

Centrale begreber

- Bayes' formel

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^n P(B|A_j)P(A_j)}$$

- Binomialfordeling $\left(p(x) = \binom{n}{x} p^x (1-p)^{n-x} \right)$

- Poissonfordeling $\left(p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \right)$

- Definition på varians

$$\left(Var(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2 \right)$$

- Definition på kovarians

$$\left(Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - (E(X)E(Y)) \right)$$

- Definition på korrelation

$$corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- Regneregler for middelværdier

$$E(X) = \sum_{i=1}^n x_i p(x_i)$$

$$E(\psi(X)) = \sum_{i=1}^n \psi(x_i) p(x_i)$$

$$E(X + Y) = E(X) + E(Y)$$

$$E(aX) = aE(X)$$

- Regneregler for varianser

Hvis X og Y er uafhængige : $Var(X + Y) = Var(X) + Var(Y)$

Ex. C.1

A basketball player shoots 10 shots and the probability of hitting is 0.5 on each shot.

- (a) What is the probability of hitting eight shots?
 - (b) What is the probability of hitting eight shots if the probability on each shot is 0.6?
 - (c) What are the expected value and variance of the number of shots if $p = 0.5$?
- (a) $X \sim Bin(10; 0, 5)$ angiver antallet af point efter 10 kast.

$$\begin{aligned} P(X = 8) &= \binom{10}{8} 0,5^8 \cdot 0,5^{10-8} \\ &= \frac{10!}{8!(10-8)!} 0,5^{10} \\ &\approx \underline{\underline{0,0439}} = 4,39\% \end{aligned}$$

- (b) $Y \sim Bin(10; 0, 6)$ angiver antallet af point efter 10 kast.

$$\begin{aligned} P(Y = 8) &= \binom{10}{8} 0,6^8 \cdot 0,4^{10-8} \\ &= \frac{10!}{8!(10-8)!} 0,6^8 \cdot 0,4^2 \\ &\approx \underline{\underline{0,1209}} = 12,09\% \end{aligned}$$

- (c) Hvis $X \sim Bin(n, p)$ gælder, at $E(X) = np$ og $Var(X) = np(1 - p)$.
Derfor gælder for $X \sim Bin(10; 0, 05)$:

$$E(X) = 10 \cdot 0,05 = \underline{\underline{5}}, \quad Var(X) = 10 \cdot 0,05 \cdot (1 - 0,05) = \underline{\underline{2,5}}$$

Ex. C.2

Let X be a random variable with discrete pdf $f(x) = \frac{x}{8}$ if $x = \{1, 2, 5\}$, and zero otherwise.
Find:

- (a) $E(X)$
- (b) $Var(X)$
- (c) $E(2X + 3)$

$$(a) E(X) = \sum_{i=1}^n x_i p(x_i) = 1 \cdot \frac{1}{8} + 2 \cdot \frac{2}{8} + 5 \cdot \frac{5}{8} = \frac{1+4+25}{8} = \underline{\underline{3,75}}$$

$$(b) Var(X) = E(X^2) - (E(X))^2 = 1^2 \cdot \frac{1}{8} + 2^2 \cdot \frac{2}{8} + 5^2 \cdot \frac{5}{8} - 3,75^2 = \frac{1+8+125}{8} - 3,75^2 = \underline{\underline{2,6875}}$$

$$(c) E(2X + 3) = 2E(X) + 3 = 2 \cdot 3,75 + 3 = \underline{\underline{10,5}}$$

Ex. C.3

At a computer store, the annual demand for a particular software package is a random variable X . The store owner orders four copies of the package at \$10 per copy and charges customers \$35 per copy. At the end of the year the package is obsolete and the owner loses the investment on the unsold copies. The pdf of X is given by the following table

x	0	1	2	3	4
$f(x)$.1	.3	.3	.2	.1

- (a) Find $E(X)$.
- (b) Find $Var(X)$.
- (c) Express the owner's net profit Y as a linear function of X , and find $E(Y)$ and $Var(Y)$.

$$(a) E(X) = \sum_{i=1}^n x_i p(x_i) = 0 \cdot 0,1 + 1 \cdot 0,3 + 2 \cdot 0,2 + 3 \cdot 0,2 + 4 \cdot 0,1 = 0,3 + 0,6 + 0,6 + 0,4 = \underline{\underline{1,9}}$$

$$(b) Var(X) = E(X^2) - (E(X))^2 = 0^2 \cdot 0,1 + 1^2 \cdot 0,3 + 2^2 \cdot 0,2 + 3^2 \cdot 0,2 + 4^2 \cdot 0,1 - 1,9^2 = \underline{\underline{1,29}}$$

- (c) Ejeren af computerforretningen har en fast udgift på \$40 og tjener \$35 per solgte softwarepakke givet ved X .

Nettoprofitten Y er derfor givet ved:

$$\underline{\underline{Y = 35X - 40}}$$

$$E(Y) = E(35X - 40) = 35E(X) - 40 = 35 \cdot 1,9 - 40 = \underline{\underline{26,5}}$$

$$Var(Y) = Var(35X - 40) = 35^2 Var(X) = 1.225 \cdot 1,29 = \underline{\underline{1580.25}}$$

Husk at kvadrere en konstant når den sættes foran en varians:

$$(Var(aY)) = E((aY)^2) - (E(aY))^2 = E(a^2 Y^2) - (aE(Y))^2 = a^2(E(Y^2) - (E(Y))^2) = a^2 Var(Y)$$

Ex. C.4

The number of calls that arrive at a switchboard during one hour is Poisson distributed with mean $\mu = 10$. Find the probability of occurrence during an hour of each of the following events:

- (a) Exactly seven calls arrive.
- (b) At most seven calls arrive.
- (c) Between three and seven calls (inclusive) arrive.

- (a) Vi indfører den stokastiske variabel X som betegner antallet af opkald per time. X er poissonfordelt med parameter 10:

$$X \sim Poiss(10)$$

$$P(X = 7) = \frac{10^7}{7!} e^{-10} \approx \underline{\underline{0,0900}} = 9\%$$

(b)

$$P(X \leq 7) = \sum_{j=0}^7 \frac{10^j}{j!} e^{-10} \approx \underline{\underline{0,2202}} = 22,02\%$$

(c)

$$P(3 \leq X \leq 7) = \sum_{j=3}^7 \frac{10^j}{j!} e^{-10} \approx \underline{\underline{0,2175}} = 21,75\%$$

Opgave 3.2

Et panel af colasmagere på fem personer skal forsøge at skelne mellem to colamærker, som vi kan betegne med C og P . De 5 colasmagere får hver serveret et glas af den samme cola, idet man ved kast med en ærlig mønt har valgt cola P eller cola C .

Antag, at hver colasmager har sandsynlighed p for at bestemme colamærket korrekt. Det viser sig, at 4 ud af 5 gættede på cola P , mens 1 gættede på cola C . Hvad er den betingede sandsynlighed, givet dette resultat, for at det var cola C der blev serveret. Brug Bayes formel.

Bayes' formel giver mulighed for at bestemme en betinget sandsynlighed ud fra kendskab til sandsynligheden for den "omvendte" betingede sandsynlighed:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^n P(B|A_j)P(A_j)} \quad (1.4.5)$$

I opgaven kan vi definere en stokastisk variabel $X \in \{0, 1, \dots, 5\}$ som angiver antallet af colasmagere der gætter på cola P

Vi kan desuden definere en anden stokastisk variabel Y som betegner den serverede cola:

$$Y = \begin{cases} C & \text{hvis Cola} = C \\ P & \text{hvis Cola} = P \end{cases}$$

Den betingede sandsynlighed for at det var cola C der blev serveret givet at 4 colasmagere gættede på cola P er da:

$$P(Y = C|X = 4) = \frac{P(X = 4|Y = C)P(Y = C)}{P(X = 4|Y = C)P(Y = C) + P(X = 4|Y = P)P(Y = P)}$$

Sandsynligheden for at colasmagerne gætter (forkert) på cola P givet at cola C blev serveret er $1 - p$. Den stokastiske variabel X som angiver antallet af gæt på cola P , givet at den serverede cola er C , er binomialfordelt med antalsparameter $n = 5$ og sandsynlighedsparameter $p = 1 - p$.

Vi har derfor:

$$\begin{aligned} P(X = 4|Y = C) &= \binom{5}{4}(1-p)^4 \cdot (1-(1-p))^1 \\ &= 5p(1-p)^4 \end{aligned}$$

og

$$\begin{aligned} P(X = 4|Y = P) &= \binom{5}{4}(p)^4 \cdot (1-p)^1 \\ &= 5p^4(1-p) \end{aligned}$$

Nu har vi de nødvendige oplysninger for at benytte Bayes formel:

$$\begin{aligned} P(Y = C|X = 4) &= \frac{P(X = 4|Y = C)P(Y = C)}{P(X = 4|Y = C)P(Y = C) + P(X = 4|Y = P)P(Y = C)} \\ &= \frac{5p(1-p)^4 \cdot \frac{1}{2}}{(5p(1-p)^4 + 5p^4(1-p)) \cdot \frac{1}{2}} \\ &= \frac{(1-p)^4}{(1-p)^4 + p^3(1-p)} \\ &= \frac{(1-p)^3}{\underline{\underline{(1-p)^3 + p^3}}} \end{aligned}$$

Opgave 3.20

Angiv middelværdi, varians og spredning for

- (a) En stokastisk variabel (X) som er ligefordelt på $\{1, 2, 3, 4, 5, 6\}$.
- (b) Summen af øjnene i et slag med to terninger (Y).
- (c) En stokastisk variabel (Z) som er ligefordelt på $\{1, 2, \dots, n\}$.
Vink: $1 + 2 + \dots + n = \frac{1}{2}n(n+1)$ og $1^2 + 2^2 + \dots + n^2 = \frac{1}{6}n(2n+1)(n+1)$.

