Empiirinen projekti

Maria Seppänen

2023-02-21

Sisällys

1 Johdanto

Tämä on Tampereen Yliopiston kurssin Empiirinen tutkimus tutkimusraportti. Tarkoituksena on hankkia sopiva tutkimusaineisto ja ratkaista siitä rajattuja tutkimusongelmia tilastollisia tutkimusmenetelmiä apuna käyttäen ja raportoida tilastollisen tutkimuksen tulokset.

Luvussa kaksi on aineiston kuvailua numeerisin ja graafisin menetelmin, luvussa kolme aineiston analysointia tutkimusmenetelmillä varianssi- ja regressioanalyysi sekä t-testaus, ja lopuksi luvussa neljä tutkimuksen tulosten yhteenveto.

Tutkimuksen tarkoituksena on tarkastella, onko koulutuksen määrällä, iällä, työkokemuksella, toimialalla, ammatilla tai sukupuolella vaikutusta tuntipalkan suuruuteen.

2 Aineiston kuvailu

2.1 Aineiston alkuperä ja tutkimukseen käytetyt muuttujat

Tutkimuksessa käytetään Tampereen Yliopiston Moodlesta "Tutkimusmenetelmien työkalupakki", linkin "Tilastotieteen peruskurssien harjoitusaineistoja 2003–2018" alta löytyvää aineistoa "kunnat_1". Analyysissä käytetään aineiston seuraavia muuttujia

- koulutus (koulutuksen määrä vuosina)
- sukupuo (sukupuoli)
- tyokok (tykokemus vuosina)
- palkka (palkka tuntipalkkana dollaria/tunnissa)
- ika (ikä vuosina)
- toimiala
- ammatti.

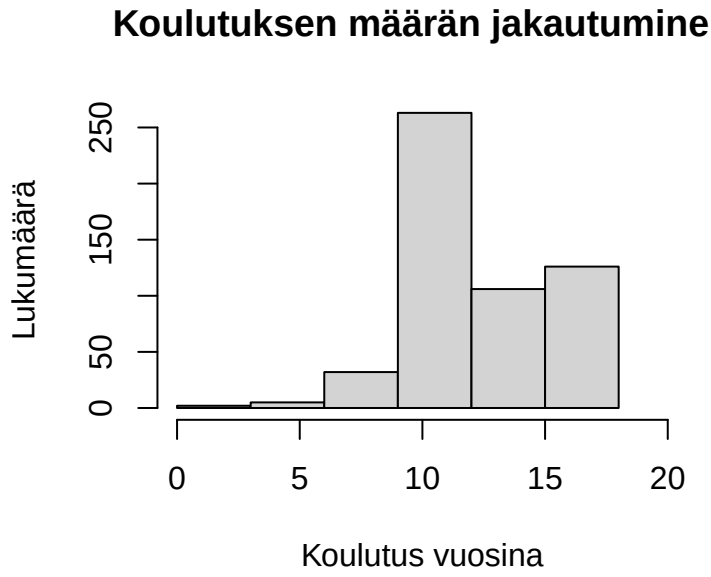
Havaintoyksikköjä on 534.

2.1.1 Muuttujien kuvailua

Tarkastellaan aineiston numeeristen muuttujien ääriarvoja, mediaania, keskiarvoa sekä 25 % ja 75 % kvartiileja. Numeerisia muuttujia ovat koulutus, työkokemus ja ikä vuosina sekä tuntipalkka.

Koulutuksen määrä vaihtelee kahden ja 18 vuoden välillä, keskiarvo on noin 13 vuotta ja mediaani 12 vuotta. Vanhempaa väestöä on hieman enemmän kuin nuorempaa. Ensimmäinen kvartiili on 12 vuotta ja kolmas kvartiili 15 vuotta. Koulutuksen määrä on jakautunut huomattavasti enemmän 10 vuoden yläpuolelle.

