TILM 3510 Tilastotieteen peruskurssi b
Lasse Rintakumpu, 63555
Harjoitustyö

TILM 3510 -kurssin harjoitustyö

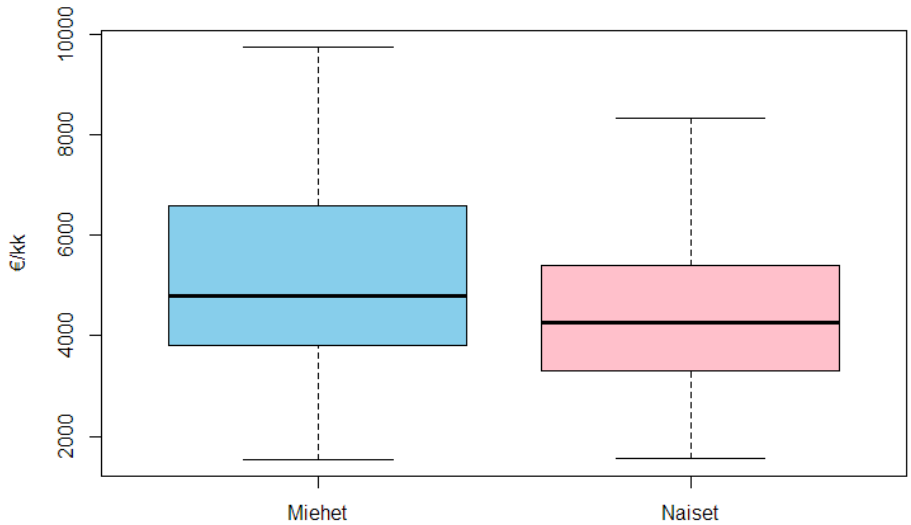
Tarkastellaan havaintoaineiston (`talous.csv`) perusteella sukupuolten välisiä eroja nettotuloissa sekä tyytyväisyydessä taloudelliseen tilanteeseen.

Nettotulot on aineistossa esitetty muodossa euroa / kuukausi. Taloudellista tyytyväisyyttä on mitattu neliportaisella ordinaaliasteikolla, jossa arvot 1 sekä 2 vastaavat tyytyväisyyttä henkilökohtaiseen taloudelliseen tilanteeseen (1 = erittäin tyytyväinen, 2 = melko tyytyväinen) ja arvot 3 sekä 4 tyytymättömyyttä taloudelliseen tilanteeseen (3 = melko tyytymätön, 4 = erittäin tyytymätön). Aineiston tilastoyksiköt sisältävät miehiä $n_m = 89$ ja naisia $n_n = 91$ kappaletta.

Kuvaillaan ensin havaintoaineiston jakaumia graafisesti, jonka jälkeen testataan havaintoaineistoissa mahdollisesti esiintyvien erojen tilastollista merkitsevyyttä.

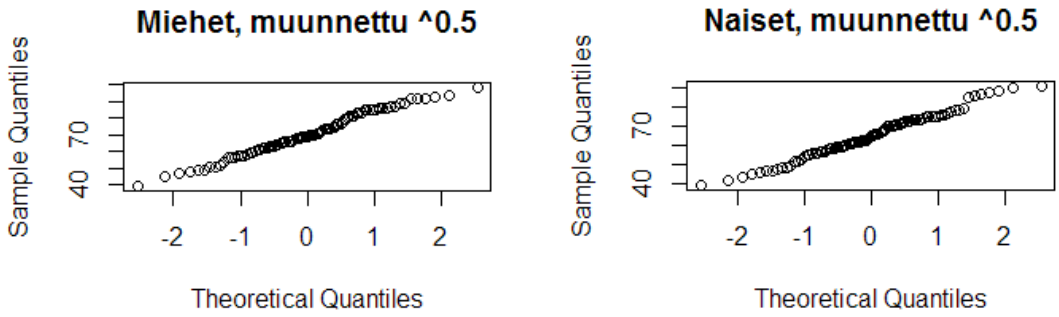
1. Sukupuolten väliset erot nettotuloissa

Kuvaillaan havaintoaineistoa Tukeyn laatikko-janakuviolla.