(a)

$$\begin{aligned} E(X) &= \frac{1+2+3+4+5+6}{6} = \underline{\underline{3,5}} \\ Var(X) &= \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} - 3,5^2 \approx \underline{\underline{2,92}} \\ Sd(X) &= \sqrt{Var(X)} \approx \underline{\underline{1,71}} \end{aligned}$$

(b)

$$\begin{aligned} E(Y) &= \frac{1 \cdot (2+12) + 2 \cdot (3+11) + 3 \cdot (4+10) + 4 \cdot (5+9) + 5 \cdot (6+8) + 6 \cdot 7}{36} = \frac{252}{36} = \underline{\underline{7}} \\ Var(Y) &= \frac{1 \cdot (2^2+12^2) + 2 \cdot (3^2+11^2) + 3 \cdot (4^2+10^2) + 4 \cdot (5^2+9^2) + 5 \cdot (6^2+8^2) + 6 \cdot 7^2}{36} - 7^2 \approx \underline{\underline{5,83}} \\ Sd(Y) &= \sqrt{Y} \approx \underline{\underline{2,41}} \end{aligned}$$

(c)

$$\begin{aligned}
 E(Z) &= \frac{1+2+\dots+n}{n} = \frac{\frac{1}{2}n(n+1)}{n} = \underline{\underline{\frac{1}{2}(n+1)}} \\
 Var(Z) &= \frac{1^2+2^2+\dots+n^2}{n} - \left(\frac{1}{2}(n+1)\right)^2 \\
 &= \frac{\frac{1}{6}n(2n+1)(n+1)}{n} - \left(\frac{1}{2}(n+1)\right)^2 \\
 &= \frac{1}{6}(2n^2+3n+1) - \frac{1}{4}(n^2+2n+1) \\
 &= \frac{1}{12}(4n^2+6n+2-3n^2-6n-3) \\
 &= \underline{\underline{\frac{1}{12}(n^2-1)}} \\
 Sd(Z) &= \sqrt{\frac{1}{12}(n^2-1)}
 \end{aligned}$$

Opgave 3.24

En stokastisk variabel X har middelværdi 5 og varians 2. Hvad er $E(7 + 8X + X^2)$?

$$\begin{aligned}
 E(7 + 8X + X^2) &= E(7) + E(8X) + E(X^2) \\
 &= 7 + 8E(X) + (Var(X) + (E(X))^2) \\
 &= 7 + 8 \cdot 5 + (2 + 5^2) \\
 &= \underline{\underline{74}}
 \end{aligned}$$

Opgave 3.27

Lad X_1, X_2 og X_3 være identisk fordelte, uafhængige stokastiske variable med strengt positiv varians.

Vis, at

$$corr(X_1 + X_2, X_2 + X_3) = \frac{1}{2}$$

Vi skal udlede

$$corr(X_1 + X_2, X_2 + X_3) = \frac{Cov(X_1 + X_2, X_2 + X_3)}{\sqrt{Var(X_1 + X_2)Var(X_2 + X_3)}}$$

Vi betragter først tælleren og benytter $Cov(X, Y) = E(XY) - E(X)E(Y)$ samt $E(X+Y) = E(X)+E(Y)$:

$$\begin{aligned}
 Cov(X_1 + X_2, X_2 + X_3) &= E\left((X_1 + X_2)(X_2 + X_3)\right) - E(X_1 + X_2)E(X_2 + X_3) \\
 &= E(X_1X_2 + X_1X_3 + X_2^2 + X_2X_3) - (E(X_1) + E(X_2))(E(X_2) + E(X_3)) \\
 &= E(X_1X_2) - E(X_1)E(X_2) + E(X_1X_3) - E(X_1)E(X_3) \\
 &\quad + E(X_2^2) - (E(X_2))^2 + E(X_2X_3) - E(X_2)E(X_3) \\
 &= Cov(X_1, X_2) + Cov(X_1, X_3) + Var(X_2) + Cov(X_2, X_3) \\
 &= Var(X_2) \quad (\text{da de er uafhængige er } Cov(X_i, X_j) = 0)
 \end{aligned}$$

I nævneren husker vi at de stokastiske variable er identisk fordelte $Var(X_1) = Var(X_2) = Var(X_3)$ og uafhængige $Var(X, Y) = Var(X) + Var(Y)$:

$$\begin{aligned}
 \sqrt{Var(X_1 + X_2)Var(X_2 + X_3)} &= \sqrt{(Var(X_1) + Var(X_2))(Var(X_2) + Var(X_3))} \\
 &= \sqrt{2Var(X_2) + 2Var(X_2)} \\
 &= 2Var(X_2)
 \end{aligned}$$

Dermed

$$\begin{aligned}
 corr(X_1 + X_2, X_2 + X_3) &= \frac{Var(X_2)}{2Var(X_2)} \\
 &\stackrel{\frac{1}{2}}{=}
 \end{aligned}$$

Opgave 1: Cykelforsikringer

Antag at du har en cykel til en værdi af 4000kr. Sandsynligheden for at cyklen bliver stjålet i løbet af et år er 5%. Du har mulighed for at tegne en forsikring, således at du får erstattet din cykel med det fulde beløb, hvis den bliver stjålet.

- Angiv (uden at regne) hvor meget du er villig til at betale for en sådan forsikring.

Hvis man regner med at 5% af cyklens værdi forsvinder ved tyveri hvert år er man nok villig til at betale

$$0,05 \times 4.000kr. = 200kr.$$

om året for fuld forsikring.

- Lad X være en stokastisk variabel som angiver værdien af cyklen i tilfældet hvor der ikke er tegnet en forsikring. Opstil udfaldene for X , angiv sandsynhederne og udregn middelværdien af X .

X er lig 4.000 med 95% sandsynlighed og 0 med 5% sandsynlighed:

$$E(X) = 4.000 \cdot 0,95 + 0 \cdot 0,05 = \underline{\underline{3.800}}$$

3. Antag nu at man kan tegne en forsikring af cyklen. Forsikringen koster 400kr. Lad Y være en stokastisk variabel som angiver værdien af cyklen minus udgifter til forsikring. Opstil udfaldene for Y , angiv sandsynlighederne og udregn middelværdien af Y .

Ligemeget om cyklen bliver stjålet eller ej har man en cykel til en værdi af 4.000kr. og man har betalt 400kr. for forsikring:

$$E(Y) = (4.000 - 400) \cdot 0,95 + (4.000 - 400) \cdot 0,05 = \underline{\underline{3.600}}$$

4. Antag at der findes en anden type forsikring. Nemlig en forsikring, hvor der er en selvisiko på 1.000kr. Prisen på denne forsikring er 150kr. Lad Z være en stokastisk variabel som angiver cyklens værdi minus udgifter til denne forsikring. Opstil udfaldene for Z , angiv sandsynlighederne og udregn middelværdien af Z .

$$Z \in \{2850, 3850\} : \quad P(Z = 3850) = 0,95 , \quad P(Z = 2850) = 0,05$$

$$E(Z) = 3850 \cdot 0,95 + 2850 \cdot 0,05 = 3657,5 + 142,5 = \underline{\underline{3800}}$$

5. Sammenligne middelværdierne for X , Y og Z og diskutér fordele og ulemper ved at benytte middelværdien når man skal vælge forsikring.

Vi har $E(X) = E(Z) = 3.800$ og $E(Y) = 3.600$ så Y er éntydigt dårligere, hvis man benytter middelværdi som udvælgelseskriterium.

Variansen af Y er dog 0, mens især X har stor varians og derved stor risiko. Valg af forsikring vil derfor ikke kun bero på forventet værdi (middelværdi) men også på forsikringstagers risikopræferencer.

Tegn evt. en konkav nyttefunktion og plot de forskellige udfald af X , Y og Z . Det er nemt at se at graden af konkavitet bestemmer hvor meget risiko man er villig til at påtage.

6. Antag i stedet at en person træffer sit valg vedr. forsikring på baggrund af en nyttefunktion. Vi antager at nyttefunktionen er givet ved:

$$u(v) = 10v - 0,001v^2, \quad v \in (0, 4000)$$

Skitsér funktionen. Udregn $E(u(X))$, $E(u(Y))$ og $E(u(Z))$. Hvis personen benytter denne nyttefunktion vil han/hun så vælge at forsikre sig og hvilken type forsikring foretrækkes?

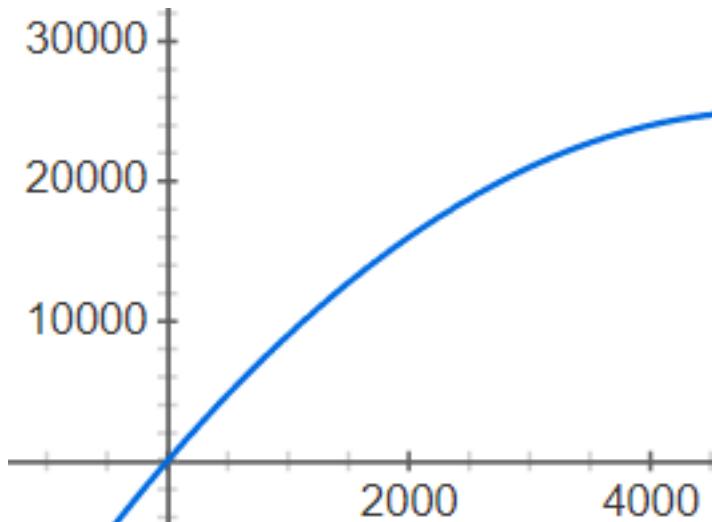
For at kunne skitsere funktionen skal kende første og anden afledte:

$$u'(v) = 10 - 0,002v , \quad u''(v) = -0,002$$

Funktionen er altså strengt konkav og et ekstremum vil være et globalt maksimum. Det globale maksimum findes hvor $u'(v) = 0$:

$$0 = 10 - 0,002v \Leftrightarrow v = 5.000$$

Funktionen er altså strengt voksende med aftagende hældning på definitionsintervallet $(0, 4000)$.



Middelværdier:

$$\begin{aligned} E(u(X)) &= 0,95(10 \cdot 4.000 - 0,001 \cdot 4.000^2) + 0,05(10 \cdot 0 - 0,001 \cdot 0^2) \\ &= 38.000 - 15.200 \\ &= \underline{\underline{22.800}} \end{aligned}$$

$$\begin{aligned} E(u(Y)) &= 1 \cdot (10 \cdot 3.600 - 0,001 \cdot 3.600^2) \\ &= 36.000 - 12.960 \\ &= \underline{\underline{23.040}} \end{aligned}$$

$$\begin{aligned} E(u(Z)) &= 0,95(10 \cdot 3.850 - 0,001 \cdot 3.850^2) + 0,05(10 \cdot 2.850 - 0,001 \cdot 2.850^2) \\ &= 22.493,625 - 1.018,875 \\ &= \underline{\underline{23.512,50}} \end{aligned}$$

Personen med ovenstående nyttefunktionen vil altså vælge at forsikre sig med selvrisko.

7. Vis at, $E(u(W)) = 10\mu - 0,001\mu^2 - 0,001\sigma^2$, hvor $\mu = E(W)$ og $\sigma^2 = Var(W)$. Giv en fortolkning af formlen.

$$\begin{aligned} E(u(W)) &= E(10W - 0,001W^2) \\ &= 10E(W) - 0,001E(W^2) \\ &= 10E(W) - 0,001E(W^2) + 0,001(E(W))^2 - 0,001(E(W))^2 \\ &= 10E(W) - 0,001(E(W))^2 - \underbrace{(0,001E(W^2) - 0,001(E(W))^2)}_{Var(W)} \\ &= 10\mu - 0,001\mu^2 - 0,001\sigma^2 \end{aligned}$$

En fortolkning af formlen er at forbrugerens nytte afhænger positivt af godet (med aftagende marginalnytte) indtil $10\mu = 0,001\mu^2 \Leftrightarrow \mu = 10.000$ mens varians er éntydigt negativ for nytten.

