Tilastotieteen peruskurssi **5 op** Tentti 16.4.2021

Tentissä on neljä kysymystä, ja jokaisesta maksimipistemäärä on 12 pistettä. 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. Ellei tehtävänannossa muuta sanota, käytä testauksessa merkitsevyystasona $\alpha = 0.05$. Menestystä tenttiin!

1. Tupakointi on monille taudeille tärkeä riskitekijä. Eräässä tutkimuksessa, jossa oli käytössä satunnaisotannalla poimittu otos ($n = 800$), kerättiin kyselylomakkeella tietoja tutkittavien tupakoinnista ja ruumiinpainosta. Lomakkeella olivat seuraavat kysymykset, joista on johdettu muuttujat `Smoker` (1. kysymys) ja `Weight` (2. kysymys).

1. Tupakoitteko päivittäin? (0=Ei, 1=Kyllä)
2. Mikä on painonne (yksikkö kg) tällä hetkellä?

Tässä on yhteenveto kyseisistä muuttujista

```
> table(dat$Smoker)

0   1
615 185
> sum(dat$Weight)
[1] 62086.6
> var(dat$Weight)
[1] 480.9695
> sd(dat$Weight)
[1] 21.93102
```

(a) Laske päivittäin tupakoivien suhteellisen osuuden piste-estimaatti ja sille 95 %:n ja 99 %:n luottamusvälit.

(b) Laske ruumiinpainon odotusarvon piste-estimaatti ja sille 95 %:n ja 99 %:n luottamusvälit.

2. Eräässä tutkimuksessa mitattiin ranteen luuntiheyttä (yksikkö g/cm^2) 50-75 -vuotiailta miehiltä ja naisilta ($n = 800$). Henkilöt oli poimittu satunnaisotannalla. Luuntiheyttä (`BMD`) mallinnettiin lineaarisella regressiolla, jossa olivat selittäjinä ikä (`Age`) ja sukupuoli (`Gender`). Alla kyseisen aineiston R-tuloste.

```
> head(bmddat)
        BMD Age Gender
1 0.3923660  66 Female
2 0.4635478  54   Male
3 0.6579392  50   Male
4 0.3142545  74   Male
5 0.4966346  59   Male
6 0.2436903  53 Female
> summary(bmddat)
      BMD               Age            Gender
Min.   :0.1006   Min.   :50.00   Length:774
```

```
1st Qu.:0.2764   1st Qu.:56.00   Class :character
Median :0.3671   Median :62.00   Mode  :character
Mean   :0.3658   Mean   :62.46
3rd Qu.:0.4501   3rd Qu.:69.00
Max.   :0.7818   Max.   :75.00
> m1 <- lm(BMD ~ I(Age-50)+factor(Gender),data=bmddat)
> summary(m1)

Call:
lm(formula = BMD ~ I(Age - 50) + factor(Gender), data = bmddat)

Residuals:
Min       1Q   Median       3Q      Max
-0.26510 -0.06344 -0.00226  0.06326  0.34220

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.3746540  0.0076665   48.87   <2e-16 ***
I(Age - 50)        -0.0062585  0.0004547  -13.76   <2e-16 ***
factor(Gender)Male  0.1239134  0.0069453   17.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09595 on 771 degrees of freedom
Multiple R-squared:  0.3957,Adjusted R-squared:  0.3942
F-statistic: 252.5 on 2 and 771 DF,  p-value: < 2.2e-16
```

(a) Kirjoita R-tulosteen perusteella malli, jonka oletetaan pätevän siinä populaatiossa, josta aineisto on poimittu.

(b) Kirjoita R-tulosteen perusteella estimoitu lauseke vasteen y odotusarvolle $E(\text{BMD}|\text{Age}, \text{Gender})$.

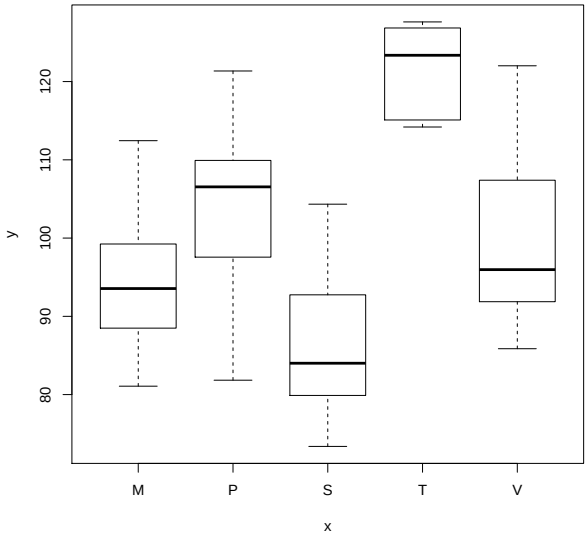
(c) Tulkitse muuttujan `Age` kerroin.