Laatikko-janakuvion perusteella aineistossa ei ole selkeästi poikkeavia havaintoja. Aineistot ovat kuitenkin oikealle vinoja sekä huipukkaita, kuten aineistolle lasketut vinous- ja huipukkuuskertoimet osoittavat. Miehet: $g_{1m} \approx 0.35$, $g_{2m} \approx 2.34$. Naiset: $g_{1n} \approx 0.46$, $g_{2n} \approx 2.34$.

Sovelletaan molempiin aineistoihin neliöjuurimuunnosta (muunnos on mahdollinen, sillä kaikki havainnot positiivisia kokonaislukuja). Muunnetuista huipukkuus- ja vinouskertoimista $g_{1m} \approx -0.01$, $g_{2m} \approx 2.35$ ja $g_{1n} \approx 0.005$, $g_{2n} \approx 2.58$ huomataan, että muunnos käytännössä eliminoi aineiston vinouden. Aineiston normaalikvantiilikuvaajat tukevat niin ikään oletusta normaalijakautuneisuudesta.



Jatketaan analyysia neliöjuurimuunnetuilla arvoilla.

Muuntamattomasta aineistosta lasketut otoskeskiarvot (oletusarvon piste-estimaatit) $\bar{x}_m \approx 5119$ ja $\bar{x}_n \approx 4392$ viittaavat tuloeroihin sukupuolten välillä. Testataan onko miesten ja naisten nettotuloissa havaittu ero tilastollisesti merkitsevä. Valitaan hypoteesipariksi

H_0: $\mu_m - \mu_n = 0$
H_v: $\mu_m - \mu_n \neq 0$.

Testataan hypoteesia tasolla 0.05. Populaatioiden varianssi ei ole tunnettu, joten käytetään testaamiseen t-testiä. Saadaan tulokseksi

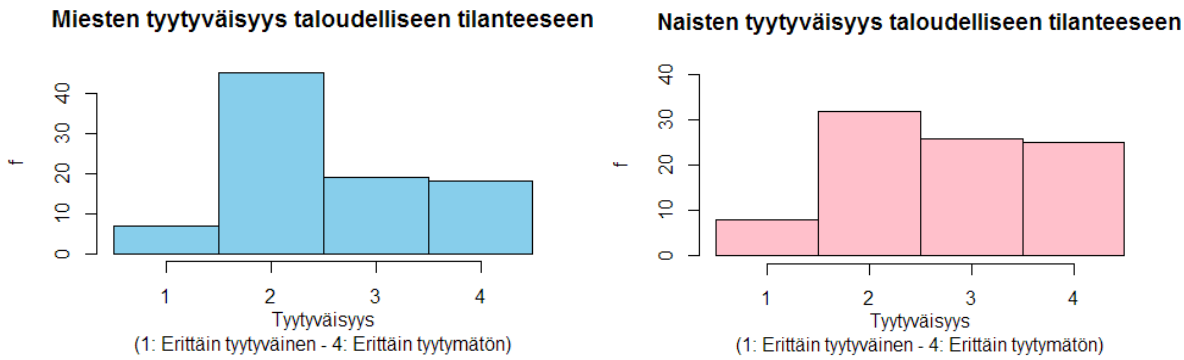
```
data: nettotulot_m_muunnettu and nettotulot_n_muunnettu
t = 2.7782, df = 174.866, p-value = 0.006064
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1.480589 8.744629
sample estimates:
mean of x mean of y
0.36786  65.25525.
```

Testin perusteella nollahypoteesi voidaan hylätä riskitasolla 0.05. Pienin riskitaso, jolla nollahypoteesi voitaisiin hylätä on 0.006.