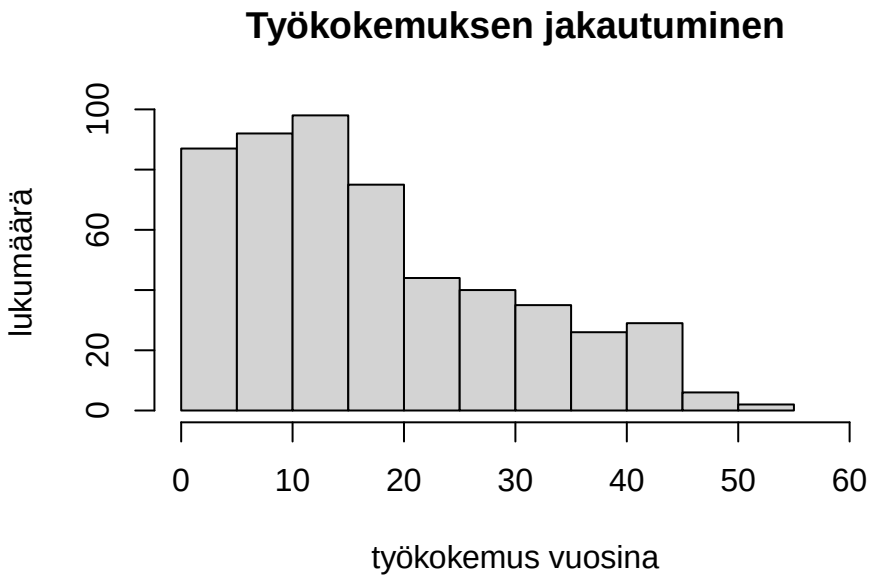
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   12.00   12.00   13.02   15.00   18.00
```



Työkokemus on jakautunut nollan ja 55 vuoden välille. Keskiarvo on 17,82 vuotta ja mediaani 15 vuotta. Ensimmäinen kvartiili on 8 vuotta ja toinen kvartiili 26 vuotta. Otoksessa on enemmän pienemmän työkokemuksen omaavia, kuin suuren työkokemuksen omaavia, mikä näkyy myös siinä, että mediaani on pienempi kuin keskiarvo. /newpage

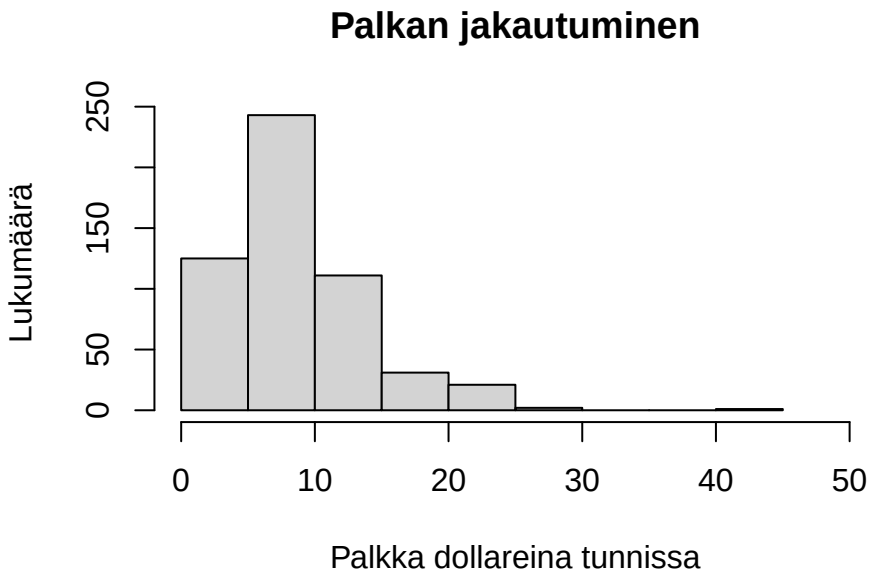
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    0.00    8.00   15.00   17.82   26.00   55.00
```