8. Udregn variansen af X, Y og Z . Forklar i relation til spørgsmål 7 hvilken forsikring som foretrækkes

Vi beregner varianserne af X , Y og Z og indsætter derefter den forventede nytte.

$$\begin{aligned}Var(X) &= E(E(X) - x_i)^2, \quad i = 1, 2 \\&= 0,95(3.800 - 4.000)^2 + 0,05(3.800 - 0)^2 \\&= \underline{\underline{760.000}} \\Var(Y) &= E(E(Y) - y_i)^2, \quad i = 1, 2 \\&= 0,95(3.600 - 3.600)^2 + 0,05(3.600 - 3.600)^2 \\&= \underline{\underline{0}} \\Var(Z) &= E(E(Z) - z_i)^2, \quad i = 1, 2 \\&= 0,95(3.800 - 3.850)^2 + 0,05(3.800 - 2.850)^2 \\&= \underline{\underline{47.500}}\end{aligned}$$

$$\begin{aligned}E(u(X)) &= 10 \cdot 3.800 - 0,001 \cdot 3.800^2 - 0,001 \cdot 760.000 \\&= \underline{\underline{22.800}} \\E(u(Y)) &= 10 \cdot 3.600 - 0,001 \cdot 3.600^2 \\&= \underline{\underline{23.040}} \\E(u(Z)) &= 10 \cdot 3.800 - 0,001 \cdot 3.800^2 - 0,001 \cdot 47.500 \\&= \underline{\underline{23.512,5}}\end{aligned}$$

Forsikringen med selvrisiko og lav præmie foretrækkes altså hvilket ikke er overraskende da vi har beregnet at høj varians mindsker nytten.

Opgave 2: Cykelforsikringer

Antag nu i stedet at du udbyder cykelforsikringer. I tilfældet af at en forsikringstager får stjålet en cykel udbetales 4.000kr. Sandsynligheden for at cyklen bliver stjålet i løbet af et år er 5%. Antag at hver forsikringstager maksimalt kan få stjålet en cykel om året. Prisen for en cykelforsikring er 400kr.

1. *Antag at der er 10 som har tegnet forsikring, og at sandsynligheden for at hver af disse personer får stjålet en cykel kan betragtes som uafhængige med den samme sandsynlighed. Lad Y være en stokastisk variabel som angiver antallet af cykler som bliver stjålet blandt de 10 forsikringstagere. Angiv fordelingen af Y og argumentér for dit valg. Diskutér om antagelserne er realistiske.*

Vi har uafhængige gentageser af et binært udfald med konstant sandsynlighed. Y er derfor binomialfordelt med antalsparameter $n = 10$ og sandsynlighedsparameter $p = 0,05$:

$$Y \sim Bin(10; 0,05)$$

Antagelserne om uafhængighed og identisk fordeling (*iid*) er problematiske. Hvis alle cykler så ens ud og stod samme sted (f.eks. på hovedbanegården) kunne man argumentere for konstant sandsynlighed. Men i dette tilfælde er uafhængighedsantagelsen ikke relistisk. Omvendt hvis cyklerne stod forskellige steder.

2. Udregn det forventede antal stjålne cykler blandt forsikringstagerne. Udregn den forventede udgift i forbindelse med udbetaling af erstatninger. Angiv hvilke "regneregler" som anvendes. Udregn de forventede indtægter.

Middelværdien af en stokastisk variabel som er fordelt med antalsparameter n og sandsynlighedsparameter p er

$$E(Y) = n \cdot p$$

Det forventede antal stjålne cykler er derfor

$$E(Y) = 10 \cdot 0,05 = \underline{\underline{0,5}}$$

Forventet udgift (U):

$$E(U) = E(4.000 \cdot Y) = 4.000 \cdot E(Y) = \underline{\underline{2.000}}$$

Forventet indtægt (I):

$$E(I) = E(10 \cdot 400) = \underline{\underline{4.000}}$$

Regneregel:

$$a \text{ er en konstant, } X \text{ er en stokastisk variabel : } E(aX) = aE(X)$$

3. Hvis der blandt de 10 forsikringstagere bliver stjålet mere end én cykel vil udgifterne overstige indtægterne. Udregn sandsynligheden for dette $P(Y > 1)$. Angiv hvilke regneregler du bruger.

$$\begin{aligned} P(Y > 1) &= 1 - P(Y \leq 1) \\ &= 1 - \binom{10}{0} 0,05^0 \cdot 0,95^{10} - \binom{10}{1} 0,05^1 \cdot 0,95^9 \\ &\approx \underline{\underline{0,0861}} = 8,61\% \end{aligned}$$

Regneregel:

$$P(A) = 1 - P(E \setminus A)$$

4. Antag i stedet at der er 100 som har tegnet en forsikring. Udregn de forventede indtægter og udgifter

Nu gælder:

$$Y \sim Bin(100; 0,05)$$

Det forventede antal stjålne cykler er derfor

$$E(Y) = 100 \cdot 0,05 = \underline{\underline{5}}$$

Forventet udgift (U):

$$E(U) = E(4.000 \cdot Y) = 4.000 \cdot E(Y) = \underline{\underline{20.000}}$$

Forventet indtægt (I):

$$E(I) = E(100 \cdot 400) = \underline{\underline{40.000}}$$

5. Udregn sandsynligheden for at de faktiske udgifter overstiger indtægterne hvis der er 100 forsikringstagere (Anvend Poissonfordelingen som approksimation og husk at tjekke forudsætningerne for at approksimationen er i orden).

Når antalsparameteren n går stiger til et højt tal i en binomialfordeling kan vi benytte følgende approksimation:

$$\lim_{n \rightarrow \infty} \binom{n}{x} p_n^x (1 - p_n)^{n-x} = \frac{\lambda^x}{x!} e^{-\lambda} \quad (4.1.1)$$

så længe det gælder, at $np_n \rightarrow \lambda \in [0, \infty)$ for $n \rightarrow \infty$.

I vores tilfælde er $\lambda = n \cdot p = 100 \cdot 0,05 = 5$ så når vi approksimerer med Poissonfordelingen er Y Poissonfordelt med parameter $\lambda = 5$:

$$Y \sim Poiss(5)$$

Sansynlighed for at udgifter overstiger indtægter er da:

$$\begin{aligned} P(U > I) &= P(4.000 \cdot Y > 40.000) \\ &= P(Y > 10) \\ &= 1 - P(Y \leq 10) \\ &= 1 - \sum_{i=0}^{10} \left(\frac{5^i}{i!} e^{-5} \right) \quad (\text{brug POISSON.DIST i Excel}) \\ &\approx \underline{\underline{0,0137}} = 1,37\% \end{aligned}$$

6. Antag nu at antallet af forsikringstagere er 200. Udregn sandsynligheden for at de faktiske udgifter overstiger indtægterne.

Y er nu poissonfordelt med parameter $\lambda = n \cdot p = 200 \cdot 0,05$:

$$Y \sim Poiss(10)$$

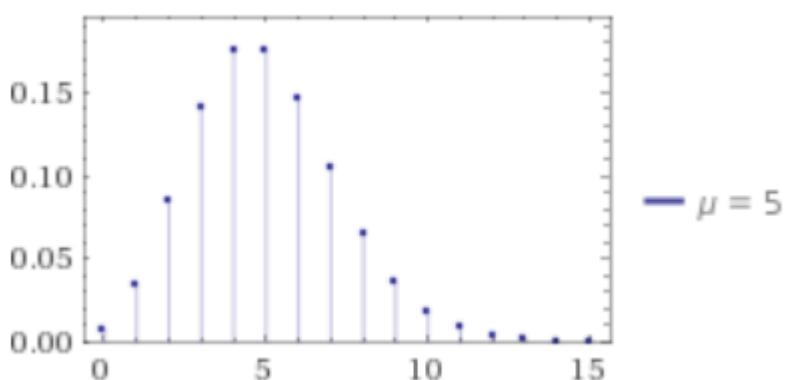
Indtægter er nu $I = 200 \cdot 400 = 80.000$. Sandsynligheden for at udgifter overstiger indtægter:

$$\begin{aligned} P(U > I) &= P(4.000 \cdot Y > 80.000) \\ &= P(Y > 20) \\ &= 1 - P(Y \leq 20) \\ &= 1 - \sum_{i=0}^{20} \left(\frac{10^i}{i!} e^{-10} \right) \quad (\text{brug POISSON.DIST i Excel}) \\ &\approx \underline{\underline{0,0016}} = 0,16\% \end{aligned}$$

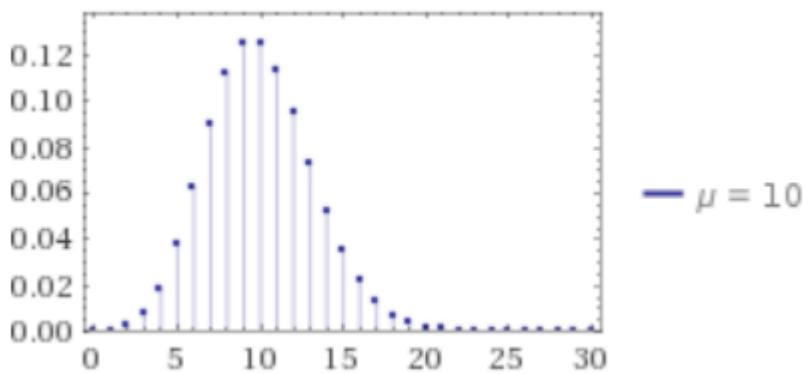
7. Forklar i ord dine resultater fra spørgsmål 3, 5 og 6.

Betrægt disse to grafer af Poissonfordelinger med parameter 5, 10 og 100.

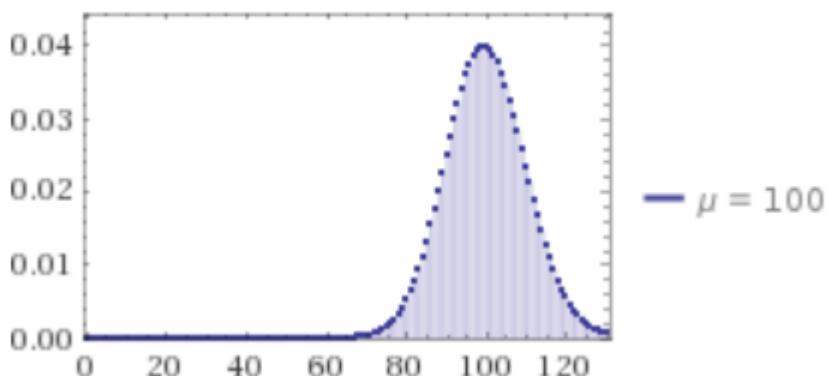
Plot of PDF:



Plot of PDF:



Plot of PDF:



Jo højere den forventede værdi λ er jo mere ”hale” er der til venstre i fordelingsfunktionen. Derfor er der også mere sandsynlighedsmasse i starten af fordelingen.

