Tilastollinen päättely R-ohjelmistolla, harjoitustyön raportti

September 26, 2017

1

a) Puuttuvat arvot voidaan huomioida käyttämällä tunnuslukuja laskiessa argumenttia na.rm = TRUE

c) Funktion yhteenveto tuloste:

```
[[1]]
[1] 1 4 5

[[2]]
[1] 35698

[[3]]
[1] 395

[[4]]
[1] 3.9336

[[5]]
[1] 0.6877915
```

Kysytyt tunnusluvut ovat siis:
Keskiarvo $\bar{y} = 3.9336$
Keskihajonta $s = 0.6877915$
Puuttuvia arvoja on 395.

Asiakkaat ovat keskimäärin melko tyytyväisiä kuljettajien toimintaan.

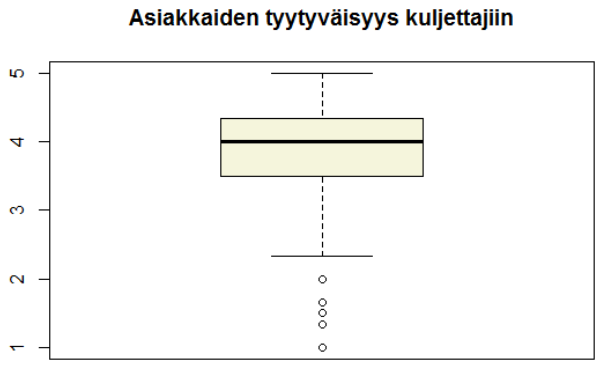

Figure 1: Summamuuttujan viiksikaavio, mediaani on paksun mustan viivan kohdalla

d) Funktion yhteenveto tuloste muuttujalle:

```
[[1]]
[1] 1.00 4.25 5.00

[[2]]
[1] 35999

[[3]]
[1] 94

[[4]]
[1] 4.164486

[[5]]
[1] 0.5253437
```

Mediaani on 4.25 ja keskiarvo $\bar{y} = 4.164486$
Asiakkaat ovat näiden perusteella tyytyväisiä HSL:n palveluihin.

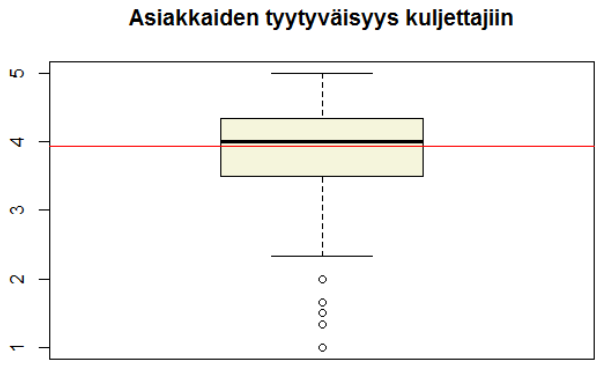

Figure 2: Keskiarvo punaisen viivan kohdalla

2

b) Muuttujien LIIKENNEMUOTO ja K1A4 ristiintaulukko:

	Lahijuna	Metro
1	40	3
2	145	13
3	527	72
4	1159	579
5	443	632

Sarakeprosentit laskettuna:

	Lahijuna	Metro
1	0.02	0.00
2	0.06	0.01
3	0.23	0.06
4	0.50	0.45
5	0.19	0.49

Taulukosta nähdään, että esimerkiksi 50% lähijunassa vastanneista kokee junien kulkevan täsmällisesti aikataulun mukaan hyvin, ja 49% metrossa vastanneista erittäin hyvin.

c)

H_0 = Kulkuneuvolla ja kysymyksellä K1A4 ei ole yhteyttä.
H_1 = Kulkuneuvolla ja kysymyksellä K1A4 on yhteys.