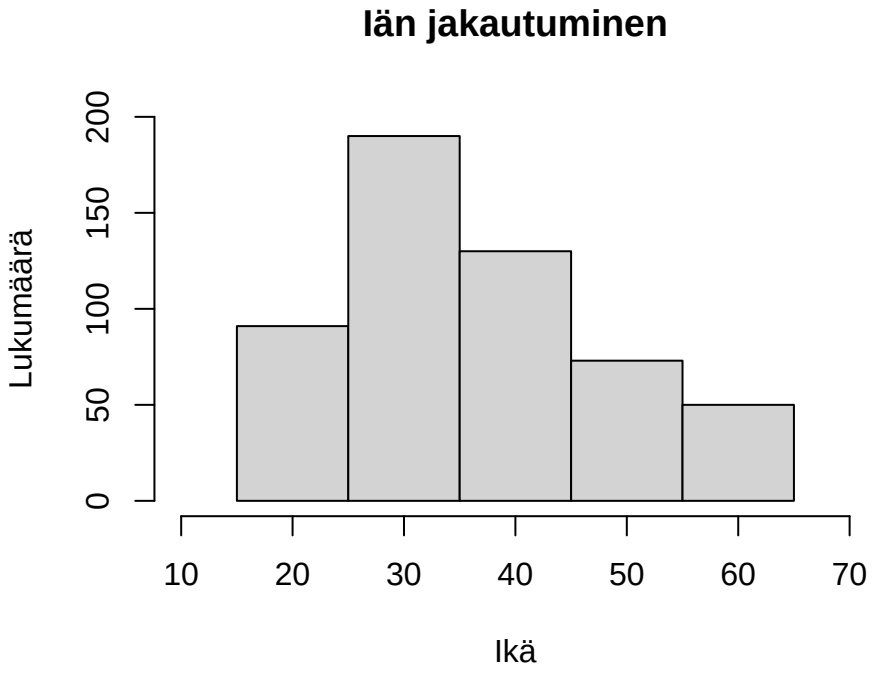
Aineiston pienin tuntipalkka on 1 $/h, kun taas suurin 44,5 $/h.Ensimmäinen kvartiili on noin 5 $/h ja kolmas noin 11 \6/h. Keskiarvo asettuu noin yhdeksään dollariin per tunti ja mediaani on 7,78 $/h. Histogrammista nähdään, että palkka on jakautunut hyvin vahvasti 25 $/h alle ja aineistosta löytyy vain muutama tätä korkeampi tuntipalkka.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.250   7.780   9.024  11.250  44.500
```



/newpage Ikä on jakautunut 18 ja 64 välille, ja sen mediaani on 35 vuotta. Keskiarvo on 36,83, ensimmäinen kvartiili 28 vuotta ja kolmas 44 vuotta. Pylväsdiagrammista näkee tarkemmin, että eniten on ikäluokkia otoksessa on väliltä 25-35 vuotta.

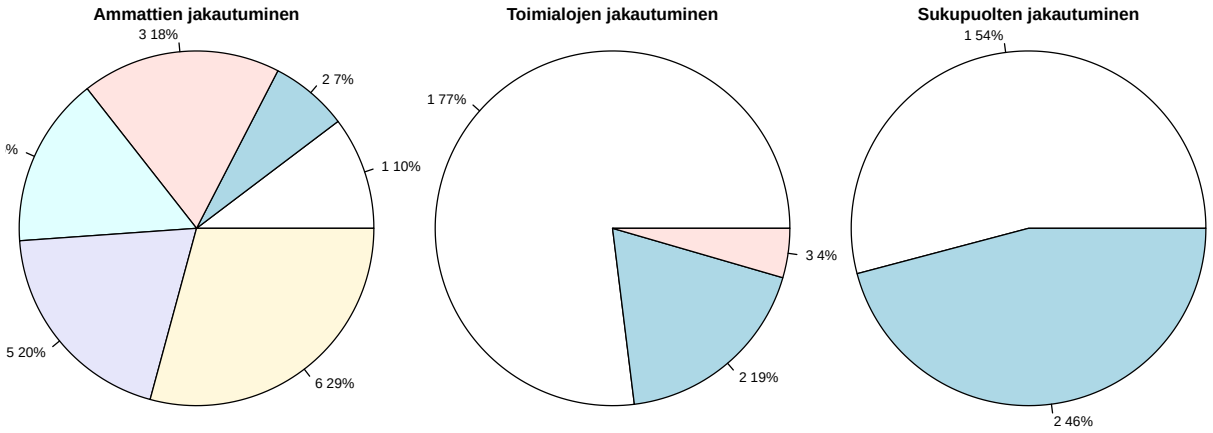
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   35.00   36.83   44.00   64.00
```



Kategorisia muuttujia tutkimuksessa on sukupuoli, toimiala, ammatti. Jokaisessa on vain numeroarvoja eri vaihtoehdoille, mutta itse vaihtoehtoja ei ole määritelty. Tarkastellaan, miten nämä ovat jakautuneet havaintoyksiköittäin.

