen muuttuja.

(d) Populaation varianssin tiedetään olevan 40. 10 havainnon otoksen keskiarvoksi saatiin arvo 5. 95% luottamusväli populaation odotusarvolle on

i. $5 \pm 1.96 \times \sqrt{40}$, eli väli $(-1.32, 11.32)$

ii. $5 \pm 1.96 \times 2$, eli väli $(1.08, 8.92)$

iii. $5 \pm 1.96 \times \sqrt{\frac{40}{10-1}}$, eli väli $(0.87, 9.13)$

(e) 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, jotka oli päivätty vuodelle 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 excel-tiedostosta 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. Kiinnostuksen kohteena oleva matkalaskujen populaatio on diskreetti

ii. Kiinnostuksen kohteena oleva matkalaskujen populaatio on ääretön

iii. Kiinnostuksen kohteena oleva matkalaskujen populaatio on luontevaa ajatella satunnaiseksi.

(f) Tutkija haluaa selvittää, miten kohotettu hiilidioksidipitoisuus vaikuttaa mustuvapajujen kasvuun. Kokeen alussa 100 kaksivuotiasta pajuntaimea siirretään laboratorioon. Näistä puolet on naaras- ja puolet urosyksilöitä. Näistä valitaan satunnaisesti 25 urosta ja 25 naarasta, joita kasvatetaan laboratoriossa normaalissa hiilidioksidipitoisuudessa. Loppuja kasvatetaan kohotetussa hiilidioksidipitoisuudessa. Kokeen alussa ja lopussa mitataan jokaisen pajuyksilön pituus, ja pituuksien erotuksena saadaan kasvu, jota analysoidaan koeasetelmaan sopivalla tilastollisella menetelmällä.

i. Se, että arvotaan sama määrä naaraita ja uroksia kumpaankin ryhmään on turhaa, koska ei olla kiinostuneita sukupuolen vaikutuksesta kasvuun.

ii. Olisi parempi arpoa käsittelyryhmät sukupuolesta välittämättä, jotta otos olisi täysin satunnaistettu.

iii. Jako sukupuolen suhteen on viisasta, koska näin varmistetaan että käsittelyryhmät ovat keskenään mahdollisimman samanlaisia.