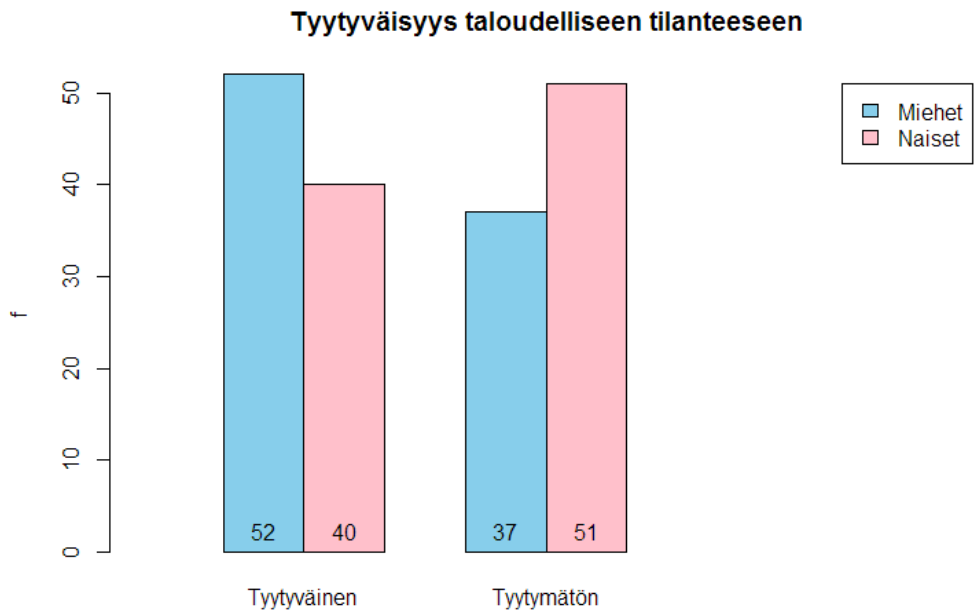
Aineiston perusteella miesten ja naisten nettopalkoissa on tilastollisesti merkitsevä ero miesten hyväksi, kun oletetaan, että sekä miesten että naisten nettotulot ovat jakautuneet normaalisti ja että havainnot ovat toisistaan riippumattomia.

2. Sukupuolten väliset erot tyytyväisyydessä taloudelliseen tilanteeseen

Käytetään aineiston kuvailuun histogrammia sekä ryhmäpylväskuviota.



Kuvaajien perusteella miehet vaikuttaisivat naisia tyytyväisemmiltä taloudelliseen tilanteeseensa. Ero näkyy vielä selkeämmin, kun ryhmitellään miesten ja naisten tyytyväisyys niin, että arvot 1 ja 2 vastaavat tyytyväistä ja arvot 3 ja 4 tyytymätöntä.



Tyytyväisten osuuden piste-estimaateiksi saamme miehille $\hat{p}_m \approx 0.58$ ja naisille $\hat{p}_n \approx 0.44$. Testataan aineiston pohjalta onko miesten ja naisten tyytyväisyydessä miesten hyväksi havaittava ero tilastollisesti merkitsevää. Valitaan hypoteesipariksi

H_0: $p_m - p_n = 0$
H_0: $p_m - p_n > 0$.

Testataan kahden suhteellisen osuuden erotusta tasolla 0.05. Saadaan tulokseksi

```
data: tyytyvaisyys
X-squared = 3.2138, df = 1, p-value = 0.03651
alternative hypothesis: greater
95 percent confidence interval:
0.0123195 1.0000000
sample estimates:
prop 1    prop 2
0.5842697 0.4395604.
```

Testin perusteella nollahypoteesi voidaan hylätä tasolla 0.05. Pienin arvo, jolla nollahypoteesi voitaisiin hylätä on 0.037. Tilastoyksikköjen määrää testissä voidaan pitää riittävänä. Aineiston pohjalta miehet vaikuttavat naisia tyytyväisemmiltä taloudelliseen tilanteeseensa, olettaen, että otokset ovat toisistaan riippumattomia.