Opgave 4.6

En terning kastes indtil den første sekser opnås. Lad X betegne antallet af kast før en sekser opnås.

Hvad er sandsynligheden for at $X < 6$?

Hændelsen $X < 6$ er lig med at man får en sekser i det første, andet, tredje, fjerde, femte eller sjette kast.

Den sandsynlighed er nem at beregne som en sum:

$$\begin{aligned}
 P(X < 6) &= \frac{1}{6} + \frac{5}{6} \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^2 \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^3 \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^4 \cdot \frac{1}{6} + \left(\frac{5}{6}\right)^5 \cdot \frac{1}{6} \\
 &= \frac{1}{6} \sum_{i=0}^5 \left(\frac{5}{6}\right)^i \\
 &= \frac{1}{6} \frac{1 - (\frac{5}{6})^6}{1 - \frac{5}{6}} \quad (\text{geometrisk række med endeligt antal}) \\
 &= 1 - \left(\frac{5}{6}\right)^6 \approx 0,6651 = 66,51\%
 \end{aligned}$$

Der findes dog også en decideret fordeling som angiver ”ventetiden” på et gunstigt udfald i et binært udfaldsrum, nemlig *den geometriske fordeling*:

$$p(x) = (1 - \theta)^x \theta, \quad x \in \mathbb{N}_0 \quad (4.1.9)$$

hvor x er antal gentagelser (eller ventetiden) inden det gunstige udfald indtræffer og θ er sandsynligheden for det gunstige udfald ved hver (uafhængige) gentagelse.

Fordelingsfunktionen for den geometriske fordeling er:

$$\begin{aligned}
 F_X(x) = P(X \leq x) &= \sum_{i=1}^x (1 - \theta)^i \theta \\
 &= \theta \frac{1 - (1 - \theta)^{x+1}}{1 - (1 - \theta)} \\
 &= 1 - (1 - \theta)^{x+1}
 \end{aligned}$$

Og ved hjælp af denne kan vi også beregne sandsynligheden for at ventetiden er mindre end seks:

$$\begin{aligned}
 P(X < 6) &= F_X(5) \\
 &= 1 - \left(1 - \frac{1}{6}\right)^6 \\
 &= 1 - \left(\frac{5}{6}\right)^6
 \end{aligned}$$

Hvad er den største værdi af $i \in \mathbb{N}$ for hvilken $P(X > i) \geq \frac{1}{2}$?

Her bruger vi igen fordelingsfunktionen (og logaritmeregnerregler):

$$\begin{aligned}
 P(X > i) &\geq \frac{1}{2} \Leftrightarrow \\
 1 - P(X \leq i) &\geq \frac{1}{2} \Leftrightarrow \\
 1 - F_X(x) &\geq \frac{1}{2} \Leftrightarrow \\
 \left(\frac{5}{6}\right)^{i+1} &\geq \frac{1}{2} \Leftrightarrow \\
 (i+1)\ln\left(\frac{5}{6}\right) &\geq \ln\left(\frac{1}{2}\right) \Leftrightarrow \\
 i &\leq \frac{\ln\left(\frac{1}{2}\right)}{\ln\left(\frac{5}{6}\right)} - 1 \approx 2,80
 \end{aligned}$$

i er et helt tal, så den største værdi i kan antage er

$$\underline{\underline{i = 2}}$$

Til og med en ventetid på 2 er sandsynligheden for en højere ventetid altså over 50%.

Opgave 4.14

Lad X være Poissonfordelt med parameter λ .

Hvad er $E(2^X)$?

Ved hjælp af formlen for middelværdi af en transformerede stokastisk variabel

$$E(\psi(X)) = \sum_{i=1}^n \psi(x_i) p(x_i) \tag{4.4.4}$$

kan vi opstille udtrykket

$$\begin{aligned}
 E(2^x) &= \sum_{x=0}^{\infty} 2^x \frac{\lambda^x}{x!} e^{-\lambda} \\
 &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(2\lambda)^x}{x!} \quad \left(\text{benyt: } e^y = \sum_{n=0}^{\infty} \frac{y^n}{n!}\right) \\
 &= e^{-\lambda} e^{2\lambda} \\
 &= \underline{\underline{e^{\lambda}}}
 \end{aligned}$$

Hvad er $E((1 + X)^{-1})$?

$$\begin{aligned}
 E((1+x)^{-1}) &= \sum_{x=0}^{\infty} \frac{1}{x+1} \frac{\lambda^x}{x!} e^{-\lambda} \\
 &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{(x+1)!} \\
 &= \frac{1}{\lambda} e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x+1}}{(x+1)!} \\
 &= \frac{1}{\lambda} e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{x!} \\
 &= \frac{1}{\lambda} e^{-\lambda} \left(\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} - 1 \right) \\
 &= \frac{1}{\lambda} e^{-\lambda} (e^\lambda - 1) \\
 &= \underline{\underline{\frac{1}{\lambda} (1 - e^{-\lambda})}}
 \end{aligned}$$

Opgave H (Aflevering uge 40)

I det følgende ser vi på husstande, hvor der er 2 voksne, som er en kvinde og en mand. I den enkelte husstand har kvinden og manden hver især en arbejdstid, som er antallet af timer, de arbejder mod betaling udenfor hjemmet. Vi antager, at de begge har mulighed for at arbejde 0 timer (ikke arbejde), 20 timer (halv tid) og 40 timer (fuld tid) pr. uge.

Vi betragter nu en tilfældig husstand og lader de stokastiske variable X_k og X_m være defineret på følgende måde:

X_k : Kvindens arbejdstid målt i timer pr. uge

X_m : Mandens arbejdstid målt i timer pr. uge

Den simultane fordeling af kvindens og mandens arbejdstid er givet ved:

		x_m		
		0	20	40
x_k	0	0,01	0,00	0,10
	20	0,02	0,06	0,20
	40	0,05	0,06	0,50

- Angiv de marginale fordelinger af X_k og X_m .

X_k :

$$P(X_k = 0) = 0,01 + 0,10 = \underline{\underline{0,11}}$$

$$P(X_k = 20) = 0,02 + 0,06 + 0,20 = \underline{\underline{0,28}}$$

$$P(X_k = 40) = 0,05 + 0,06 + 0,50 = \underline{\underline{0,61}}$$

X_m :

$$\begin{aligned} P(X_m = 0) &= 0,01 + 0,02 + 0,05 = \underline{\underline{0,08}} \\ P(X_m = 20) &= 0,06 + 0,06 = \underline{\underline{0,12}} \\ P(X_m = 40) &= 0,10 + 0,20 + 0,50 = \underline{\underline{0,80}} \end{aligned}$$

2. Udregn middelværdi og varians af X_k og X_m .

X_k

$$\begin{aligned} E(X_k) &= 20 \cdot 0,28 + 40 \cdot 0,61 = \underline{\underline{30}} \\ Var(X_k) &= 20^2 \cdot 0,28 + 40^2 \cdot 0,61 - 30^2 = \underline{\underline{188}} \end{aligned}$$

X_m

$$\begin{aligned} E(X_m) &= 20 \cdot 0,12 + 40 \cdot 0,80 = \underline{\underline{34,4}} \\ Var(X_m) &= 20^2 \cdot 0,12 + 40^2 \cdot 0,80 - 34,4^2 = \underline{\underline{144,64}} \end{aligned}$$

3. Udregn korrelationen mellem X_k og X_m .

$$\begin{aligned} Cov(X_k, X_m) &= E(X_k X_m) - E(X_k)E(X_m) \\ &= 20 \cdot 20 \cdot 0,06 + 20 \cdot 40 \cdot 0,20 + 40 \cdot 20 \cdot 0,06 + 40 \cdot 40 \cdot 0,50 - 30 \cdot 34,4 \\ &= 1032 - 1032 \\ &= 0 \end{aligned}$$

$$\begin{aligned} Corr(X_k, X_m) &= \frac{Cov(X_k, X_m)}{\sqrt{Var(X_k)Var(X_m)}} \\ &= \underline{0} \end{aligned}$$

4. Angiv de betingede fordelinger af kvindens arbejdstid givet, at mandens arbejdstid er henholdsvis 0, 20 og 40 timer pr. uge.

$$\begin{aligned}
 P(X_k = 0|X_m = 0) &= \frac{P(X_k = 0, X_m = 0)}{P(X_m = 0)} = \frac{0,01}{0,08} = \underline{\underline{0,125}} \\
 P(X_k = 20|X_m = 0) &= \frac{P(X_k = 20, X_m = 0)}{P(X_m = 0)} = \frac{0,02}{0,08} = \underline{\underline{0,25}} \\
 P(X_k = 40|X_m = 0) &= \frac{P(X_k = 40, X_m = 0)}{P(X_m = 0)} = \frac{0,05}{0,08} = \underline{\underline{0,625}} \\
 P(X_k = 0|X_m = 20) &= \frac{P(X_k = 0, X_m = 20)}{P(X_m = 20)} = \frac{0,00}{0,12} = \underline{0} \\
 P(X_k = 20|X_m = 20) &= \frac{P(X_k = 20, X_m = 20)}{P(X_m = 20)} = \frac{0,06}{0,12} = \underline{\underline{0,50}} \\
 P(X_k = 40|X_m = 20) &= \frac{P(X_k = 40, X_m = 20)}{P(X_m = 20)} = \frac{0,06}{0,12} = \underline{\underline{0,50}} \\
 P(X_k = 0|X_m = 40) &= \frac{P(X_k = 0, X_m = 40)}{P(X_m = 40)} = \frac{0,10}{0,80} = \underline{\underline{0,125}} \\
 P(X_k = 20|X_m = 40) &= \frac{P(X_k = 20, X_m = 40)}{P(X_m = 40)} = \frac{0,20}{0,80} = \underline{\underline{0,25}} \\
 P(X_k = 40|X_m = 40) &= \frac{P(X_k = 40, X_m = 40)}{P(X_m = 40)} = \frac{0,50}{0,80} = \underline{\underline{\underline{0,625}}}
 \end{aligned}$$

5. Er X_k og X_m uafhængige?

Hvis X_k og X_m er uafhængige gælder

$$\forall x_k, x_m \in \{0, 20, 40\} : P(X_k = x_k, X_m = x_m) = P(X_k = x_k) \cdot P(X_m = x_m).$$

Det vil sige at X_k og X_m er uafhængige hvis, og kun hvis, det for alle simultane sandsynligheder gælder, at den simultane sandsynlighed er lig med produktet af de marginale sandsyigheder.
Fra de marginale fordelinger ved vi:

$$P(X_k = 20) = 0,28 \text{ & } P(X_m = 0,12).$$

Da

$$P(X_k = 20) \cdot P(X_m = 20) = 0,0336 \neq P(X_k = 20, X_m = 20) = 0,06$$

er X_k og X_m ikke uafhængige.

6. En ny stokastisk variabel defineres som $\tilde{X} = \frac{X_k + X_m}{2}$.

(a) Hvad angiver den stokastiske variabel \tilde{X} ?