(d) Tulkitse muuttujan `Gender` kerroin.

(e) Tulkitse vakiotermin estimaatti.

(f) Laske mallin perusteella ennuste 60-vuotiaalle naiselle ja 70-vuotiaalle miehelle.

3. Tutkimuksessa selvitettiin eri ohralajien tuottavuutta. Peltoalue jaettiin 50 ruutuun ja kuhinkin kylvettiin satunnaisesti yhtä viidestä lajikkeesta (lajikkeet `M, P, S, T, V`) niin, että kutakin lajiketta kasvatettiin kymmenessä ruudussa. Kasvukauden lopussa mitattiin ohrasato (`tuotos`, kg/ruutu). Kuva saadusta aineistosta on alla. Testaa sopivaa menetelmää käyttäen, onko lajikkeiden välillä eroa tuotoksessa ja jos on, mitkä lajikkeet poikkeavat toisistaan ja miten.




```
> malli<-lm(tuotos~lajike,data=ohra)
> aggregate(ohra$tuotos,list(ohra$lajike),mean)
  Group.1         x
1       M  95.17728
2       P 104.73791
3       S  86.66385
4       T 121.59028
5       V  99.09158
> anova(malli)
Analysis of Variance Table

Response: tuotos
          Df Sum Sq Mean Sq F value    Pr(>F)
lajike     4 6799.8 1699.95  17.821 7.818e-09 ***
Residuals 45 4292.5   95.39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(malli)

Call:
lm(formula = tuotos ~ lajike, data = ohra)

Residuals:
     Min       1Q   Median       3Q      Max
-22.8995  -6.6673  -0.9085   5.2409  22.9251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   95.177      3.089  30.816  < 2e-16 ***
lajikeP        9.561      4.368   2.189   0.0338 *
lajikeS       -8.513      4.368  -1.949   0.0575 .
lajikeT       26.413      4.368   6.047 2.66e-07 ***
lajikeV        3.914      4.368   0.896   0.3749
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.767 on 45 degrees of freedom
Multiple R-squared:  0.613,Adjusted R-squared:  0.5786
F-statistic: 17.82 on 4 and 45 DF,  p-value: 7.818e-09

> pairwise.t.test(ohra$tuotos,ohra$lajike,p.adj="none")
```

```
Pairwise comparisons using t tests with pooled SD

data:  ohra$tuotos and ohra$lajike

  M       P       S       T
P 0.03384 -       -       -
S 0.05753 0.00015 -       -
T 2.7e-07 0.00036 3.5e-10 -
V 0.37493 0.20271 0.00666 5.6e-06

P value adjustment method: none
> TukeyHSD(aov(tuotos~lajike,data=ohra))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = tuotos ~ lajike, data = ohra)

$lajike
          diff          lwr        upr     p adj
P-M   9.560628  -2.85034018  21.971597 0.2025365
S-M  -8.513432 -20.92440010   3.897537 0.3072179
T-M  26.413003  14.00203430  38.823971 0.0000026
V-M   3.914304  -8.49666496  16.325272 0.8968242
S-P -18.074060 -30.48502841  -5.663091 0.0013640
T-P  16.852374   4.44140599  29.263343 0.0031748
V-P  -5.646325 -18.05729327   6.764644 0.6969557
T-S  34.926434  22.51546591  47.337403 0.0000000
V-S  12.427735   0.01676665  24.838704 0.0495405
V-T -22.498699 -34.90966775 -10.087731 0.0000528
```