Liitteet

Liitteenä analyysissa käytetty R-koodi.

```
# Funktio kirjastojen asentamiselle / lataamiselle

lataa_kirjasto <- function(kirjasto) {
if(kirjasto %in% rownames(installed.packages()) == FALSE)
  {install.packages(kirjasto)}
library(kirjasto, character.only = TRUE)
}

# Ladataan/asennetaan käytetyt kirjastot

lapply(c("moments","car", "alr3"), lataa_kirjasto)

# Ladataan havaintoaineisto

talous <-
read.csv("https://raw.github.com/rintakumpu/tilm3510/master/talous.csv",
```

```
sep=";", dec=",");

# Ollaan kiinnostuttu nettotuloista (1.) sekä taloudellisesta tilanteesta
(2.)

miehet <- subset(talous, supu==1)
naiset <- subset(talous, supu==2)

# Erotetaan havaintoaineistosta sukupuolittain nettotulot

nettotulot_miehet <- miehet$oma_tulo
nettotulot_naiset <- naiset$oma_tulo

# sekä tyytyväisyys taloudelliseen tilanteeseen
# Ryhmitellään tyytyvaisyys niin,
# että 1--2 => Tyytyväinen
# 3--4 => Tyytymätön

tyytyvaisyys <- matrix(nrow=2,ncol=2)
rownames(tyytyvaisyys) <- c("Mies", "Nainen")
colnames(tyytyvaisyys) <- c("Tyytyväinen", "Tyytymätön")
tyytyvaisyys[1,1] <- sum(miehet$taltyyt==1 | miehet$taltyyt==2)
tyytyvaisyys[1,2] <- sum(miehet$taltyyt==3 | miehet$taltyyt==4)
tyytyvaisyys[2,1] <- sum(naiset$taltyyt==1 | naiset$taltyyt==2)
tyytyvaisyys[2,2] <- sum(naiset$taltyyt==3 | naiset$taltyyt==4)

##################
# 1. Nettotulot #
##################

# Käytetään normaalikvantiilikuvaajaa sekä laatikko-janakuviota

boxplot(nettotulot_miehet, nettotulot_naiset, col = c("skyblue", "pink"),
xaxl="n", ylab = "€/kk")
axis(1, at = c(1,2), labels = c("Miehet", "Naiset"))

# Maksimoidaan kuvaajien luotettavuus tallentamalla pdf-muodossa
# http://xkcd.com/1301/

pdf('boxplot_tulot.pdf')
dev.off()

skewness(nettotulot_miehet)
kurtosis(nettotulot_miehet)
# Miehet: g1, 0.3483119, g2: 2.343662
skewness(nettotulot_naiset)
kurtosis(nettotulot_naiset)
# Naiset: g2, 0.4554618, g2: 2.343662
```

```
# Haetaan potenssimuunnosta paremman normaalijakautuneisuuden löytämiseksi

ptm <- powerTransform(nettotulot_miehet)
# Estimated transformation parameters
# nettotulot_miehet
# 0.4683207

ptn <- powerTransform(nettotulot_naiset)
# Estimated transformation parameters
# nettotulot_naiset
# 0.4279722

# Sovelletaan molempiin aineistoihin neliöjuurimuunnosta

nettotulot_m_muunnettu <- nettotulot_miehet^0.5
nettotulot_n_muunnettu <- nettotulot_naiset^0.5

skewness(nettotulot_m_muunnettu)
kurtosis(nettotulot_m_muunnettu)
# Miehet: g1: -0.01, g2: 2.35
skewness(nettotulot_n_muunnettu)
kurtosis(nettotulot_n_muunnettu)
# Naiset: g1: 0.005, g2: 2.58

qqnorm(nettotulot_m_muunnettu, main = "Miehet, muunnettu ^0.5")
pdf('qqnorm_m_muunnettu.pdf')
dev.off()
qqnorm(nettotulot_n_muunnettu, main = "Naiset, muunnettu ^0.5")
pdf('qqnorm_n_muunnettu.pdf')
dev.off()