(b) Udregn middelværdi og varians af \tilde{X} .

(a) \tilde{X} angiver gennemsnittet mellem X_k og X_m .

(b)

$$\begin{aligned}
 E(\tilde{X}) &= E\left(\frac{X_k + X_m}{2}\right) = \frac{1}{2}(E(X_k) + E(X_m)) = \frac{1}{2}(15 + 17, 2) = \underline{\underline{32, 2}} \\
 Var(\tilde{X}) &= Var\left(\frac{X_k + X_m}{2}\right) \\
 &= \left(\frac{1}{2}\right)^2 Var(X_k) + Var(X_m) + 2Cov(X_k, X_m) \\
 m &= \left(\frac{1}{2}\right)^2 Var(X_k) + Var(X_m) \\
 &= \frac{1}{4}(188 + 144, 64) = \underline{\underline{83, 16}}
 \end{aligned}$$

7. Vi vil nu se på forholdet mellem kvindens og mandens fritid. Vi definerer nye stokastiske variable $Y_k = 168 - X_k$ og $Y_m = 168 - X_m$, som angiver henholdsvis kvindens og mandens fritid. Fritid er her defineret som antallet af timer pr. uge, der ikke bruges på lønnet arbejde udenfor hjemmet. Vi vil nu se på den stokastiske variabel Z , som er defineret som:

$$Z = \frac{Y_k}{Y_m} = \frac{168 - X_k}{168 - X_m}$$

- (a) Hvad angiver variablen Z ?
(b) Udregn middelværdien af Z og giv en fortolkning af resultatet.

- (a) Z angiver forholdet mellem kvindens og mandens fritid. Hvis Z er over 1 har kvinden mere fritid end manden og omvendt hvis Z er under 1.
(b) $E(Z)$ kan ikke beregnes som

$$E(Z) = \frac{168 - E(Y_K)}{168 - E(Y_m)}$$

da $E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$. Dette gælder approksimativt og benyttes i stikprøveteori, men er ikke eksakt.

Middelværdien af Z kan i stedet beregnes udfra vores viden om de simultane udfald af X_k og X_m .

Z kan antage følgende 9 værdier:

$$\begin{aligned}
 \frac{168}{168} &\text{ ved } (X_k, X_m) = (0, 0) \\
 \frac{148}{168} &\text{ ved } (X_k, X_m) = (20, 0) \\
 \frac{128}{168} &\text{ ved } (X_k, X_m) = (40, 0) \\
 \frac{168}{148} &\text{ ved } (X_k, X_m) = (0, 20) \\
 \frac{148}{148} &\text{ ved } (X_k, X_m) = (20, 20) \\
 \frac{128}{148} &\text{ ved } (X_k, X_m) = (40, 20) \\
 \frac{168}{128} &\text{ ved } (X_k, X_m) = (0, 40) \\
 \frac{148}{128} &\text{ ved } (X_k, X_m) = (20, 40) \\
 \frac{128}{128} &\text{ ved } (X_k, X_m) = (40, 40)
 \end{aligned}$$

Da vi kender sandsynligheden for de simultane udfald af (X_k) og (X_m) kan vi beregne mid-delværdien af Z .

$$\begin{aligned}E(Z) &= \frac{168}{168} \cdot 0,01 + \frac{148}{168} \cdot 0,02 + \frac{128}{168} \cdot 0,05 + \frac{148}{148} \cdot 0,06 \frac{128}{148} \cdot 0,06 + \\&+ \frac{168}{128} \cdot 0,10 + \frac{148}{128} \cdot 0,20 + \frac{128}{128} \cdot 0,50 \\&\approx \underline{\underline{1,04}}\end{aligned}$$

Kvinder har altså ca. 4% mere fritid end mænd.

Centrale begreber

- Ligefordelingen på (a, b) $\left(p(x) = \frac{1_{(a,b)}(x)}{b-a}\right)$
- Eksponentialfordelingen $\left(p(x) = \lambda e^{-\lambda x}\right)$
- Normalfordelingen $\left(p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)$
- Fordelingsfunktion for en kontinuert fordeling:

$$F(x) = \int_{-\infty}^x p(y) dy$$

- Fordeling af $Y = t(X)$:

$$q(y) = \begin{cases} p(t^{-1}(y)) |\frac{d}{dy}t^{-1}(y)| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.1)$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y))/|t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.2)$$

- Middelværdi af en kontinuert stokastisk variabel:

$$E(X) = \int_{-\infty}^{\infty} xp(x) dx \quad (5.2.2)$$

- Varians af en kontinuert stokastisk variabel:

$$Var(X) = \int_{-\infty}^{\infty} x^2 p(x) dx - \left(\int_{-\infty}^{\infty} xp(x) dx \right)^2$$

- Middelværdi af en transformerede stokastiske variabel $t(X)$ som er koncentreret på intervallet I :

$$E(t(X)) = \int_I t(x)p(x) dx \quad (5.2.4)$$

hvor p er sandsynlighedstætheden for X .

Opgave 5.2

Antag at X er en kontinuert stokastisk variabel på $(1, \infty)$ med sandsynlighedstæthed $p(x) = \alpha x^{-(\alpha+1)}$ for $x > 1$, hvor $\alpha > 0$. Find fordelingsfunktionen for X

Definitionsintervallet for X kan indarbejdes i sandsynlighedstæthedens med en indikatorfunktion på følgende måde:

$$p(x) = \alpha x^{-(\alpha+1)} 1_{(1, \infty)}(x)$$

Fordelingsfunktionen er da

$$\begin{aligned} F(x) &= \int_{-\infty}^x \alpha x^{-(\alpha+1)} 1_{(1, \infty)}(x) dx = \begin{cases} \int_1^x \alpha x^{-(\alpha+1)} dx \text{ for } x \geq 1 \\ \int_{-\infty}^x 0 dx \text{ for } x < 1 \end{cases} \\ &= \begin{cases} \left[-x^{-\alpha} \right]_1^x \text{ for } x \geq 1 \\ 0 \text{ for } x < 1 \end{cases} \\ &= \begin{cases} 1 - x^{-\alpha} \text{ for } x \geq 1 \\ 0 \text{ for } x < 1 \end{cases} \end{aligned}$$

Tænk over hvorfor X først er defineret for tal højere eller lig med 1.

Opgave 5.3

Lad X være en stokastisk variabel med fordelingsfunktionen

$$F(x) = \begin{cases} 0 \text{ for } x \leq 0 \\ \frac{x}{3} \text{ for } 0 < x \leq 1 \\ \frac{(2x-1)}{3} \text{ for } 1 < x \leq 2 \\ 1 \text{ for } x > 2 \end{cases}$$

Find $P\left(\frac{1}{2} < X < 1\right)$, $P\left(1 \leq X < \frac{3}{2}\right)$ og $P\left(\frac{2}{3} < X \leq \frac{4}{3}\right)$. Gør rede for, at X er kontinuert, og find sandsynlighedstæden for X .

$$\begin{aligned}
 P\left(\frac{1}{2} < X < 1\right) &= F(1) - F\left(\frac{1}{2}\right) \\
 &= \frac{1}{3} - \frac{\frac{1}{2}}{3} \\
 &= \frac{1}{6} \\
 &\underline{\underline{=}}
 \end{aligned}$$

$$\begin{aligned}
 P\left(1 \leq X \leq \frac{3}{2}\right) &= F\left(\frac{3}{2}\right) - F(1) \\
 &= \frac{2 \cdot \frac{3}{2} - 1}{3} - \frac{1}{3} \\
 &= \frac{1}{3}
 \end{aligned}$$

$$\begin{aligned}
 P\left(\frac{2}{3} < X \leq \frac{4}{3}\right) &= P\left(\frac{2}{3} < X \leq 1 \cup 1 < X \leq \frac{4}{3}\right) \\
 &= F(1) - F\left(\frac{2}{3}\right) + F\left(\frac{4}{3}\right) - F(1) \\
 &= F\left(\frac{4}{3}\right) - F\left(\frac{2}{3}\right) \\
 &= \frac{2 \cdot \frac{4}{3} - 1}{3} - \frac{\frac{2}{2}}{3} \\
 &= \frac{1}{3}
 \end{aligned}$$

Kontinuitet for en afbildung $f : D \rightarrow R$ med $x_0 \in D$ er:

For ethvert $\varepsilon > 0$ eksisterer der et $\delta > 0$ så $|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$
eller
 $\lim_{x \rightarrow c^+} f(x) = \lim_{x \rightarrow c^-} f(x)$

De relevante punkter at undersøge for kontinuitet er de punkter, hvor funktionen skifter forskrift.
For vores funktion, $F(x)$, gælder:

$$\begin{aligned}
 \lim_{x \rightarrow 0^-} f(x) &= 0 = \lim_{x \rightarrow 0^+} f(x) \\
 \lim_{x \rightarrow 1^-} f(x) &= \frac{1}{3} = \lim_{x \rightarrow 1^+} f(x) \\
 \lim_{x \rightarrow 2^-} f(x) &= 1 = \lim_{x \rightarrow 2^+} f(x)
 \end{aligned}$$

$F(x)$ er altså kontinuert.

Sandsynlighedstætheden for X er:

$$p(x) = F'(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{1}{3} & \text{for } 0 < x \leq 1 \\ \frac{2}{3} & \text{for } 1 < x \leq 2 \\ 0 & \text{for } x > 2 \end{cases}$$

Opgave 5.7

Vis, at fordelingen med sandsynlighedstæthed givet ved

$$p(x) = \beta x^{\beta-1}, \quad x \in (0, 1), \quad \beta > 0$$

har middelværdi $\frac{\beta}{\beta+1}$. Hvad er variansen?

