Sandsynlighedsregning og Statistik

Januar 2025: Vejledende besvarelse

Opgave 1

Vi starter med at indlæse datasættet som beskrevet i opgaveteksten:

```
mydata <- read.csv("jan2025.txt", sep="")
```

De 9 korrekte svar er:

- **C**, **A**, **D**, **B**, **D**, **D**, **C**, **A**, **B**

Spørgsmål 1.1: Korrekt svar C

Normalitetsantagelsen vedrører fejlleddene. Denne antagelse valideres dermed ved at man undersøger normalitet af de standardiserede residualer.

Spørgsmål 1.2: Korrekt svar = A

Vi fitter den lineære regression, og finder estimatet s under *Residual standard error*. Derefter fås den centrale estimat for variansparameteren ved at beregning s^2. Vi finder korrekt svar **A** efter afrunding.

```
model <- lm(reaktionstid~alder,data=mydata)
summary(model)
```

```
Call:
lm(formula = reaktionstid ~ alder, data = mydata)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-424.41 -109.57   -2.41  121.28  485.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 794.9165    21.1506  37.584  < 2e-16 ***
alder         1.9781     0.5558   3.559 0.000457 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 160.7 on 219 degrees of freedom
Multiple R-squared:  0.05467,   Adjusted R-squared:  0.05035
F-statistic: 12.66 on 1 and 219 DF,  p-value: 0.0004568
```

```
160.7^2
```

```
[1] 25824.49
```

Spørgsmål 1.3: Korrekt svar = D

Fra `summary(model)` kan vi også aflæse p-værdien for nulhypotesen H_0: $\beta = 0$, hvor β er hældningen mht alder. Denne p-værdi er $0.000457 = 0.0457\% < 0.1\%$. Der er således stærk evidens for at en persons alder har betydning for personens reaktionstid.

Spørgsmål 1.4: Korrekt svar = B

Estimatet $\hat{\beta} = 1.9781$ aflæses også fra `summary(model)`. Idet enheden for *reaktionstid* og for *alder* er henholdsvis *ms* og *år* passer dette med 1.98 ms/år. 95% konfidensinterval findes via følgende R kode, og vi ser at dette også passer med svarmulighed **B**.

```
confint(model)["alder",]
```

```
    2.5 %    97.5 %
0.8826047 3.0735296
```

Spørgsmål 1.5: Korrekt svar = D

Der spørges til en prædiktion for et enkelt ny person (arrangøren af videnskabsfestivalen). Der skal altså beregnes et 95% konfidensinterval, hvilket gøres via kodne nedenfor. Bemærk, at den nedre grænse skal rundes ned samtidig med at den øvre grænse rundes op, for at vi kan være sikker på at få dækningsgrad på mindst 95%. Dette er dog en detalje, og vi ser at den korrekte svarmulighed er **D**.

```
predict(model,newdata=data.frame(alder=53),interval="prediction")
```

```
       fit      lwr      upr
1 899.7541 581.6375 1217.871
```

Spørgsmål 1.6: Korrekt svar = D

I bogen (DS) siges det, at man kan antage ens varians for situationen med to uafhængige stikprøver når de tilhørende stikprøvevarianser er af samme størrelsesorden. Dette er tilfældet her. Det sammenvejede variansestimat fås via en sammenvejning ift frihedsgraderne i de to stikprøver, dvs. med vægte $128 - 1 = 127$ og $93 - 1 = 92$. Følgende beregning giver dermed, at den korrekte svarmulighed er **D**.

```
(24931*127 + 30137*92)/(127 + 92)
```

```
[1] 27118
```

Spørgsmål 1.7: Korrekt svar = C

Estimatet for forskellen findes som forskellen af estimaterne, altså forskellen mellem stikprøvegennemsnit. Følgende beregning viser, at dette passer med svarmulighed **C** og **D**.

```
870 - 844
```

```
[1] 26
```

Standard error findes som kvadratroden af den estimerede spredning på variansen på forskellen af populationsmiddelværdier. Idet de to stikprøver er uafhængige skal de tilhørende varianser lægges sammen. Efter afrunding til 3'de decimal passer dette med svarmulighed **C**.

```r
sqrt(27118 * (1/128 + 1/93))
```

```
[1] 22.43771
```

Spørgsmål 1.8: Korrekt svar = A

Først bemærker vi, at det samlede antal frihedsgrader er $128 + 93 - 2 = 219$.

