STK1110 - Statistiske metoder og dataanalyse høsten 2019
Løsningsforslag til oppgaver i læreboka for uke 34

Vinnie Ko & Ørnulf Borgan

August 29, 2019

Lærebok: *Modern Mathematical Statistics with Applications (2nd ed.)* - Devore & Berk

Exercise 7.1

a) Vi bruker gjennomsnittet (sample mean) til å estimere forventingen (population mean):

$$\widehat{\mu} = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3753}{33} = 113.7$$

```
> # Leser data fra web:
> IQ.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe7_
    01.txt", header=T)
>
> # Trekker ut vektoren med IQ-verdiene:
> IQ = IQ.data$IQ1
>
> # Beregner gjennomsnittet:
> IQ.mean = mean(IQ)
> print(IQ.mean)
[1] 113.7273
```

b) Vi bruker empirisk median (sample median) til å estimere populasjonsmedianen:

```
> # Beregner empirisk median:
> IQ.median = median(IQ)
> print(IQ.median)
[1] 113
```

c) Vi bruker empirisk standardavvik (sample standard deviation) til å estimere populasjonsstandard-avviket (population standard deviation):

```
> # Beregn empirisk standardavvik:
> IQ.sd = sd(IQ)
> print(IQ.sd)
[1] 12.74331
```

d) Vi bruker andelen i utvalget som har IQ over 100 til å estimere andelen populasjonen som har IQ over 100:

```
> # Andelen i utvalget med IQ > 100:
> print(sum(IQ>100)/length(IQ))
[1] 0.9090909
```

e) Vi bruker den empiriske variasjonskoeffisienten til å estimere variasjonskoeffisienten i populasjonen:

```
> # Estimerer variasjonskoeffisienten:
> coeff.of.var = sd(IQ)/mean(IQ)
> print(coeff.of.var)
[1] 0.1120515
```

Exercise 7.4

a) Vi har at

$$\mathrm{E}\left[\overline{X}-\overline{Y}\right] = \mathrm{E}[\overline{X}] - \mathrm{E}[\overline{Y}] = \mu_1 - \mu_2$$

Altså er $\overline{X}-\overline{Y}$ en forventningsrett estimator for $\mu_1 - \mu_2$.

Estimatet blir 7.6, se R-koden:

```
> # Fjerner variabler fra forrige oppgave fra minnet:
> rm(list=ls(all=TRUE))
>
> # Leser data fra web:
> IQ.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe7_
    04.txt", header=T)
>
> # Trekker ut IQ-verdiene for menn og kvinner:
> IQ.menn=IQ.data$VERBIQ[IQ.data$mf=="male"]
> IQ.kvinner=IQ.data$VERBIQ[IQ.data$mf=="female"]
>
> # Beregner estimatet:
> IQ.diff = mean(IQ.menn)-mean(IQ.kvinner)
> print(IQ.diff)
[1] 7.644444
```

b) Vi finner

$$\begin{aligned}
\mathrm{V}\left(\overline{X}-\overline{Y}\right) &= \mathrm{V}(\overline{X}) + \mathrm{V}(\overline{Y}) && \text{(på grunn av uavhengighet)} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{V}(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(Y_i) && \text{(på grunn av u.i.f.)} \\
&= \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.
\end{aligned}$$

Standardfeilen til estimatoren er dermed

$$\sqrt{\mathrm{V}\left(\overline{X}-\overline{Y}\right)} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}.$$

Vi bruker empirisk varians til å estimere populasjonsvariansene for menn og kvinner (dvs. σ_1^2 og σ_2^2). Estimatet for standardfeilen til estimatoren blir 4.11, se R-koden.

```
> # Beregner estimert standardfeil:
> var.diff = var(IQ.menn)/length(IQ.menn) + var(IQ.kvinner)/length(IQ.kvinner)
> se.diff = sqrt(var.diff)
> print(se.diff)
[1] 4.108677
```

c) Vi bruker empirisk standardavvik for estimere populasjonsstandardavviket for menn og kvinner, og finner at et estimat for forholdet mellom de to populasjonsstandardavvikene er 1.09, se R-koden.

```
> # Beregner forholdet mellom de empiriske standardavvikene
> sd.ratio = sd(IQ.menn)/sd(IQ.kvinner)
> print(sd.ratio)
[1] 1.089362
```

d) Vi har at

$$\mathrm{V}(X_i - Y_i) = \mathrm{V}(X_i) + \mathrm{V}(Y_i) = \sigma_1^2 + \sigma_2^2$$

Vi bruker empirisk varians til å estimere populasjonsvariansene og får estimatet 278.4, se R-koden.

```
> # Finner estimat for variansen til differenasen:
> diff.var = var(IQ.menn) + var(IQ.kvinner)
> print(diff.var)
[1] 278.3993
```

Exercise 7.6

a) Fra STK1100 har vi at hvis $Y \sim N(\mu, \sigma^2)$, så er $X = \exp(Y) \sim LN(\mu, \sigma^2)$, dvs. lognormalt-fordelt. Merk at parameterne i $LN(\mu, \sigma^2)$-fordelingen er forventning og varians til $Y = \log(X)$, ikke til X.