Middelværdi:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xp(x)1_{(0,1)}(x)dx \\ &= \int_0^1 xp(x)dx \\ &= \int_0^1 x\beta x^{\beta-1}dx \\ &= \int_0^1 \beta x^{\beta}dx \\ &= \left[\frac{\beta}{\beta+1} x^{\beta+1} \right]_0^1 \\ &= \frac{\beta}{\beta+1} \end{aligned}$$

Varians:

$$\begin{aligned} Var(X) &= \int_{-\infty}^{\infty} x^2 p(x)1_{(0,1)}(x)dx - \left(\frac{\beta}{\beta+1} \right)^2 \\ &= \int_0^1 \beta x^{\beta+1}dx - \left(\frac{\beta}{\beta+1} \right)^2 \\ &= \frac{\beta}{\beta+2} \left[x^{\beta+2} \right]_0^1 - \left(\frac{\beta}{\beta+1} \right)^2 \\ &= \frac{\beta}{\beta+2} - \left(\frac{\beta}{\beta+1} \right)^2 \end{aligned}$$

Opgave U41.1

Antag at X og Y er to uafhængige ligefordelte variable på intervallet $[0, 1]$

1. Opskriv tæthederne $p_x(x)$ og $p_y(y)$ for X hhv. Y .
2. Find $E(6X + 32Y)$
3. Find $E(X^3)$ og $E(X^3 + Y^3)$
4. Find $V(X) = E(X^2) - [E(X)]^2$
5. Find tætheden for $Z = X - \frac{1}{2}$

6. Find $E(Z)$ og $F(z) = P(Z \leq z)$

1. Generelt for en ligefordeling på intervallet (a, b) :

$$X \sim U(a, b) \Leftrightarrow p(x) = \frac{1_{(a,b)}(x)}{b - a}$$

Tæthedsfunktionerne for X og Y er derfor:

$$p(x) = \frac{1_{(0,1)}(x)}{1 - 0} = \begin{cases} 1 & \text{for } x \in (0, 1) \\ 0 & \text{for } x \notin (0, 1) \end{cases}$$

$$p(y) = \frac{1_{(0,1)}(y)}{1 - 0} = \begin{cases} 1 & \text{for } y \in (0, 1) \\ 0 & \text{for } y \notin (0, 1) \end{cases}$$

2.

$$\begin{aligned} E(6X + 32Y) &= 6E(X) + 32E(Y) \\ &= 6 \int_0^1 x \, dx + 32 \int_0^1 y \, dy \\ &= 6 \left[\frac{1}{2}x^2 \right]_0^1 + 32 \left[\frac{1}{2}y^2 \right]_0^1 \\ &= 3 + 16 \\ &= \underline{\underline{19}} \end{aligned}$$

3.

$$\begin{aligned} E(X^3) &= \int_0^1 x^3 \, dx \\ &= \left[\frac{1}{4}x^4 \right]_0^1 \\ &= \underline{\underline{\frac{1}{4}}} \end{aligned}$$

Da X og Y er identisk fordelt er $E(X^3) = E(Y^3) \Leftrightarrow E(X^3 + Y^3) = 2E(X^3) = \underline{\underline{\frac{1}{2}}}$

4.

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= \int_0^1 x^2 \, dx - \left(\int_0^1 x \, dx \right)^2 \\ &= \left[\frac{1}{3}x^3 \right]_0^1 - \left(\left[\frac{1}{2}x^2 \right]_0^1 \right)^2 \\ &= \frac{1}{3} - \frac{1}{4} \\ &= \underline{\underline{\frac{1}{12}}} \end{aligned}$$

5. Tæthedens for $Z = X - \frac{1}{2}$ kan findes via fordelingsfunktionen for X på intervallet $(0, 1)$, $F_X(x) = 1_{(0,1)}(x)x$:

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X - \frac{1}{2} \leq z) = P(X \leq z + \frac{1}{2}) = F_X\left(z + \frac{1}{2}\right) = 1_{(0,1)}\left(z + \frac{1}{2}\right)\left(z + \frac{1}{2}\right) \\ &= 1_{(-\frac{1}{2}, \frac{1}{2})}(z)\left(z + \frac{1}{2}\right) \end{aligned}$$

Udfra udtrykket for F_Z kan vi finde $p(z)$:

$$\begin{aligned} p(z) &= F'(z) = 1_{(-\frac{1}{2}, \frac{1}{2})}(z) \\ &= \begin{cases} 1 & \text{for } z \in (-\frac{1}{2}, \frac{1}{2}) \\ 0 & \text{for } z \notin (-\frac{1}{2}, \frac{1}{2}) \end{cases} \end{aligned}$$

$p(z)$ kan naturligvis også findes ved at benytte (5.4.1) eller (5.4.2).

- 6.

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} \left(x - \frac{1}{2}\right) 1_{(0,1)}(x) dx \\ &= \int_0^1 \left(x - \frac{1}{2}\right) dx \\ &= \left[\frac{1}{2}x^2 - \frac{1}{2}x\right]_0^1 \\ &= \left(\frac{1}{2} - \frac{1}{2}\right) \\ &= \underline{\underline{0}} \end{aligned}$$

$F_Z(z)$ er givet ved:

$$F_Z(z) = \begin{cases} 0 & \text{for } z < -\frac{1}{2} \\ z + \frac{1}{2} & \text{for } -\frac{1}{2} \leq z \leq \frac{1}{2} \\ 1 & \text{for } z > \frac{1}{2} \end{cases}$$

Opgave U41.2

1. Med X eksponentialfordelt er tæthedens $p(x) = \lambda \exp(-\lambda x)$. Opskriv fordelingsfunktionen $F(x)$ og vis, at $Y = F(x)$ er ligefordelt på $[0, 1]$

Fordelingsfunktion for X :

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x \lambda \exp(-\lambda x) 1_{(0,1)}(x) dx \\ &= \begin{cases} \left[-\exp(-\lambda x)\right]_0^x & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \\ &= \begin{cases} 1 - \exp(-\lambda x) & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \end{aligned}$$

Det gælder, at $Y = F(X) = 1 - \exp(-\lambda X)1_{(0,\infty)}(X)$ er ligefordelt på $(0, 1)$, hvis $F_Y(y) = y$ på $(0, 1)$.

$$\begin{aligned}
 y \in (0, 1) : F_Y(y) &= P(Y \leq y) = P((1 - \exp(-\lambda X)) \leq y) \\
 &= P(\exp(-\lambda X) \geq 1 - y) \\
 &= P(-\lambda X \geq \log(1 - y)) \\
 &= P(X \leq -\frac{\log(1 - y)}{\lambda}) \\
 &= F_X\left(-\frac{\log(1 - y)}{\lambda}\right) \\
 &= 1 - \exp\left(-\lambda\left(-\frac{\log(1 - y)}{\lambda}\right)\right) \\
 &= 1 - \exp(\log(1 - y)) \\
 &= 1 - (1 - y) \\
 &= y
 \end{aligned}$$

Hermed er det vist, at $Y = F(x)$ er ligefordelt på $(0, 1)$.

2. Med X standard normalfordelt er $P(X \leq x) = \Phi(x)$. Vis, at $Y = \Phi(x)$ er ligefordelt på $[0, 1]$.

Ifølge *Sætning 5.4.2* i bogen er tæthedens $q(y)$ af $Y = t(X)$ givet ved:

$$q(y) = \begin{cases} \frac{p(t^{-1}(y))}{|t'(t^{-1}(y))|} & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.2)$$

X er $N(0, 1)$, så $p(x) = \phi(x) = \Phi'(x)$. Med $Y = t(X) = \Phi(X)$ får vi dermed:

$$\begin{aligned}
 q(y) &= \frac{p(\Phi^{-1}(y))}{|\Phi'(\Phi^{-1}(y))|} \\
 &= \frac{\phi(\Phi^{-1}(y))}{|\phi(\Phi^{-1}(y))|} \\
 &= 1
 \end{aligned}$$

Da en variabel Y med tæthed $q(y) = 1$ er ligefordelt på intervallet $(0, 1)$ har vi vist at $Y = \Phi(x)$ er ligefordelt på $[0, 1]$.

Opgave 5.1

Antag at den stokastiske variable X er eksponentialfordelt med parameter λ . Find $P(X \geq x)$ for alle $x \geq 0$. Find for $\lambda = 1$ sandsynligheden $P(1 < X < 2)$.

Sandsynlighedstætheden for X er:

$$p(x) = \lambda e^{-\lambda x}$$

Læg mærke til, at i kontinuerte fordelinger er punktsandsynligheden altid 0. Derfor skelnes der ikke mellem skarpe og svage uligheder.

$$\begin{aligned}
 P(X \geq x) &= 1 - P(X \leq x) \\
 &= 1 - \int_0^x \lambda e^{-\lambda x} dx \\
 &= 1 + \left[e^{-\lambda x} \right]_0^x \\
 &= 1 + e^{-\lambda x} - 1 \\
 &= \underline{\underline{e^{-\lambda x}}}
 \end{aligned}$$

Nu sættes $\lambda = 1$:

$$\begin{aligned}
 P(1 < X < 2) &= \int_1^2 e^{-x} dx \\
 &= \left[-e^{-x} \right]_1^2 \\
 &= \underline{\underline{e^{-1} - e^{-2}}}
 \end{aligned}$$

Opgave 5.5

(a) Vis, at

$$p(x) = \frac{1}{2} e^{-|x|}, \quad x \in \mathbb{R}$$

er en sandsynlighedstæthed på \mathbb{R} . Den tilsvarende fordeling kaldes Laplace-fordelingen eller den tosidede eksponentialfordeling.

(b) Find fordelingsfunktionen.

(c) Vis, at fordelingen har middelværdi og varians, og find disse størrelser.

(a) $p(x)$ er en sandsynlighedstæthed på \mathbb{R} , hvis $\int_{-\infty}^{\infty} p(x) dx = 1$.

$$\begin{aligned}
 \int_{-\infty}^{\infty} \frac{1}{2} e^{-|x|} dx &= \frac{1}{2} \left(\int_{-\infty}^0 e^x dx + \int_0^{\infty} e^{-x} dx \right) \\
 &= \frac{1}{2} \left([e^x]_{-\infty}^0 - [e^{-x}]_0^{\infty} \right) \\
 &= \frac{1}{2} ((1 - 0) - (0 - 1)) \\
 &= 1
 \end{aligned}$$

(b)

$$\begin{aligned}
 F(x) &= \frac{1}{2} \int_{-\infty}^x e^{-|x|} dx = \begin{cases} \frac{1}{2} \int_{-\infty}^x e^x dx, & x < 0 \\ \frac{1}{2} \int_{-\infty}^0 e^x dx + \frac{1}{2} \int_0^x e^{-x} dx, & x \geq 0 \end{cases} \\
 &= \begin{cases} \frac{1}{2} [e^x]_{-\infty}^x, & x < 0 \\ \frac{1}{2} \left([e^x]_{-\infty}^0 - [e^{-x}]_0^x \right), & x \geq 0 \end{cases} \\
 &= \begin{cases} \frac{1}{2} e^x, & x < 0 \\ 1 - \frac{1}{2} e^{-x}, & x \geq 0 \end{cases}
 \end{aligned}$$

(c) En fordeling har middelværdi, hvis

$$\int_{-\infty}^{\infty} |x| p(x) dx < \infty, \quad (5.2.1)$$

og middelværdien er da

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx < \infty. \quad (5.2.2)$$

En fordeling har varians, hvis

$$\int_{-\infty}^{\infty} x^2 p(x) dx < \infty, \quad (5.2.8)$$

og variansen er da

$$Var(X) = \int_{-\infty}^{\infty} x^2 p(x) dx - (E(X))^2$$

Påvisning af, at fordelingen har middelværdi:

$$\begin{aligned} \frac{1}{2} \int_{-\infty}^{\infty} |x| p(x) dx &= \frac{1}{2} \left(\int_{-\infty}^0 -xe^x dx + \int_0^{\infty} xe^{-x} dx \right) \left(\text{benyt } \int_b^a f(x)g(x) dx = [f(x)G(x)]_b^a - \int_b^a f'(x)G(x) dx \right) \\ &= \frac{1}{2} \left([-xe^x]_{-\infty}^0 + \int_{-\infty}^0 e^x dx - [xe^{-x}]_0^{\infty} + \int_0^{\infty} e^{-x} dx \right) \\ &= \frac{1}{2} ((-0 - 0) + (1 - 0) - (0 - 0) + (0 - 1)) \\ &= \frac{1}{2} (1 - 1) \\ &= 0 < \infty \end{aligned}$$