Bruges estimat og standard error som angivet i svarmulighed **C** for spørgsmål 1.7 fås 95% konfidensinterval som estimat plus/minus t-fraktil gange standard error. Dette giver:

```r
26 + c(-1,1)*qt(0.975,df=219)*22.438
```

```
[1] -18.22205  70.22205
```

Vi kan naturligvis også lave beregningen direkte udfra antal observationer og stikprøvegennemsnit og -varianser som angivet i tabellen:

```r
870 - 844 + c(-1,1) * qt(0.975,df=128+93-2) *
  sqrt((24931*(128-1)+30137*(93-1))/(128+93-2)) * sqrt(1/128 + 1/93)
```

```
[1] -18.22147  70.22147
```

vi ser, at afrundingerne giver forskellige resultater på 3'de decimal (og her er den sidste beregning den korrekte). Men afrundes der til 1 decimal, så ser vi af den korrektesvarmulighed ved begge beregninger er **A**.

Spørgsmål 1.9: Korrekt svar = B

T-teststørrelsen for nulhypotesen $H_0 : \mu_{\text{kvinde}} - \mu_{\text{mand}} = 0$ beregnes som estimat divideret med standard error. Nedenfor bruger vi bare svarmulighed **C** for spørgsmål 1.7, men man kunne naturligvis også lave beregningen via antal observetioner og stikprøvegennemsnit og -varianser som angivet i tabellen. Vi ses, at t-teststørrelsen passer med svarmulighederne **A** og **B**:

```r
26/22.438
```

```
[1] 1.158749
```

Derefter beregner vi p-værdien ved at findes de to halesandsynligheder i en t-fordeling med 219 frihedsgrader:

```
2*(1-pt(1.159,df=219))
```

```
[1] 0.2477184
```

Dette passer med svarmulighed **B**. Det bemærkes også, at en p-værdi på omkring 25% på ingen måde indikerer en sammenhæng mellem biologisk køn og reaktionstid.

Opgave 2

Spørgsmål 2.1

Sandsynlighederne $P(X = 3)$ og $P(Y = 3)$ beregnes som henholdsvis den 3'de rækkemarginal og den 3'de søjlemarginal. Altså

$$P(X = 3) = 0.02 + 0.02 + 0.23 + 0.03 = 0.3$$
$$P(Y = 3) = 0.02 + 0.06 + 0.23 + 0.02 = 0.33$$

Idet der således gælder $P(X = 3) \neq P(Y = 3)$ kan vi også konkludere, at X og Y *ikke har samme fordeling*.

Spørgsmål 2.2

De stokastiske variable X og Y er *ikke stokastisk uafhængige*. For at argumentere for dette skal vi bare finde en fælles hændelse hvis sandsynlighed ikke er produktet af sandsynlighederne for de to marginale hændelser. Og der gælder f.eks.

$$P(X = 3, Y = 3) = 0.23 \neq 0.099 = 0.3 \cdot 0.33 = P(X = 3) \cdot P(Y = 3)$$

Spørgsmål 2.3

For at beregne middelværdien og variansen for X finder vi først punktsandsynlighederne for X. Dette gøres ved en tilsvarende marginalisering som anvendt i spørgsmål 2.1:

$$P(X = 1) = 0.20 + 0.07 + 0.02 + 0.01 = 0.3$$
$$P(X = 2) = 0.03 + 0.20 + 0.06 + 0.01 = 0.3$$
$$P(X = 3) = 0.02 + 0.02 + 0.23 + 0.03 = 0.3$$
$$P(X = 4) = 0.01 + 0.01 + 0.02 + 0.06 = 0.1$$

Vi kan derefter beregning middelværdi og andet moment:

$$\mathbf{E}(X) = 0.3 \cdot 1 + 0.3 \cdot 2 + 0.3 \cdot 3 + 0.1 \cdot 4 = 2.2$$
$$\mathbf{E}(X^2) = 0.3 \cdot 1^2 + 0.3 \cdot 2^2 + 0.3 \cdot 3^2 + 0.1 \cdot 4^2 = 5.8$$