Eri ammatteja on 6 kappaletta ja toimialoja kolme. Eniten on ammattia numero kuusi, jota on 156. Vähiten on numeroa kaksi, jota on vain 38 kappaletta. Toimialoissa huomattavasti eniten on numeroa nolla, jota on 411 ja vähiten on numeroa kaksi, jota on vain 24 kappaletta ja jäljelle jääneet 99 kappaletta ovat numeroa kolme.

Aineistossa sukupuolet ovat jakautuneet hyvin tasaisesti - numeroa yksi on 289 kappaletta ja numeroa kaksi 245 kappaletta.



3 Aineiston analysointi

Tutkimusongelmana on tarkastella, mitkä tekijät vaikuttavat henkilön tuntipalkkaan sekä onko sukupuolten välillä eroa ammateissa. Tutkimusongelmien tutkimiseen käytetään kolmea menetelmää:

- regressioanalyysi, jolla tutkitaan vaikuttavatko ikä, koulutuksen määrä tai työkokemus tuntipalkkaan
- t-testaus, jolla tutkitaan onko ammateissa eroa sukupuolten välillä
- varianssianalyysi, jolla tutkitaan onko tuntipalkassa eroa eri toimialojen välillä.

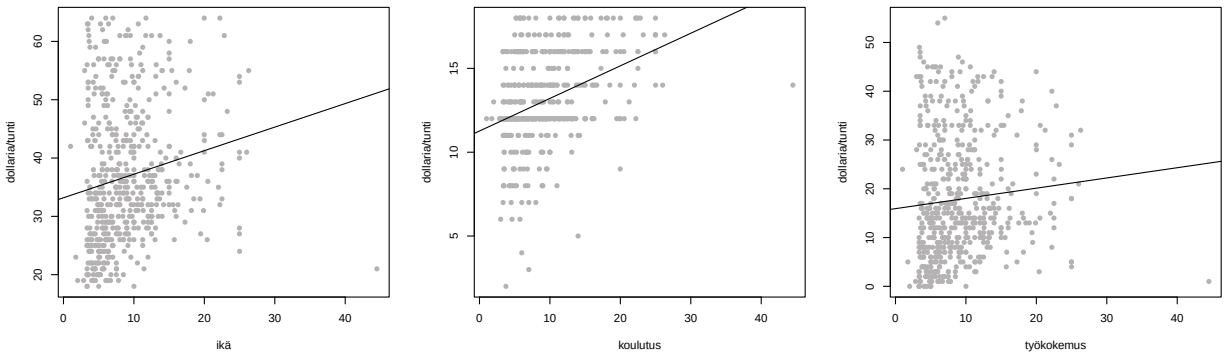
3.1 Regressioanalyysi

Tutkitaan, vaikuttaako koulutus, ikä tai työkokemus tuntipalkkaan käyttäen apuna lineaarista regressiomallia. Lineaarisella regressioanalyysillä voi havaita, vaikuttaako selittävät muuttujat (koulutus, ikä, työkokemus) selitettävän muuttujan (palkka) odotusarvoon.

Lineaarisen regressiomallin yhtälö löytyy alta. Yhtälössä y on palkka, $\beta_i, i = 0, 1, 2, 3$ ovat regressiokertoimia, x_1 on ikä, x_2 on koulutus ja x_3 on työkokemus. Tarkoituksena on selvittää regressiokertoimien arvot.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Tarkastellaan aineistosta selittäville muuttujille erikseen piirrettyjä regressiosuoria palkan suhteen. Pelkästään katsomalla nähdään, että kaikki suorat ovat nousevia. Suurin lineaarinen korrelaatio vaikuttaisi olevan tuntipalkan suuruudella ja koulutuksen määrällä ja pienin tuntipalkalla ja työkokemuksella.