# Otoskeskiarvot viittaavat tuloeroihin sukupuolten välillä.

nettotulot_m_viiva <- mean(nettotulot_miehet) # 5118.921
nettotulot_n_viiva <- mean(nettotulot_naiset) # 4391.956

# Testataan onko miesten ja naisten nettotuloissa tilastollista eroa.

# Valitaan nollahypoteesiksi H0: mu_m - mu_n == 0 (populaatioiden
odotusarvot samat)
# ja vastahypoteesiksi       Hv: mu_m - mu_n != 0

t.test(nettotulot_m_muunnettu, nettotulot_n_muunnettu, alternative =
"two.sided", conf.level = 0.95)

# data:  nettotulot_m_muunnettu and nettotulot_n_muunnettu
# t = 2.7782, df = 174.866, p-value = 0.006064
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
```

```
#  1.480589 8.744629
# sample estimates:
#   mean of x mean of y
# 70.36786  65.25525

# => Hylätään nollahypoteesi riskitasolla 0.05. Testin p-arvosta näemme,
# että pienen riskitaso, jolla testi voitaisiin hylätä on 0.006064.

################################################
# 2. Tyytyväisyys taloudelliseen tilanteeseen #
################################################

# Käytetään aineiston kuvailuun histogrammia sekä ryhmäpylväskuviota.

tyytyvaisyys_m <- hist(miehet$taltyyt, 0:4, breaks =
c(0.5,1.5,2.5,3.5,4.5))
plot(tyytyvaisyys_m,
     xlab="Tyytyväisyys\n(1: Erittäin tyytyväinen - 4: Erittäin
tyytymätön)",
     ylab="f",
     main="Miesten tyytyväisyys taloudelliseen tilanteeseen",
     col="skyblue")

pdf('tyytyvaisyys_mies.pdf')
dev.off()

tyytyvaisyys_n <- hist(naiset$taltyyt, 0:4, breaks =
c(0.5,1.5,2.5,3.5,4.5))
plot(tyytyvaisyys_n,
     xlab="Tyytyväisyys\n(1: Erittäin tyytyväinen - 4: Erittäin
tyytymätön)",
     ylab="f",
     main="Naisten tyytyväisyys taloudelliseen tilanteeseen",
     col="pink", ylim = c(0,40))

pdf('tyytyvaisyys_nainen.pdf')
dev.off()

tyytyvaisyys_bp <- barplot(tyytyvaisyys, beside = TRUE,
        main=c("Tyytyväisyys taloudelliseen tilanteeseen"),
        col=c("skyblue","pink"),
        ylab="f", legend.text=c("Miehet","Naiset"), xlim = c(0,10))
text(tyytyvaisyys_bp, 0, round(tyytyvaisyys, 1),cex=1,pos=3)

pdf('tyytyvaisyys_yhteinen.pdf')
dev.off()

# Lasketaan estimaatit
```

```r
pm <- tyytyvaisyys[1,1] / sum(tyytyvaisyys[1,]) # 0.5842697
pn <- tyytyvaisyys[2,1] / sum(tyytyvaisyys[2,]) # 0.4395604

# Testataan suhteellisten osuuksien erotusta tasolla 0.05
# Valitaan vastahypoteesiksi tehtävänannon mukaan
# oletus, että miehet ovat naisia tyytyväisempiä.
# H0: pm-pn = 0 / Hv: pm-pn > 0

prop.test(tyytyvaisyys, alternative = "greater", conf.level=0.95)

# data:  tyytyvaisyys
# X-squared = 3.2138, df = 1, p-value = 0.03651
# alternative hypothesis: greater
# 95 percent confidence interval:
#  0.0123195 1.0000000
# sample estimates:
#  prop 1    prop 2
# 0.5842697 0.4395604

# Pienin arvo, jolla nollahypoteesi voitaisiin
# hylätä on 0.03651, hylätään nollahypoteesi tasolla 0.05.
```