Og dette giver variansen:

$$\mathbf{Var}(X) = \mathbf{E}(X^2) - \big(\mathbf{E}(X)\big)^2 = 5.8 - 2.2^2 = 0.96$$

Spørgsmål 2.4

Vi kan bruge de stokastiske variable X og Y til at beskrive synet for en tilfældig kvinde. De stokastiske variabel X og Y angiver scorer på henholdsvis højre og venstre øje. Og idet en lavere scorer svarer til et bedre syn fås dermed

$$\begin{aligned}
&P(\text{kvinden ser mindst ligeså godt på højre øje som på venstre øje}) \\
&\quad = P(X \leq Y) \\
&\quad = (0.20 + 0.07 + 0.02 + 0.01) + (0.20 + 0.06 + 0.01) + (0.23 + 0.03) + (0.06) \\
&\quad = 0.89
\end{aligned}$$

Spørgsmål 2.5

Vi beregner først sandsynligheden for at kvinder har perfekt syn på enten højre eller venstre øje:

$$\begin{aligned}
P\big((X = 1) \cup (Y = 1)\big) &= 0.20 + 0.07 + 0.02 + 0.01 + 0.03 + 0.02 + 0.01 \\
&= 0.36
\end{aligned}$$

Derefter beregner vi sandsynligheden for fælleshændelsen

$$\begin{aligned}
P\Big((X = 1) \cap \big((X = 1) \cup (Y = 1)\big)\Big) &= P\Big((X = 1) \cup \big((X = 1) \cap (Y = 1)\big)\Big) \\
&= P(X = 1) \\
&= 0.30
\end{aligned}$$

Dette giver dermed den betingede sandsynlighed:

$$
\begin{aligned}
&P(\text{perfekt syn på højre øje} \mid \text{perfekt syn på mindst et øje}) \\
&= P(X = 1 \mid (X = 1) \cup (Y = 1)) \\
&= \frac{P\Big((X = 1) \cap ((X = 1) \cup (Y = 1))\Big)}{P((X = 1) \cup (Y = 1))} \\
&= \frac{0.30}{0.36} \\
&= \frac{5}{6} \\
&\approx 83\%
\end{aligned}
$$

Spørgsmål 2.6

Der er tale om antallet af successer (= perfekt syn på begge øjne) ud 160 uafhængige gentagelser (=160 kvinder). Dette er altså en binomialfordeling med antalsparameter $n = 160$ og sandsynlighedsparameter

$$p = P(X = 1, Y = 1) = 0.20$$

I kursets notation er svaret dermed: *bin(160,0.2)*.

Spørgsmål 2.7

Dette spørgsmål skal besvares i forlængelse af situationen fra spørgsmål 2.6. Der er nu ialt 200 kvinder med samme successandsynlighed. Vi betinger med hændelsen $U + V = 35$.

Hvis succes indkodes som en *hvid kugle*, så svarer det til en situation hvor der trækkes *40 kugler* (=kvinderne fra den anden fabrik) fra en pose med *200 kugler* (=alle kvinderne), hvoraf *35 kugler er hvide* ($U + V$=antal kvinder med perfekt syn), og vi spørger til sandsynligheden for at trække *10 hvide kugler* (V=10). Vi spørger altså til sandsynligheden for udfaldet 10 i en hypergeometrisk fordeling med $m = 35$ hvide kugler, $n = 200 - 35 = 165$ sorte kugler, og hvor der trækkes $k = 40$ kugler. Dette kan beregnes i R via koden:

```
dhyper(10,35,165,40)
```

```
[1] 0.06805507
```