Fordelingen har altså middelværdi:

$$\begin{aligned} E(X) &= \frac{1}{2} \left(\int_{-\infty}^0 -xe^x dx + \int_0^{\infty} xe^{-x} dx \right) \quad (\text{benyt resultatet ovenfor}) \\ &= \frac{1}{2} [e^x - xe^x]_{-\infty}^0 + [e^{-x} - xe^{-x}]_0^{\infty} \\ &= \frac{1}{2} ((1 - 0 - 0 + 0) + (0 - 0 - 1 + 0)) \\ &= \frac{1}{2} (1 - 1) \\ &= \underline{\underline{0}} \end{aligned}$$

Påvisning af, at fordelingen har varians:

$$\begin{aligned} \frac{1}{2} \int_{-\infty}^{\infty} x^2 p(x) dx &= \frac{1}{2} \left(\int_{-\infty}^0 x^2 e^x dx + \int_0^{\infty} x^2 e^{-x} dx \right) \left(\text{benyt } \int_b^a f(x)g(x) dx = [f(x)G(x)]_b^a - \int_b^a f'(x)G(x) dx \right) \\ &= \frac{1}{2} \left([x^2 e^x]_{-\infty}^0 - \int_{-\infty}^0 2xe^x dx - [x^2 e^{-x}]_0^{\infty} + \int_0^{\infty} 2xe^{-x} dx \right) \\ &= \frac{1}{2} \left([x^2 e^x]_{-\infty}^0 - [2xe^x]_{-\infty}^0 + \int_{-\infty}^0 2e^x dx - [x^2 e^{-x}]_0^{\infty} + [2xe^{-x}]_0^{\infty} - \int_0^{\infty} 2e^{-x} dx \right) \\ &= \frac{1}{2} \left([x^2 e^x]_{-\infty}^0 - [2xe^x]_{-\infty}^0 + [2e^x]_{-\infty}^0 - [x^2 e^{-x}]_0^{\infty} + [2xe^{-x}]_0^{\infty} + [2e^{-x}]_{-\infty}^0 \right) \\ &= \frac{1}{2} ((0 - 0) - (0 - 0) + (2 - 0) - (0 - 0) + (0 - 0) + (2 - 0)) \\ &= 2 < \infty \end{aligned}$$

Fordelingen har altså varians:

$$\begin{aligned} Var(X) &= \frac{1}{2} \int_{-\infty}^{\infty} x^2 p(x) dx - E(X)^2 \\ &= 2 - 0^2 \\ &= \underline{\underline{2}} \end{aligned}$$

Opgave 5.13

Lad X være en kontinuert stokastisk variabel, som er koncentreret på et interval (a, b) , og antag at X 's sandsynlighedstæthed p er kontinuert på (a, b) . Find sandsynlighedstætheden for

- (a) $\exp(X)$

Antag nu, at $a \geq 0$, og find tætheden for fordelingen af

- (b) \sqrt{X}

- (c) $\frac{1}{X}$

- (d) X^2

- (e) Antag til slut, at X kan antage værdier på hele den reelle akse fraregnet nul. Hvad er da tætheden for fordelingen af $\frac{1}{X}$? Hvad er svaret, hvis også nul er en mulig værdi for X ?

Fordeling af $Y = t(X)$ er givet ved:

$$q(y) = \begin{cases} p(t^{-1}(y)) \left| \frac{d}{dy} t^{-1}(y) \right| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.1)$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y)) / |t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.2)$$

Hvilken af de to formler man benytter afhænger af kontekst. I det følgende laver vi udregninger med begge formler og viser at de giver samme resultat.

- (a) $Y = t(X) = \exp(X) \Leftrightarrow X = t^{-1}(Y) = \log(Y)$, $\frac{d}{dy} t^{-1}(y) = \frac{1}{y}$, $t'(t^{-1}(y)) = \exp(t^{-1}(y))$:

$$q(y) = \begin{cases} p(t^{-1}(y)) \left| \frac{d}{dy} t^{-1}(y) \right| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\log(y))}{|y|} & \text{hvis } y \in (\exp(a), \exp(b)) \\ 0 & \text{hvis } y \notin (\exp(a), \exp(b)) \end{cases}$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y)) / |t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\log(y))}{|y|} & \text{hvis } y \in (\exp(a), \exp(b)) \\ 0 & \text{hvis } y \notin (\exp(a), \exp(b)) \end{cases}$$

- (b) $Y = t(X) = \sqrt{X} \Leftrightarrow X = t^{-1}(Y) = Y^2$, $\frac{d}{dy} t^{-1}(y) = 2y$, $t'(t^{-1}(y)) = \frac{1}{2y}$:

$$q(y) = \begin{cases} p(t^{-1}(y)) \left| \frac{d}{dy} t^{-1}(y) \right| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} p(y^2) 2y & \text{hvis } y \in (\sqrt{a}, \sqrt{b}) \\ 0 & \text{hvis } y \notin (\sqrt{a}, \sqrt{b}) \end{cases}$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y))/|t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\log(y))}{|y|} & \text{hvis } y \in (\sqrt{a}, \sqrt{b}) \\ 0 & \text{hvis } y \notin (\sqrt{a}, \sqrt{b}) \end{cases}$$

(c) $Y = t(X) = \frac{1}{X} \Leftrightarrow X = t^{-1}(Y) = \frac{1}{Y}$, $\frac{d}{dy}t^{-1}(y) = -\frac{1}{y^2}$, $t'(t^{-1}(y)) = -y^2$:

$$q(y) = \begin{cases} p(t^{-1}(y))|\frac{d}{dy}t^{-1}(y)| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\frac{1}{y})}{y^2} & \text{hvis } y \in (\frac{1}{b}, \frac{1}{a}) \\ 0 & \text{hvis } y \notin (\frac{1}{b}, \frac{1}{a}) \end{cases}$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y))/|t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\frac{1}{y})}{y^2} & \text{hvis } y \in (\frac{1}{b}, \frac{1}{a}) \\ 0 & \text{hvis } y \notin (\frac{1}{b}, \frac{1}{a}) \end{cases}$$

(d) $Y = t(X) = X^2 \Leftrightarrow X = t^{-1}(Y) = \sqrt{Y}$, $\frac{d}{dy}t^{-1}(y) = \frac{1}{2\sqrt{y}}$, $t'(t^{-1}(y)) = 2\sqrt{y}$:

$$q(y) = \begin{cases} p(t^{-1}(y))|\frac{d}{dy}t^{-1}(y)| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\sqrt{y})}{2\sqrt{y}} & \text{hvis } y \in (a^2, b^2) \\ 0 & \text{hvis } y \notin (a^2, b^2) \end{cases}$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y))/|t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} = \begin{cases} \frac{p(\sqrt{y})}{2\sqrt{y}} & \text{hvis } y \in (a^2, b^2) \\ 0 & \text{hvis } y \notin (a^2, b^2) \end{cases}$$

(e) Ifølge *Sætning 5.4.1* gælder de ovenfor benyttede transformationssætninger kun, hvis den reelle funktion t er kontinuert differentiabel for all $x \in (a, b)$.

$t(X) = \frac{1}{X}$ er ikke kontinuert i $X = 0$ da $\lim_{x \rightarrow 0^-} \frac{1}{X} = -\infty$ og $\lim_{x \rightarrow 0^+} \frac{1}{X} = \infty$.

Tæthedens fordelingen af $\frac{1}{X}$ dermed kun beregnes for intervallerne $]-\infty, 0[$ og $]0, \infty[$, og vil da være givet ved svaret i delopgave (c).

Opgave 5.15

Lad X være normalfordelt med middelværdi μ og varians σ^2 . Fordelingen af $Y = \exp(X)$ kaldes den logaritmiske normalfordeling med parametre (μ, σ^2) .

- (a) Find sandsynlighedstætheden for fordelingen af Y .
- (b) Vis, at hvis Y er logaritmisk normalfordelt, så er også βY logaritmisk normalfordelt for ethvert $\beta > 0$. En klasse af fordelinger med denne egenskab kaldes skalainvariant. Opskriv den formel, som viser, hvorledes den logaritmiske normalfordelings parametre ændrer sig ved en sådan skalatransformation.
- (c) Vis, at middelværdien i den logaritmiske normalfordeling med parametre $\mu = 0$ og $\sigma^2 = 1$ er $\sqrt{e} = 1,6487$.

- (a) X er normalfordelt med middeværdi μ og varians σ^2 :

$$X \sim N(\mu, \sigma^2).$$

Fordelingen af X er således givet ved sandsynlighedstætheden

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Y er dannet ved en transformation af X :

$$Y = t(X) = \exp(X) \Leftrightarrow X = t^{-1}(Y) = \log(Y), Y \in \mathbb{R}_+$$

Y er logaritmisk normalfordelt, hvilket noteres som:

$$Y \sim \log N(\mu, \sigma^2)$$

Vi kan bruge (5.4.2) til beregne sandsynlighedstætheden for fordelingen af Y :

$$\begin{aligned} q(y) &= \begin{cases} p(t^{-1}(y))/|t'(t^{-1}(y))| & \text{hvis } y \in \mathbb{R}_+ \\ 0 & \text{hvis } y \notin \mathbb{R}_+ \end{cases} \\ &= \frac{\mathbb{I}_{y \in \mathbb{R}_+}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right) / \left|\frac{d \exp(\log(y))}{d \log(y)}\right| \\ &= \frac{\mathbb{I}_{y \in \mathbb{R}_+}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right) \left(\frac{1}{y}\right) \end{aligned}$$

- (b) I det følgende benytter vi

$$X \sim N(\mu, \sigma^2) \Leftrightarrow Y = a + bX \sim N(a + b\mu, b^2\sigma^2).$$

For at vise, at βY også er log-normalfordelt benytter vi følgende mellemregning:

$$\log(\beta Y) = \log(\beta) + \log(Y) = \log(\beta) + X \sim N(\mu + \log(\beta), \sigma^2)$$

Udtrykket $\log(\beta Y)$ er altså normalfordelt. $\exp(\log(\beta Y))$ er således logaritmisk normalfordelt:

$$\exp(\log(\beta Y)) = \beta Y \sim \log N(\mu + \log(\beta), \sigma^2)$$

Hermed ser vi også hvorledes den logaritmiske normalfordelings parametre ændrer sig ved en skalatransformation.