χ^2 - testin tulokset:

```
Pearson's Chi-squared test

data:  table(asty_osa$LIIKENNEMUOTO, asty_osa$K1A4)
X-squared = 466.16, df = 4, p-value < 2.2e-16
```

Koska $p < 2.2 \cdot 10^{16} < 0.01$,
nollahypoteesi hylätään, kulkuneuvolla, jossa kysely on tehty on siis yhteys kysymyksen K1A4 tulokseen.

d)
$H_0 = \mu_{metro} = 4.3$
$H_1 = \mu_{metro} \neq 4.3$

T-testin tulokset:

```
One Sample t-test

data:  asty_osa$K1A4[asty$LIIKENNEMUOTO == 2]
t = -21.949, df = 3171, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 4.3
95 percent confidence interval:
 3.927869 3.988902
sample estimates:
mean of x
 3.958386
```

H_0 voidaan hylätä, sillä $p < 2.2 \cdot 10^{16} < 0.05$
Täsmällisyysmuuttujan keskiarvo on siis erisuuri kuin 4.3.

3

c) Keskiarvot ja 95% luottamusvälit ikäryhmittäin:

Alle 20-vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "<20"], 0.95)
[1] 3.357763 3.622951 3.888139
```

20-29- vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "20-29"], 0.95)
[1] 3.558746 3.708075 3.857403
```

30-39- vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "30-39"], 0.95)
[1] 3.829682 3.991803 4.153925
```

40-49- vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "40-49"], 0.95)
[1] 3.777584 3.950000 4.122416
```

50-59- vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "50-59"], 0.95)
[1] 3.496350 3.697368 3.898387
```

60-69- vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "60-69"], 0.95)
[1] 3.687218 3.914894 4.142569
```

Yli 70- vuotiaat:

```
KeskiarvoJaVali(asty_juna$K1A4[asty_juna$ikaluokka == "70<="], 0.95)
[1] 3.222788 3.785714 4.348640
```

Yli 70-vuotiailla luottamusväli on suurin, joka saattaa johtua siitä, että tämä on suurin ikäluokka, jolloin vaihtelua on myös enemmän.

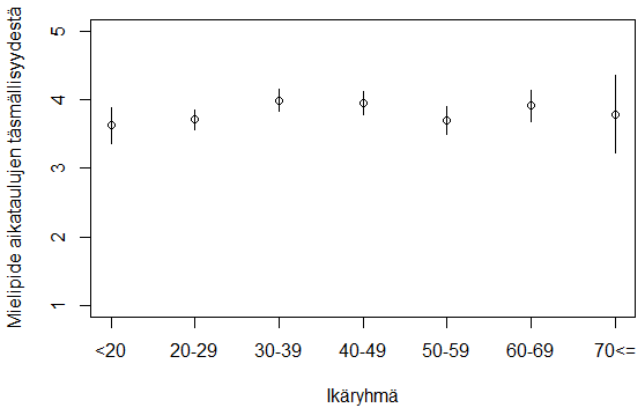

Figure 3: Täsmällisyysmuuttujan keskiarvot ikäryhmittäin ja 95% luottamusvälit

4

a) LogBKT- ja hedelmallisyys- muuttujien lineaarinen malli:

```
summary(fit)

Call:
lm(formula = yk$hedelmallisyys ~ yk$LogBKT)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9154 -0.7052 -0.1052  0.6040  3.3484

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.71165    0.52627   16.55   <2e-16 ***
yk$LogBKT   -0.70678    0.06309  -11.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.993 on 129 degrees of freedom
Multiple R-squared:  0.4931,Adjusted R-squared:  0.4892
F-statistic: 125.5 on 1 and 129 DF,  p-value: < 2.2e-16
```

Regressiosuoran yhtälö on siis:

$y = -0.70678x + 8.71165$

Muuttujilla näyttäisi olevan käänteinen yhteys, logaritmisen bkt:n kasvaessa hedelmällisyys laskee.
Logaritminen bkt myös selittää hedelmällisyyttä noin 49% selitysasteella.

b)

