

STK1110 - Statistiske metoder og dataanalyse høsten 2022

Løsningsforslag til oppgaver i læreboka for uke 39

Lærebok: *Modern Mathematical Statistics with Applications (3rd ed.)* - Devore, Berk & Carlton

Exercise 9.49

Ved hypotesetesting med signifikansnivå α , forkaster vi nullhypotesen H_0 hvis P -verdien er mindre enn eller lik α (jf. side 457 i læreboka). Vi har her gitt $\alpha = 0.05$.

- a) Siden P -verdi $= 0.01 \leq 0.05$, forkaster vi H_0 .
- b) Siden P -verdi $= 0.021 \leq 0.05$, forkaster vi H_0 .
- c) Siden P -verdi $= 0.078 > 0.05$, forkaster vi ikke H_0 .
- d) Siden P -verdi $= 0.047 \leq 0.05$, forkaster vi H_0 .
- e) Siden P -verdi $= 0.148 > 0.05$, forkaster vi ikke H_0 .

Exercise 9.52

Siden vi tester nullhypotesen $H_0 : \mu = 30$ mot det tosidige alternativet $H_a : \mu \neq 30$, er P -verdien som svarer til test observatoren z gitt ved

$$P\text{-verdi} = 2 \{1 - \Phi(|z|)\},$$

jf. side 460 i læreboka.

a) P -verdi $= 2 \{1 - \Phi(|2.10|)\} = 0.0357$

```
1 > 2*(1-pnorm(abs(2.10), mean=0, sd=1))
2 [1] 0.03572884
```

b) P -verdi $= 2 \{1 - \Phi(|-1.75|)\} = 0.0801$

```
1 > 2*(1-pnorm(abs(-1.75), mean=0, sd=1))
2 [1] 0.08011831
```

c) P -verdi $= 2 \{1 - \Phi(|-0.55|)\} = 0.5823$

```
1 > 2*(1-pnorm(abs(-0.55), mean=0, sd=1))
2 [1] 0.5823194
```

d) P -verdi $= 2 \{1 - \Phi(|1.41|)\} = 0.1585$

```
1 > 2*(1-pnorm(abs(1.41), mean=0, sd=1))
2 [1] 0.1585397
```

e) P -verdi $= 2 \{1 - \Phi(|-5.3|)\} = 1.16 \cdot 10^{-7}$

```

1 > 2*(1-pnorm(abs(-5.3), mean=0, sd=1))
2 [1] 1.158027e-07

```

Exercise 9.59

La μ = faktisk gjennomsnitt av organisk materiale (%) i jorda. Vi vil teste $H_0 : \mu = 3$ mot $H_a : \mu \neq 3$. Basert på kommentaren i slutten av oppgaven, antar vi at dataene er normalfordelt. Siden σ er ukjent, bruker vi her en t-test for ett utvalg, jf. side 444 i læreboka.

Av tallene i oppgaven får vi testobservatoren

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.481 - 3}{1.616/\sqrt{30}} = -1.759$$

Dermed blir P -verdien (jf. sidene 461-462 i læreboka)

$$P\text{-verdi} = 2 \{1 - \Phi_{t,29}(|-1.759|)\} = 0.089,$$

hvor $\Phi_{t,29}$ er den kumulative fordelingsfunksjonen for t -fordelingen med $n-1 = 30-1 = 29$ frihetsgrader.

Når $\alpha = 0.10$, har vi $P\text{-verdi} = 0.089 \leq 0.10$, som betyr at vi forkaster H_0 .

Når $\alpha = 0.05$, har vi $P\text{-verdi} = 0.089 > 0.05$, som betyr at vi ikke forkaster H_0 .

```

1 > soil.data = c(1.10, 5.09, 0.97, 1.59, 4.60, 0.32, 0.55, 1.45,
2 +             0.14, 4.47, 1.20, 3.50, 5.02, 4.67, 5.22, 2.69,
3 +             3.98, 3.17, 3.03, 2.21, 0.69, 4.47, 3.31, 1.17,
4 +             0.76, 1.17, 1.57, 2.62, 1.66, 2.05)
5 > mean.soil = mean(soil.data)
6 > print(mean.soil)
7 [1] 2.481333
8 >
9 > sd.soil = sd(soil.data)
10 > print(sd.soil)
11 [1] 1.615641
12 >
13 > T.value = (mean.soil - 3)/(sd.soil/sqrt(length(soil.data)))
14 > print(T.value)
15 [1] -1.758345
16 >
17 > p.value = 2*(1-pt(abs(T.value), df=length(soil.data)-1))
18 > print(p.value)
19 [1] 0.08923962

```

Exercise 9.70

Hvis $X_1, \dots, X_n$ er uavhengige og $N(\mu, \sigma_0^2)$ -fordelte med kjent varians σ_0^2 , og $x_1, \dots, x_n$ er de observerte verdiene av X_i -ene, er likelihooden gitt ved

$$\begin{aligned}
 L(\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma_0^2} \right\} \\
 &= (2\pi\sigma_0^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\
 &= (2\pi\sigma_0^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_0^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right\}.
 \end{aligned}$$

Vi ser at uansett verdi av σ_0^2 , er maximum likelihood estimatet for μ gitt ved $\hat{\mu} = \bar{x}$.