Spørgsmålet kan også besvares ved brug af uafhængige binomialfordelte stokastiske variable $V \sim \text{bin}(40, 0.2)$ og $U \sim \text{bin}(160, 0.2)$. Dermed fås $U + V \sim \text{bin}(200, 0.2)$ og

$$\begin{aligned} P(V = 10 \mid V + U = 35) &= \frac{P(V = 10, U = 25)}{P(U + V = 35)} \\ &= \frac{P(V = 10) \cdot P(U = 25)}{P(U + V = 35)} \\ &= \frac{\binom{40}{10} \cdot 0.2^{10} \cdot 0.8^{30} \cdot \binom{160}{25} \cdot 0.2^{25} \cdot 0.8^{135}}{\binom{200}{35} \cdot 0.2^{35} \cdot 0.8^{165}} \\ &= \frac{\binom{40}{10} \cdot \binom{160}{25}}{\binom{200}{35}} \end{aligned}$$

Vi genkender dette som den hypergeometriske sandsynlighed. Og en konkret beregning i R giver svaret:

```
choose(40,10) * choose(160,25) / choose(200,35)
```

```
[1] 0.06805507
```

Altså cirka 6.8%.

Opgave 3

Spørgsmål 3.1

Funktionen $f_{X,Y}$ er ikke negativ. Så den kan bruges som tætheden for en sandsynlighedsfordeling på $\mathbb{R}^2$ hvis den integrerer til 1. For at vise dette starter vi med at opskrive støtten M som et *Nord-Syd domæne*:

$$\begin{aligned} M &\overset{\text{def}}{=} \{(x, y) \in \mathbb{R}^2 \mid f_{X,Y}(x, y) > 0\} \\ &= \{(x, y) \in \mathbb{R}^2 \mid -1 < x + y < 1,\ -1 < x - y < 1\} \\ &= \{(x, y) \in \mathbb{R}^2 \mid -1 < x < 1,\ -1 + |x| < y < 1 - |x|\} \end{aligned}$$

Derefter kan vi beregne planintegralet:

$$\begin{aligned}
\int_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}(x,y) &= \int_M f_{X,Y}(x,y)\,\mathrm{d}(x,y) \\
&= \int_{-1}^{1}\int_{-1+|x|}^{1-|x|} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-1}^{1}\int_{-1+|x|}^{1-|x|} \tfrac{3}{2}\cdot(x+y)^2\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-1}^{1}\left[\tfrac{(x+y)^3}{2}\right]_{y=-1+|x|}^{y=1-|x|}\mathrm{d}x \\
&= \int_{-1}^{1}\frac{(1-|x|+x)^3-(-1+|x|+x)^3}{2}\,\mathrm{d}x \\
&= \int_{-1}^{1}\frac{(1-|x|+x)^3+(1-|x|-x)^3}{2}\,\mathrm{d}x \\
&= \int_{-1}^{1}\Big((1-|x|)^3+3\cdot(1-|x|)\cdot x^2\Big)\,\mathrm{d}x \\
&= 2\int_{0}^{1}\Big((1-x)^3+3\cdot(1-x)\cdot x^2\Big)\,\mathrm{d}x \\
&= 2\int_{0}^{1}\Big(1-3x+3x^2-x^3+3x^2-3x^3\Big)\,\mathrm{d}x \\
&= \int_{0}^{1}\Big(2-6x+12x^2-8x^3\Big)\,\mathrm{d}x \\
&= \big[2x\big]_{x=0}^{x=1}-\big[3x^2\big]_{x=0}^{x=1}+\big[4x^3\big]_{x=0}^{x=1}-\big[2x^4\big]_{x=0}^{x=1} \\
&= 2-3+4-2 \\
&= 1
\end{aligned}$$

Idet $f_{X,Y}(x,y) \geq 0$ og $\int_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}(x,y) = 1$ kan $f_{X,Y}(x,y)$ dermed bruges som tætheden for en sandsynlighedsfordeling på $\mathbb{R}^2$.

Spørgsmål 3.2

Der gælder, at $f_{X,Y}(x,y) \approx 0$ når $x+y \approx 0$, altså når vi er tæt på linjen $y = -x$, og at $f_{X,Y}(x,y)$ er størt når $x+y$ ligger lige under linjen $y = 1-x$ eller lige over linjen $y = -1-x$.

Dermed skal der være færrest punkter omkring linjen $y = -x$, og flest punkter lige under linjen $y = 1-x$ og lige over linjen $y = -1-x$.