3.1.1 Palkan regressiomalli

Tutkitaan muuttujien välisiä yhteyksiä tarkemmin laskemalla regressiomallin regressiokertoimet ja tarkastelemalla selittävien muuttujien merkitsevyyttä. Alla on tutkittava regressiomalli.

```
##
## Call:
## lm(formula = palkka ~ ika + koulutus + tyokok)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.351 -2.857 -0.599  1.994 36.336
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.76987    7.04271  -0.677    0.499
## ika         -0.02241    1.15475  -0.019    0.985
## koulutus     0.94833    1.15524   0.821    0.412
```

```
## tyokok       0.12756    1.15571   0.110    0.912
##
## Residual standard error: 4.604 on 530 degrees of freedom
## Multiple R-squared:  0.202,  Adjusted R-squared:  0.1975
## F-statistic: 44.73 on 3 and 530 DF,  p-value: < 2.2e-16
```

Regressiokertoimiksi saatiin $\beta_0 = -4,77$, $\beta_1 = -0,22$, $\beta_2 = 0,95$, $\beta_3 = 0,13$. Kun nämä laitetaan alkuperäiseen regressiomalliin, saadaan sovitefunktio, jolla voi halutessaan laskea sovitearvon, eli estimaatin selitettävälle muuttujalle (tässä tapauksessa palkalle). Sovitefunktio on seuraavanlainen:

$$\hat{\mu} = -4,77 - 0,22x_1 + 0,95x_2 + 0,13x_3.$$

3.1.2 Selittävien muuttujien vaikutus palkkaan

Testataan 5% merkitsevyystasolla hypoteeseja

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0,$$
$$H_1 : \text{ainakin yksi } \beta_i \neq 0, i = 1, 2, 3.$$