Vi vil teste $H_0 : \mu = 0$ mot $H_a : \mu \neq 0$.

Vi ser på tilfellet med kjent varians σ_0^2 (selv om oppgaven først antar at $\sigma_0^2 = 1$)

Likelihood ratioen blir

$$LR = \frac{L(0)}{L(\hat{\mu})} = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2} [\sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2]\right\}}{(2\pi\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2\right\}} = \exp\left\{-\frac{n}{2\sigma_0^2} \bar{x}^2\right\}.$$

Likelihood ratio testen forkaster H_0 hvis

$$\begin{aligned} LR &\leq k \\ \iff \exp\left\{-\frac{n}{2\sigma_0^2} \bar{x}^2\right\} &\leq k \\ \iff -\frac{n}{2\sigma_0^2} \bar{x}^2 &\leq \log k \\ \iff \bar{x}^2 &\geq -\frac{2\sigma_0^2}{n} \log k \\ \iff |\bar{x}| &\geq \sqrt{-\frac{2\sigma_0^2}{n} \log k} \\ \iff \bar{x} &\leq -\sqrt{-\frac{2\sigma_0^2}{n} \log k} \quad \text{eller} \quad \bar{x} \geq \sqrt{-\frac{2\sigma_0^2}{n} \log k} \\ \iff \bar{x} &\leq -c \quad \text{eller} \quad \bar{x} \geq c, \end{aligned}$$

hvor $c = \sqrt{-\frac{2\sigma_0^2}{n} \log k}$.

Vi har at $\bar{X} \sim N(\mu, \sigma_0^2/n)$, og dermed $(\bar{X} - \mu)/(\sigma_0/\sqrt{n}) \sim N(0, 1)$

Vi har derfor at

$$\begin{aligned} P(\text{forkast } H_0 \mid \mu = 0) &= P(\bar{X} \leq -c \mid \mu = 0) + P(\bar{X} \geq c \mid \mu = 0) \\ &= P\left(\frac{\bar{X} - 0}{\sigma_0/\sqrt{n}} \leq -\frac{c}{\sigma_0/\sqrt{n}} \mid \mu = 0\right) + P\left(\frac{\bar{X} - 0}{\sigma_0/\sqrt{n}} \geq \frac{c}{\sigma_0/\sqrt{n}} \mid \mu = 0\right) \\ &= \Phi\left(-\frac{c}{\sigma_0/\sqrt{n}}\right) + 1 - \Phi\left(\frac{c}{\sigma_0/\sqrt{n}}\right) \end{aligned}$$

For at testen skal få nivå α må vi ha

$$\frac{c}{\sigma_0/\sqrt{n}} = z_{\alpha/2}$$

som gir

$$c = z_{\alpha/2} \frac{\sigma_0}{\sqrt{n}}.$$

Exercise 9.75

a) Vi bruker formelen for alternativet $H_a : p > p_0$ på side 452 i læreboka. Sannsynligheten for feil av type II er:

$$\beta(p') = \Phi\left(\frac{p_0 - p' + z_{\alpha}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right)$$

Med $\alpha = 0.01$, $p_0 = 0.2$ og $p' = 0.21$, får vi (se R-kode nedenfor):

- når $n = 100$ er $\beta(0.21) = 0.9793$
- når $n = 2500$ er $\beta(0.21) = 0.8548$
- når $n = 10000$ er $\beta(0.21) = 0.4323$
- når $n = 40000$ er $\beta(0.21) = 0.0043$
- når $n = 90000$ er $\beta(0.21) = 0.00000019$

```

1 > z.alpha = qnorm(0.99, mean=0, sd=1)
2 > p.0 = 0.2
3 > p = 0.21
4 > n=c(100,2500,10000,40000,90000)
5 > pnorm((p.0-p+z.alpha*sqrt(p.0*(1-p.0)/n))/sqrt(p*(1-p)/n), mean=0, sd=1)
6 [1] 9.792796e-01 8.547522e-01 4.322942e-01 4.323811e-03 1.879021e-07

```

b) Vi bruker testobservatoren gitt side 451 av læreboka:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Her blir P -verdien (jf. side 460 i læreboka)

$$P\text{-verdi} = 1 - \Phi(z).$$

Med $\hat{p} = 0.21$ og $p_0 = 0.2$, får vi (se R-kode nedenfor):

- når $n = 100$ er P -verdi = 0.401
- når $n = 2500$ er P -verdi = 0.106.
- når $n = 10000$ er P -verdi = 0.0062.
- når $n = 40000$ er P -verdi = 0.00000029.

```

1 > p.0 = 0.2
2 > p.hat = 0.21
3 > n=c(100,2500,10000,40000)
4 > z=(p.hat-p.0)/sqrt(p.0*(1-p.0)/n)
5 > 1-pnorm(z, mean=0, sd=1)
6 [1] 4.012937e-01 1.056498e-01 6.209665e-03 2.866516e-07

```

c) Nei, det er ikke rimelig å bruke en test med signifikansnivå 1% når vi har et utvalg på $n = 40000$. For med et så stort utvalg vil selv et ubetydelig avvik fra nullhypotesen føre til forkastning.