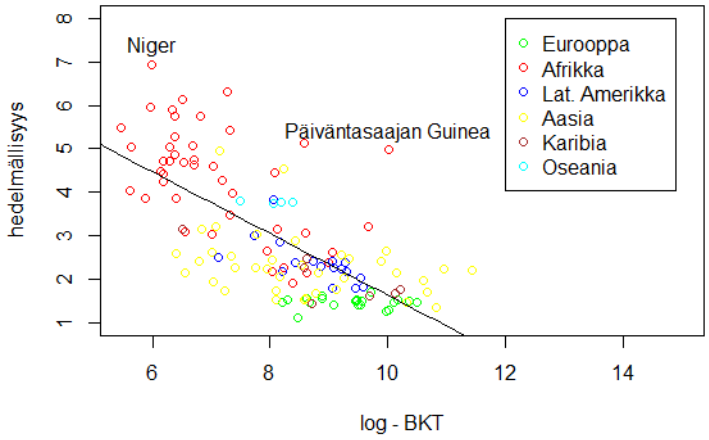

Figure 4: Logaritmisen bkt:n ja hedelmällisyysluvun välinen hajontakuvio ja regressiosuora + poikkeavimmat valtiot.

c) Lineaarinen malli, kun lisätään imeväiskuolleisuusmuuttuja:

```
summary(fit2)

Call:
lm(formula = yk$hedelmallisyys ~ yk$LogBKT + yk$imevaiskuolleisuus)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43841 -0.42471 -0.03409  0.38178  1.94075

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)            2.486054   0.647197   3.841 0.000192 ***
yk$LogBKT             -0.106735   0.067659  -1.578 0.117139
yk$imevaiskuolleisuus  0.036534   0.003126  11.686  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6934 on 128 degrees of freedom
Multiple R-squared:  0.7547,Adjusted R-squared:  0.7509
F-statistic:   197 on 2 and 128 DF,  p-value: < 2.2e-16
```

99% luottamusvälit:

```
confint(fit2, level = 0.99)
                             0.5 %     99.5 %
(Intercept)            0.79377180 4.17833573
yk$LogBKT             -0.28364767 0.07017843
yk$imevaiskuolleisuus  0.02835905 0.04470825
```

Imeväiskuolleisuus näyttäisi olevan parempi selittävä muuttuja hedelmällisyysluvulle parametrien p-arvojen perusteella.
Myös mallin seltitysaste on nyt parempi, noin 75%.
Tämä saattaa johtua siitä, että niissä maissa, missä imeväiskuolleisuus on suurempi tehdään myös enemmän lapsia.

5

a)

Tunnuslukuja:
Keskiarvo $\bar{y} = 0.2986064$
Mediaani $= 0.2852378$
Keskihajonta $s = 0.137342$

Parametrin jakauma muistuttaa hieman khiin neliön jakaumaa ja sen todennäköisin arvo näyttää olevan noin 0.2.

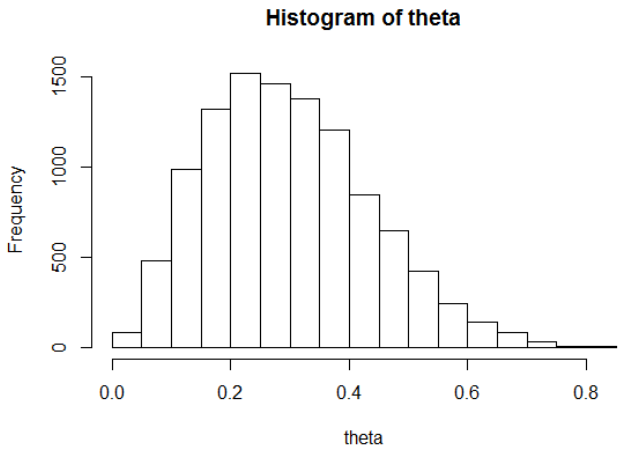

Figure 5: Funktion tuloksesta saatu histogrammi otoskoolla 100000

b) Funktion tuloste otoskoolla 100000:

```
simul2(100000)
havainnot
         1          2          3          4          5
0.19365835 0.38313609 0.26858760 0.14085413 0.01376383
```

Pekka heittelee siis todennäköisimmin kolikkoa 2.