Kun nollahypoteesi on voimassa, selittävät muuttujat eivät tilastollisesti eroa nollasta ja täten vaikuta palkan suuruuteen. Kun vaihtoehtoinen hypoteesi on voimassa, selittävillä muuttujilla on vaikutus palkkaan. Yllä olevasta taulukosta saadaan F-testisuureeksi 44,73, jota vastaavaksi p-arvoksi $< 2,2 \times 10^{-16}$, joka on hyvin pieni. Tällöin nollahypoteesi H_0 hylätään ja vaihtoehtoinen hypoteesi H_1 astuu voimaan. Selittävillä muuttujilla on siis tilastollinen vaikutus tuntipalkkaan. Tutkitaan samalla, mitkä muuttujista ovat merkitseviä.

```
## Analysis of Variance Table
##
## Response: palkka
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## ika         1   440.8  440.84  20.8004 6.338e-06 ***
## koulutus    1  2402.7 2402.75 113.3689 < 2.2e-16 ***
## tyokok      1     0.3    0.26   0.0122    0.9122
## Residuals 530 11232.8   21.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yllä olevasta taulukosta nähdään, että muuttujien F-testisuureet ja niitä vastaavat p-arvot ovat

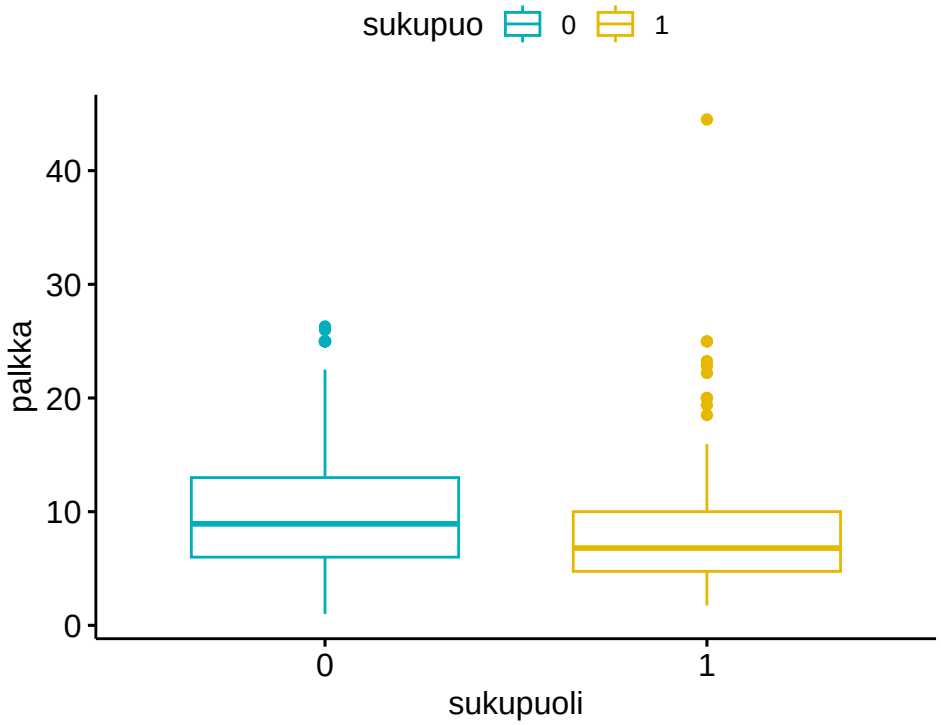
$$F_{ikä} = 20,8004;\ p_{ikä} = 0.000006338 < 0,05$$
$$F_{koulutus} = 113,3689;\ p_{koulutus} < 2.2 \times 10^{16} < 0,05$$
$$F_{työkokemus} = 0,0122;\ p_{työkokemus} = 0,9122 > 0,05.$$

Iän sekä koulutuksen p-arvot ovat merkitsevyystasoa 0,05 pienemmät, eli ne vaikuttavat tilastollisesti merkittävästi 5% merkitsevyystasolla tuntipalkan suuruuteen, kun taas työkokemuksen p-arvo ylittää merkitsevyystason 0.05 eikä täten vaikuta tilastollisesti merkittävästi tuntipalkan suuruuteen.

3.2 Welchin t-testi

Tarkastellaan, onko sukupuolella merkitystä palkkaan t-testiä apuna käyttäen.

Alla on boxplot molempien sukupuolten jakaumista palkkojen suhteen. Näennäisesti voisi sanoa, että eroa näidenkeskiarvossa voisi olla. Tutkitaan asiaa itse t-testin ja -testisuureen avulla.



Oletetaan, että sukupuolten osapopulaatiot noudattavat normaalijakaumaa ja käytetään testaamiseen Welchin t-testiä, jolloin varianssien ei tarvitse olla samat. Testataan hypoteeseja

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2,$$

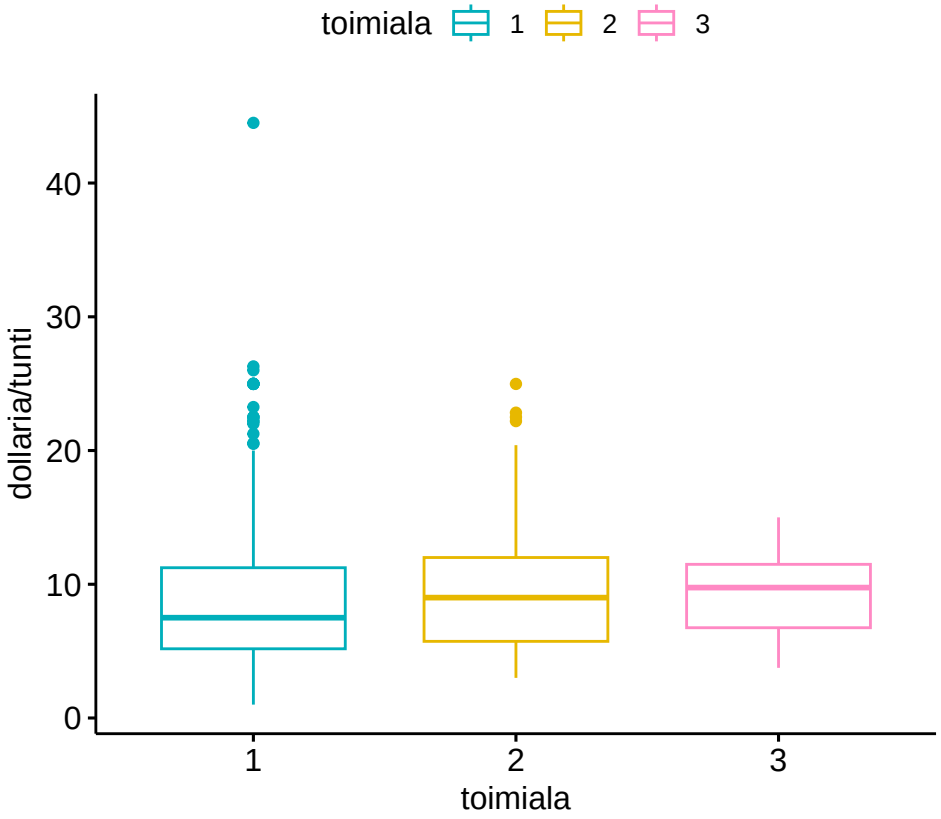
missä μ_1 on sukupuolen "1" keskiarvo ja μ_2 sukupuolen "2" keskiarvo.