Dette passer kun med fordeling nummer 3, som dermed er det rigtige svar.

Spørgsmål 3.3

De stokastiske variable X og Y er ikke uafhængige. Dette følger umiddelbart af, at støtten M ikke kan skrives som en produktmængde.

Spørgsmål 3.4

Tæthedsfunktionen $f_X(x)$ for X kan findes ved at marginalisere over det indre integral i *Nord-Syd* opskrivningen af planintegralet. Fra det sidste integral over $x \in (-1,1)$ i udledningerne i *spørgsmål 3.1* har vi dermed

$$\begin{aligned} f_X(x) &= 1_{(-1,1)}(x) \cdot \Big((1-|x|)^3 + 3 \cdot (1-|x|) \cdot x^2 \Big) \\ &= \begin{cases} 1 - 3 \cdot |x| + 6 \cdot x^2 - 4 \cdot |x|^3 & \text{for } x \in (-1,1) \\ 0 & \text{ellers} \end{cases} \end{aligned}$$

Spørgsmål 3.5

Vi starter med at bestemme andet momentet af $\frac{1}{X+Y}$ via LOTUS. Ved tilsvarende overvejelser som i *spørgsmål 3.1* fås dermed

$$\begin{aligned} \mathbf{E}\left(\left(\frac{1}{X+Y} \right)^2 \right) &= \int_M \frac{1}{(x+y)^2} \cdot f_{X,Y}(x,y) \, \mathrm{d}(x,y) \\ &= \int_{-1}^{1} \int_{-1+|x|}^{1-|x|} \tfrac{3}{2} \cdot \frac{(x+y)^2}{(x+y)^2} \, \mathrm{d}y \, \mathrm{d}x \\ &= \tfrac{3}{2} \int_{-1}^{1} \int_{-1+|x|}^{1-|x|} 1 \, \mathrm{d}y \, \mathrm{d}x \\ &= \tfrac{3}{2} \int_{-1}^{1} (2 - 2 \cdot |x|) \, \mathrm{d}x \\ &= 6 \int_{0}^{1} (1-x) \, \mathrm{d}x \\ &= 6 \int_{0}^{1} u \, \mathrm{d}u \\ &= 3 \end{aligned}$$

Middelværdien af $\frac{1}{X+Y}$ beregnes tilsvarende via LOTUS:

$$\begin{aligned}
\mathbf{E}\left(\frac{1}{X+Y}\right) &= \int_M \frac{1}{x+y}\cdot f_{X,Y}(x,y)\,\mathrm{d}(x,y)\\
&= \int_{-1}^{1}\int_{-1+|x|}^{1-|x|} \tfrac{3}{2}\cdot\frac{(x+y)^2}{x+y}\,\mathrm{d}y\,\mathrm{d}x\\
&= \tfrac{3}{2}\int_{-1}^{1}\int_{-1+|x|}^{1-|x|} (x+y)\,\mathrm{d}y\,\mathrm{d}x\\
&= \tfrac{3}{4}\int_{-1}^{1}\Big[(x+y)^2\Big]_{y=-1+|x|}^{y=1-|x|}\mathrm{d}x\\
&= \tfrac{3}{4}\int_{-1}^{1}\Big((1-|x|+x)^2-(-1+|x|+x)^2\Big)\,\mathrm{d}x\\
&= \tfrac{3}{4}\int_{-1}^{1}\Big((1-|x|+x)^2-(1-|x|-x)^2\Big)\,\mathrm{d}x\\
&= 3\int_{-1}^{1}(1-|x|)\cdot x\,\mathrm{d}x\\
&= 0
\end{aligned}$$

Dermed gælder altså

$$\begin{aligned}
\mathbf{E}\left(\frac{1}{X+Y}\right) &= 0,\\
\mathbf{Var}\left(\frac{1}{X+Y}\right) &= \mathbf{E}\left(\left(\frac{1}{X+Y}\right)^2\right)-\left(\mathbf{E}\left(\frac{1}{X+Y}\right)\right)^2 = 3-0^2 = 3
\end{aligned}$$

Slut på den vejledende besvarelse.