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) Ao. tuloste saatiin sovittamalla kaksisuuntaisen varianssianalyysin malli aineistoon, jossa vastemuuttuja `y` on suhdeasteikollinen, muuttuja `x1` voi saada arvon `A`, `B` tai `C` ja muuttuja `x2` arvon `YES` tai `NO`.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.00    0.28931  17.420  < 2e-16 ***
x1B             3.00    0.40914   7.278 1.46e-09 ***
x1C             2.50    0.40914   5.970 1.90e-07 ***
x2YES          -2.00    0.40914  -5.377 1.66e-06 ***
x1B:x2YES       0.00    0.57861   0.126    0.900
x1C:x2YES       1.00    0.57861   1.569    0.122
```

Kun `x1=B` ja `x2=YES`, muuttujan y populaariokeskiarvon (=odotusarvon) estimaatti on

i. 5

ii. 6

iii. 0

(b) Regressiomallin selitysasteeksi saatiin $R^2 = 0.14$.

i. Regressiorelaatio ei ole tilastollisesti merkitsevä.

ii. Regressiorelaatio on merkitsevä, koska selitysaste on suurempi kuin merkitsevyystaso ($\alpha = 0.05$).

iii. Vastemuuttujan kokonaisvaihtelusta 86% on sellaista, jota malli ei selitä.

(c) Puut valmistautuvat talven mm. väkevöittämällä solunesteitä talven lähestyessä. Siksi puut kestävät koviakin pakkasia talvisaikaan, kun taas pakkaset kesällä tai alkusyksyllä tappaisivat puun. Tutkija halusi selvittää, miten pakkaskestävyys kehittyy syksyn aikana. Kokeessa männyntaimia laitettiin pakastearkkuun viikoksi eri ajankohtina ja tutkittiin, selvisikö taimi pakastamiskäsittelystä. Analyysi tehtiin logistisella regressiomallilla, jossa selittäjänä oli pakastamisen ajankohta (päivää kokeen alusta; koe aloitettiin syyskuun alussa ja sitä jatkettiin 50 päivän ajan) ja vastemuutujana oli binaarinen muuttuja, joka kertoo kuoliko taimi kokeen seurauksena vai ei (kuolema on koodattu ykkösenä). Estimoiduksi malliksi saatiin

$$\log\left(\frac{p}{1-p}\right) = 19.76 - 0.75 aika\,.$$

Tarkastellaan taimien pakkaskestävyyttä syyskuun viimeisenä päivänä, eli kun $aika = 30$.

i. Syyskuun 30. päivänä pakastettu taimi kuolee todennäköisemmin kuin säilyy hengissä.

ii. Syyskuun 30. päivänä pakastetulla taimella vedonlyöntisuhde kuolemisen puolesta on 0.065.

iii. Syyskuun 30. päivänä pakastettu taimi kuolee todennäköisyydellä 0.065.

(d) Mikä seuraavista väittämistä on totta

i. Epäparamerisia testejä tulee käyttää aina kun populaation jakauma ei ole normaalinen.

ii. Parametrisia testejä voidaan käyttää isoissa otoksissa, vaikka populaation jakauma ei olekaan normaalinen.

iii. Koska epäparametriset testit eivät tee mitään oletusta populaation jakaumasta, niitä voidaan käyttää turvallisesti kaikissa tilanteissa.

(e) muuttujan y populaatiokeskiarvon (=odotusarvon) piste-estimaatiksi saatiin luku 15. 95% luottamusvälin alarajaksi saatiin 10 ja ylärajaksi 20.

i. Muuttujan y populaatiokeskiarvo on 15.

ii. Muuttujan y populaatiokeskiarvo on välillä $(10, 20)$.

iii. Muuttujan y populaatiokeskiarvo on aika varmasti välillä $(10, 20)$.

(f) χ^2 riippumattomuustestillä testattiin sukupuolen (`M`=mies ja `F`=nainen) ja puoluekannan yhteyttä USA:ssa ja saatiin seuraava tuloste:

```
> CrossTable(M,chisq=TRUE)
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
```

```
|         N / Table Total |
|-------------------------|

=====================================================
          party
gender    Democrat   Independent   Republican   Total
-----------------------------------------------------
F              762           327          468    1557
             4.835         0.169        8.084
             0.489         0.210        0.301   0.565
             0.612         0.578        0.495
             0.276         0.119        0.170
-----------------------------------------------------
M              484           239          477    1200
             6.273         0.220       10.489
             0.403         0.199        0.398   0.435
             0.388         0.422        0.505
             0.176         0.087        0.173
-----------------------------------------------------
Total         1246           566          945    2757
             0.452         0.205        0.343
=====================================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 30.07015      d.f. = 2      p = 2.95e-07
```

i. Sukupuolen ja puoluekannan väillä ei näyttäisi olevan yhteyttä.
ii. Miehet näyttävät suosivan demokreetteja.
iii. Naiset näyttävät suosivan demokraatteja.