```
##
##  Welch Two Sample t-test
##
## data:  palkka by sukupuo
## t = 4.8853, df = 530.55, p-value = 1.369e-06
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  1.265164 2.966949
## sample estimates:
## mean in group 0 mean in group 1
##        9.994913        7.878857
```

T-testisuure on 4,8853 ja 95%-luottamusväli [1,265, 2,967]. Testisuure ei asetu luottamusväliin ja p-arvo on $1,369 \times 10^{-6} < 0{,}05$. Nollahypoteesi hylätään ja vaihtoehtoinen hypoteesi astuu voimaan, mikä tarkoittaa, että sukupuolten välillä on 5%-merkitsevyystasolla eroa palkassa.

3.3 Varianssianalyysi

Tutkitaan varianssianalyysillä, vaikuttaako toimiala tuntipalkkaan. Tarkastellaan aineiston kolmen eri toimialan palkkojen jakautumista boxplotilla. Keskiarvot ovat kaikki hyvin lähellä toisiaan, kuten myös ensimmäiset ja kolmannet kvantiilit. Toimialoilla "1" ja "2" neljäs kvantiili on korkeampi kuin toimialalla "3". Oletetaan, että toimialojen tuntipalkat ovat normaalisti jakautuneita.



3.3.1 Varianssien yhtäsuuruudet

Tarkistetaan Levenen testillä, eroavatko varianssit tilastolliseesti merkitsevästi toisistaan. Nollahypoteesi on, että varianssit ovat yhtä suuria ja vaihtoehtoinen hypoteesi, että ne ovat erisuuria.

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.8924 0.4103
##       531
```

Levenen testin F-arvo on 0,8924 ja sitä vastaaava p-arvo 0,4103, joka on suurempi kuin merkitsevyystaso 0,05. Nollahypoteesi jää voimaan ja varianssit eivät eroa toisistaan 5% merkitsevyystasolla. Jatketaan siis itse varianssianalyysiin.

3.3.2 Toimialan vaikutus palkkaan

Tutkitaan varianssianalyysillä eroa eri toimialojen tuntipalkkojen keskiarvoissa ja testataan hypoteeseilla

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

$$H_1 : \text{keskiarvot ovat erisuuria,}$$

missä μ_1 on toimialan "1" keskiarvo palkassa, μ_2 toimialan "2" ja μ_3 toimialan "3" keskiarvo.

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(toimiala)   2     44   21.84   0.826  0.438
## Residuals        531  14033   26.43
```

Yllä olevasta taulukosta nähdään, että muuttujan toimiala F-testisuure on 0,826 ja p-arvo 0,438 $>$ 0.05. P-arvo on suurempi kuin merkitsevyystaso, eli nollahypoteesi jää voimaan ja eri toimialojen palkkojen keskiarvot eivät eroa toisistaan tilastollisesti 5% merkitsevyystasolla.

4 Liitteet

