STK1000: Løsningsforslag Uke 44

2023-10-30

Oppgave 7.86(R)

a)

Vi regner ut gjennomsnitt og varians i hver av gruppene:

```
data = read.csv('../../ips10e_csv_data_sets/ips10e_ch7_csv_data_sets/EX07-086PAIRED.csv')
g1 = data$Group1
g2 = data$Group2
xb_1 = mean(g1)
xb_2 = mean(g2)
s2_1 = var(g1)
s2_2 = var(g2)
c(xb_1, s2_1)
```

```
[1] 49.69200  5.37264
```

```
c(xb_2, s2_2)
```

```
[1] 50.545000  3.703161
```

Vi utfører en to-utvalgs t-test (two-sample t-test) fra side 440 med $H_0 : \mu_1 = \mu_2$ eller $H_0 : \mu_1 - \mu_2 = 0$ og $H_a : \mu_1 \neq \mu_2$. Testobservatoren er da

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

og vi kan bruker `t.test` i R for å få testobservatoren, frihetsgradene og p-verdien:

```r
t.test(g1, g2)
```

```
	Welch Two Sample t-test

data:  g1 and g2
t = -0.89538, df = 17.411, p-value = 0.3828
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.859351  1.153351
sample estimates:
mean of x mean of y 
   49.692    50.545 
```

Ekstra:

Alternativt kan vi regne dette ut selv i R:

```r
n_1 = length(g1)
n_2 = length(g2)
denom = s2_1 / n_1  + s2_2 / n_2
d1 = (s2_1 / n_1)^2 / (n_1-1)
d2 = (s2_2 / n_2)^2 / (n_2-1)
df = denom^2 / (d1 + d2) # finner frihetsgrader (side 447)
t = (xb_1 - xb_2) / sqrt(denom) # finner t
p = 2 * pt(t, df) # ganger p-verdien med 2 fordi vi har tosidig.
c(t, df, p)
```

```
[1] -0.8953783 17.4108684  0.3827970
```

b)

Vi regner nå ut gjennomsnitt og varians av differansene og foretar så en parvis t-test

```r
diff = g1 - g2
xb = mean(diff)
s2 = var(diff)
c(xb, s2)
```

```
[1] -0.853000  1.610668
```

```r
t.test(diff)
```

```

	One Sample t-test

data:  diff
t = -2.1254, df = 9, p-value = 0.06248
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.76087438  0.05487438
sample estimates:
mean of x 
   -0.853 
```

Prøv gjerne å kode t-testen selv her i stedet for å bruke `t.test`.

c)

Vi ser at p-verdiene er veldig forskjellige. I a) er testen vår langt fra signifikant, mens i b) vil vi faktisk forkaste H_0 hvis vi har et signifikansnivå på 0.1 (som kanskje er litt høyt). Uansett viser resultatene i b) at det kanskje er en forskjell, mens i resultatene i a) er det ingen ting som tilsier at det er en forskjell.

Oppgave 7.87(R)

a) I outputen fra `t.test` er det inkludert et 95%-konfidensintervall. Vi kan derfor se fra Oppgave 7.86 a) at vi har konfidensintervallet $(-2.848, 1.158)$ for to-utvalgs testen

b) Fra `t.text` i Oppgave 7.126 b) har vi konfidensintervallet $(-1.757, 0.067)$ for parvis sammenligning.

c) I a) har vi intervallet -0.845 $\pm$ 2.003 og i b) har vi intervallet -0.845 $\pm$ 0.912. Altså har vi samme senter for intervallene, mens marginen (bredden) for intervallet i a) er større enn det i b). Dette er naturlig siden a) er kraftig påvirket av variansen mellom individene i hver gruppe, mens i b) bryr vi oss kun om variansen i differansene.

Check In - Oppgave 10.3

a) $\widehat{y} = 29.578 - 0.655 \cdot 9.5 = 23.3555$.

b) $e = y - \widehat{y} = 24.3 - 23.3555 = 0.9445$.

c) Fra figur 10.3 (side 559) ser vi at det er bare to observasjoner med mindre enn 4000 skritt. Man må derfor være litt forsiktig med å predikere i dette området ettersom vi har lite bevis på at regresjonslinjen passer bra her. Videre har vi ingen observasjoner over 16000 skritt, så her må vi ekstrapolere. Vi har derfor begrenset tillit til prediksjonene her. For 10000 skritt har vi mange nærliggende observasjoner, så her er det rimelig å predikere. Men merk at det er forholdsvis stor usikkerhet i prediksjonene (punktene er spredd ganske langt fra linjen), så vi det er viktig å formidle denne usikkerheten (og ikke bare oppgi punktestimatet $\widehat{y}$).

Check In - Oppgave 10.4

Her bruker vi $t = \frac{b_1}{\mathrm{SE}_{b_1}}$, med $n-2$ frihetsgrader. Testene utføres med et signifikansnivå på 5%, med $H_0 : \beta_1 = 0$, og $H_a : \beta_1 \neq 0$.

a) $t = \frac{1.4}{0.65} = 2.154$, $df = 20 - 2 = 18$ som gir en p-verdi $2 \cdot 0.0225 = 0.045$. Vi forkaster derfor H_0 til fordel for H_a. R-kode for å finne p-verdi: 2*pt(2.154, df=18, lower.tail = FALSE)

b) $t = \frac{2.2}{1.05} = 2.095$, $df = 30$ som gir en p-verdi $= 0.045$. Vi forkaster derfor H_0 til fordel for H_a.

c) $t = \frac{2.2}{1.05} = 2.095$, $df = 14$ som gir en p-verdi $= 0.055$. Vi forkaster derfor *ikke* H_0.

Check In - Oppgave 10.5

95% konfidensintervallene har formen $b_1 \pm t^* \mathrm{SE}_{b_1}$, hvor t^* har $n-2$ frihetsgrader og representerer verdien der $\alpha/2$ av t-verdiene er høyere enn denne, altså 0.975-persentilen. R-kode qt(0.975, df=18) i oppgave a, osv.

a) $t^* = 2.101$, gir intervallet $1.4 \pm 2.101 \cdot 0.65 = (0.034, 2.766)$.

b) $t^* = 2.042$, gir intervallet $2.2 \pm 2.048 \cdot 1.05 = (0.056, 4.344)$.

c) $t^* = 2.145$, gir intervallet $2.2 \pm 2.145 \cdot 1.05 = (-0.052, 4.452)$.

Vi ser at vi vil trekke samme konklusjoner fra intervallene som fra hypotesetestene i Oppgave 10.3 (som forventet!).

Eksamen V-2008 oppg. 2 og H-2015 oppg. 3 a,b,c

Løsningsforslag til eksamensoppgavene er på emnets semesterside

https://www.uio.no/studier/emner/matnat/math/STK1000/oppgaver/losningsforslag/