La $x_1, \ldots, x_n$ betegne de observerte målingene av vannstrømmen, og la $y_i = \ln(x_i)$.

Da er y_i-ene observerte verdier fra $N(\mu, \sigma^2)$-fordelingen.

Fra STK1100 har vi dermed at estimater for μ og σ^2 er gitt ved

$$\widehat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{og} \quad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Vi finner estimatene ved R-kommandoene:

```
> # Leser dataene:
> flow.data = read.table("https://www.uio.no/studier/emner/matnat/math/STK1110/data/exe7_06.txt", header=T)
>
> # Trekker ut vektoren med observasjonene:
> flow=flow.data$flow
>
> # Bestemmer estimatene:
> my.hat=mean(log(flow))
> sigma2.hat=var(log(flow))
> print(c(my.hat,sigma2.hat))
[1] 5.1016519 0.2460772
```

b) Fra side 205 i læreboka, har vi for lognormal-fordeling at $\mathrm{E}[X] = e^{\mu+\sigma^2/2}$.

Ved å sette inn estimatene fra punkt (a) får vi et estimatet for $\mathrm{E}[X]$, se R-koden.

```
> # Finner estimat for E(X):
> print(exp(my.hat + sigma2.hat/2))
[1] 185.8036
```

Exercise 7.15

a) Fra oppgaven har vi at $\theta = \frac{1}{2}\mathrm{E}[X^2]$.

Det er derfor rimelig å bruke estimatoren $\widehat{\theta} = \frac{1}{2n}\sum_{i=1}^{n} X_i^2$.

Vi har nå at

$$\mathrm{E}\left[\widehat{\theta}\right] = \mathrm{E}\left[\frac{1}{2n}\sum_{i=1}^{n} X_i^2\right] = \frac{1}{2n}\sum_{i=1}^{n}\mathrm{E}[X_i^2] = \frac{1}{2n}\sum_{i=1}^{n} 2\theta = \theta$$

Siden $\mathrm{E}\left[\widehat{\theta}\right] = \theta$, er $\widehat{\theta}$ en forventningsrett estimator for θ.

b) Vi bruker R til bestemme $\widehat{\theta}$ for blodplasma konsentrasjonene:

```
> # Lager en vektor med dataene:
> kons = c(16.88, 10.23, 4.59, 6.66, 13.68, 14.23, 19.87, 9.40, 6.51, 10.95)
>
> # Beregner estimatet:
> theta.hat = mean(kons^2)/2
> print(theta.hat)
[1] 74.50529
```

Exercise 7.19

a) En estimator for λ er $\widehat{\lambda} = Y/n$. Vi vil bruke sammenhengen mellom λ og p for å finne en estimator for p. Siden $\lambda = 0.50p + 0.15$, har vi at $p = 2\lambda - 0.30$. En estimator for p er dermed $\widehat{p} = 2(Y/n) - 0.3$.

Ved å bruke data gitt i oppgaven, finner vi estimatet $\widehat{p} = 2 \cdot (20/80) - 0.30 = 0.20$.

b) Vi har at $\mathrm{E}\left[\widehat{\lambda}\right] = \lambda$. Dermed er

$$\mathrm{E}\left[\widehat{p}\right] = 2\mathrm{E}\left[\widehat{\lambda}\right] - 0.30 = 2\lambda - 0.30 = p$$

Altså er $\widehat{p}$ en forventningsrett estimator for p.

c) Av regelen om total sannsynlighet har vi nå at

$$\begin{aligned}
\lambda &= P(\text{svarer ja}) \\
&= P(\text{kort type I}) \cdot P(\text{svarer ja} \,|\, \text{kort type I}) + P(\text{kort type II}) \cdot P(\text{svarer ja} \,|\, \text{kort type II}) \\
&= (70/100) \cdot p + (30/100) \cdot 0.30 \\
&= 0.70p + 0.09
\end{aligned}$$

Ved å uttrykke p ved hjelp av λ, får vi $p = (10/7) \cdot \lambda - 9/70$.

En forventningsrett estimator for p er nå

$$\widehat{p} = \frac{10}{7} \cdot \frac{Y}{n} - \frac{9}{70}.$$