```
library(readxl)
library(ggpubr)
library(MASS)
library(dplyr)
library(reshape2)
library(ggplot2)
library(car)

palkat <- read_excel("filepath")
attach(palkat)

sukupuo <- factor(sukupuo)
levels(sukupuo) <- c("0", "1")

ammatti <- factor(ammatti)
levels(ammatti) <- c("1", "2", "3", "4", "5", "6")


toimiala <- factor(toimiala)
levels(toimiala) <- c("1", "2", "3")

# Koulutuksen tunnusluvut.
summary(koulutus)
hist(koulutus, seq(0, 18, 3), xlim = c(0,20), main = "Koulutuksen määrän jakautuminen",
      xlab= "Koulutus vuosina", ylab="Lukumäärä")

# Työkokemuksen tunnusluvut.
summary(tyokok)
hist(tyokok, c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55), xlim = c(0,60),
    main = "Työkokemuksen jakautuminen", xlab= "työkokemus vuosina", ylab="lukumäärä")

# Palkan tunnusluvut.
summary(palkka)
hist(palkka, c(0, 5, 10, 15, 20, 25, 30, 35, 40, 45), xlim = c(0,50),
      main = "Palkan jakautuminen", xlab= "Palkka dollareina tunnissa", ylab="Lukumäärä")

# Iän tunnusluvut.
summary(ika)
hist(ika, c(15, 25, 35, 45, 55, 65), ylim = c(0, 200), xlim = c(10,70),
      main="Iän jakautuminen", xlab = "Ikä", ylab = "Lukumäärä")

lbls <- seq(1:6)
pct <- round(table(ammatti)/sum(table(ammatti))*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")

lbls2 <- seq(1:6)
pct2 <- round(table(toimiala)/sum(table(toimiala))*100)
lbls2 <- paste(lbls2, pct2)
lbls2 <- paste(lbls2,"%",sep="")
```

```
lbls3 <- seq(1:6)
pct3 <- round(table(sukupuo)/sum(table(sukupuo))*100)
lbls3 <- paste(lbls3, pct3)
lbls3 <- paste(lbls3,"%", sep="")
par(mar=c(0.5,0,2,2))
par(mfrow=c(1,3))

pie(table(ammatti), labels = lbls, radius = 1, main="Ammattien jakautuminen")

pie(table(toimiala), labels = lbls2, radius =1, main="Toimialojen jakautuminen")

pie(table(sukupuo), labels=lbls3,radius =1, main="Sukupuolten jakautuminen")

piirräReg <- function(y, x, ...) {
    plot(x, y, ...)
    abline(lm(y ~ x))
}

par(mfrow=c(1,3))
piirräReg(ika, palkka, xlab= "ikä", ylab= "dollaria/tunti", pch=16, col = "#B4B2B3")
piirräReg(koulutus, palkka, ylab= "dollaria/tunti", xlab= "koulutus", pch=16, col = "#B4B2B3")
piirräReg(tyokok, palkka, ylab="dollaria/tunti", xlab="työkokemus", pch=16, col="#B4B2B3")

palkka.lm <- lm(palkka~ika+koulutus+tyokok)
summary(palkka.lm)

anova(palkka.lm)

amm_suk <- cbind.data.frame(as.double(ammatti), sukupuo)

ggboxplot(amm_suk, x = "sukupuo", y = "as.double(ammatti)", color = "sukupuo",
      palette = c("#00AFBB", "#E7B800"),
          order = c("0", "1"),
          ylab = "ammatti", xlab = "sukupuoli")

ggboxplot(pal_suk, x = "sukupuo", y = "palkka", color = "sukupuo", palette = c("#00AFBB", "#E7B800"),
          order = c("0", "1"),
          ylab = "palkka", xlab = "sukupuoli")

t.test(palkka~sukupuo, var.equal = FALSE)

pal_toim <- cbind.data.frame(palkka, toimiala)


ggboxplot(pal_toim, x = "toimiala", y = "palkka", color = "toimiala", palette = c("#00AFBB", "#E7B800",
          order = c("1", "2", "3"),
          ylab = "dollaria/tunti", xlab = "toimiala")

leveneTest(palkka~toimiala)

pal_toim.aov <- aov(palkka~factor(toimiala), data=palkat)
summary(pal_toim.aov)
```

```
```