- (c) $Y \sim \log N(0, 1) \Leftrightarrow Y = \exp(X), X \sim N(0, 1)$

Sandsynlighedstætheden for X er

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

og vi kan benytte (5.2.4) til at beregne middelværdien for $Y = \exp(X)$:

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \exp(x) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} + 1\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2 - 2x}{2}\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x^2 - 2x + 1) + \frac{1}{2}\right) dx \\
&= \exp\left(\frac{1}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-1)^2\right) dx \\
&= \sqrt{e} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2}\right) dx}_{=1, \text{ da } det \text{ er } X \sim N(1,1)} \\
&= \underline{\underline{\sqrt{e}}}
\end{aligned}$$

Opgave 5.17

Antag, at X er ligefordelt på intervallet $(0, 1)$. Find fordelingen af $Y = -\log(X)$.

Fordelingen for X er:

$$p(x) = 1_{(0,1)}(x)$$

Den stokastiske variable X er transformeret ved $t(X) = Y = -\log(X)$. Transformationen udspænder intervallet $(0, 1)$ på intervallet $(0, \infty[$

Den inverse transformation er $t^{-1}(Y) = X = \exp(-Y)$.

Fordelingen af Y , $q(y)$ kan da beregnes ved (5.4.1) eller (5.4.2):

$$\begin{aligned}
q(y) &= 1_{(0,\infty)}(y) p(t^{-1}(y)) \left| \frac{d}{dy} t^{-1}(y) \right| \\
&= \underline{\underline{1_{(0,\infty)}(y) \exp(-y)}}
\end{aligned}$$

eller

$$\begin{aligned}
q(y) &= 1_{(0,\infty)}(y) p(t^{-1}(y)) / \left| t'(t^{-1}(y)) \right| \\
&= 1_{(0,\infty)}(y) / \underline{\underline{\frac{1}{\exp(-y)}}} \\
&= \underline{\underline{1_{(0,\infty)}(y) \exp(-y)}}
\end{aligned}$$

Fordelingen af Y kan også udledes via fordelingsfunktionerne for X , $F_X(x)$ og Y , $F_Y(y)$:

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(-\log(X) \leq y) \\
&= P(X \geq \exp(-y)) \\
&= 1 - F_X(\exp(-y)) \quad (\text{husk } X \sim U(0, 1) \Leftrightarrow F(x) = x, \text{ for } x \in (0, 1)) \\
&= 1 - \exp(-y)
\end{aligned}$$

$X \in (0, 1) \Leftrightarrow Y = -\log(X) \in (0, \infty[.$ Dermed:

$$\begin{aligned} q(y) &= F'(y) \\ &= \underline{\underline{1_{(0,\infty)}(y) \exp(-y)}} \end{aligned}$$

Vi tester om det udledte udtryk for $q(y)$ er en sandsynlighedsfunktion ved at teste, om $\int_0^\infty q(y)dy = 1$:

$$\begin{aligned} \int_0^\infty 1_{(0,\infty)}(y) \exp(-y) dy &= \left[-\exp(-y) \right]_0^\infty \\ &= 0 - (-1) \\ &= 1 \end{aligned}$$

Opgave U41.3

Lad X være $N(\mu, \sigma^2)$ fordelt

1. Hvad er fordelingen af $Y = \frac{X-\mu}{\sigma}$?

2. Find tætheden for $Z = \left(\frac{X-\mu}{\sigma}\right)^2$.

1. For at beregne $Y = \frac{X-\mu}{\sigma}$ bruger vi igen regnereglen:

$$X \sim N(\mu, \sigma^2) \Leftrightarrow Y = a + bX \sim N(a + b\mu, b^2\sigma^2).$$

Vi ved, at $X \sim N(\mu, \sigma^2)$ og hvis vi danner en ny stokastisk variabel $W = X - \mu$ gælder $W \sim N(0, \sigma^2)$. Nu kan vi udtrykke Y som:

$$Y = \frac{W}{\sigma}$$

og

$$Var(Y) = Var\left(\frac{W}{\sigma}\right) = \frac{1}{\sigma^2}Var(W) = \frac{\sigma^2}{\sigma^2} = 1$$

Fordelingen af $Y = \frac{X-\mu}{\sigma}$ er altså $N(0, 1)$ med sandsynlighedstætheden

$$q(y) = \underline{\underline{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)}}$$

Dette kan også nemt vises med (5.4.1) og (5.4.2).

2. Udfaldsrummet for $Z = \left(\frac{X-\mu}{\sigma}\right)^2$ er \mathbb{R}_+ .

$Z = t(Y) = Y^2$ er ikke en monoton transformation. $p(z)$ for $z > 0$ udledes derfor via fordelingsfunktionen for Y :

$$\begin{aligned}
 F_Z(z) &= P(Z \leq z) \\
 &= P(Y^2 \leq z) \\
 &= P(-\sqrt{z} \leq Y \leq \sqrt{z}) \\
 &= F_Y(\sqrt{z}) - F_Y(-\sqrt{z}) \Leftrightarrow \\
 p(z) &= \frac{d}{dz} \left(F_Y(\sqrt{z}) - F_Y(-\sqrt{z}) \right) \quad \left(\text{Husk, at } \frac{d}{dy} F_Y(y) = q(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \right) \\
 &= \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\sqrt{z}^2}{2}\right\} + \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(-\sqrt{z})^2}{2}\right\} \\
 &= \frac{2}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \left(\exp\left\{-\frac{z^2}{2}\right\} \right) \\
 &= \begin{cases} \frac{1}{\sqrt{2\pi}z} \exp\left(-\frac{z}{2}\right) & \text{hvis } z > 0 \\ 0 & \text{hvis } z \leq 0 \end{cases}
 \end{aligned}$$

hvilket er χ^2 -fordelingen med 1 frihedsgrad.

Centrale begreber

- Ligefordelingen på $B \in \mathbb{R}^n$ $\left(p(x_1, \dots, x_n) = \frac{1_B(x_1, \dots, x_n)}{|B|}\right)$ (6.1.2)
- Lad (X, Y) være en to-dimensional kontinuert stokastisk vektor med sandsynlighedstæthed $p(x, y)$.
Den marginale sandsynlighed q for X er:

$$q(x) = \int_{\mathbb{R}} p(x, y) dy \quad (6.1.5)$$

- Lad (X, Y) være to uafhængige kontinuerte stokastiske variable med marginale tætheder p_1 og p_2 .
 $Z = X + Y$ er da en kontinuert stokastisk variabel med sandsynlighedstæthed:

$$q(z) = \int_{-\infty}^{\infty} p_1(x) p_2(z - x) dx \quad (6.3.2)$$

- Regneregler for kovarianser:

$$Cov(X, Y) = E(XY) - E(X)E(Y) \quad (3.8.2)$$

$$Cov(a + bX, c + dY) = bd Cov(X, Y) \quad (3.8.3)$$

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \quad (3.8.4)$$

$$Cov(X, Y) = Cov(Y, X) \quad (3.8.5)$$

$$Cov(X, X) = Var(X)$$

Opgave U43.1.1

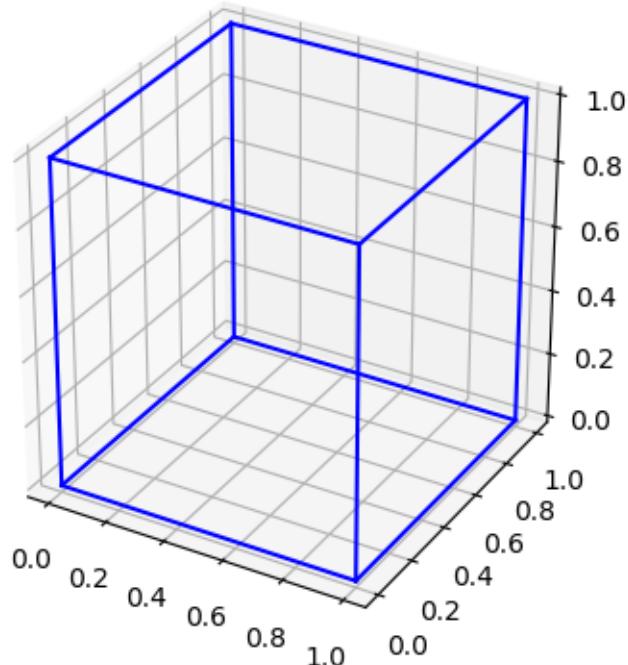
Lad (X, Y) være en to-dimensional stokastisk vektor på $A = (0, 1) \times (0, 1)$ med tæthedsfunktion givet ved

$$p(x, y) = 1_A(x, y)$$

- Udregn sandsynligheden for, at $P(X < 0,1; Y < 0,6)$. Forklar hvordan man grafisk kan vise sandsynligheden for en ligefordeling på A .

$$\begin{aligned} P(X < 0,1; Y < 0,6) &= \int_0^{0,1} \int_0^{0,6} dy dx \\ &= \int_0^{0,1} [y]_0^{0,6} dx \\ &= [0, 6x]_0^{0,1} \\ &= \underline{\underline{0,06}} \end{aligned}$$

Grafisk ser ligefordelingen således ud:



2. Udregn sandsynligheden for $P(0, 25 \leq X < 0, 75; 0, 4 \leq Y < 0, 6)$

$$\begin{aligned} P(0, 25 \leq X < 0, 75; 0, 4 \leq Y < 0, 6) &= \int_{0,25}^{0,75} \int_{0,4}^{0,6} dy dx \\ &= \int_{0,25}^{0,75} [y]_{0,4}^{0,6} dx \\ &= [0, 2x]_{0,25}^{0,75} \\ &= \underline{\underline{0, 1}} \end{aligned}$$

3. Udregn sandsynligheden for, at $P(X < 0, 1)$

$$\begin{aligned} P(X < 0, 1) &= \int_0^1 \int_0^{0,1} dx dy \\ &= \int_0^1 [x]_0^{0,1} dy \\ &= \underline{\underline{0, 1}} \end{aligned}$$

4. Find den marginale fordeling for X

Den marginale fordeling for den ene variabel findes ved at integrere i forhold til den anden variabel.

$$\begin{aligned} p(x) &= \int_{\mathbb{R}} 1_A(x, y) dy \\ &= \int_{\mathbb{R}} 1_{(0,1)}(x) 1_{(0,1)}(y) dy \\ &= 1_{x \in (0,1)} \int_0^1 dy \\ &= \mathbb{I}_{x \in (0,1)} [y]_0^1 \\ &= \mathbb{I}_{x \in (0,1)} \\ &= \begin{cases} 1 & \text{for } x \in (0, 1) \\ 0 & \text{ellers} \end{cases} \end{aligned}$$

5. Vis at X og Y er uafhængige

Den marginale fordeling for Y er $\mathbb{I}_{y \in (0,1)}$, og vi har dermed:

$$p(x) \cdot p(y) = \mathbb{I}_{x \in (0,1)} \cdot \mathbb{I}_{y \in (0,1)} = \mathbb{I}_{x \in (0,1), y \in (0,1)} = p(x, y)$$

hvilket viser, at X og Y er uafhængige.

Opgave U.43.1.2

Antag, at X og Y er uafhængige ligefordelte variable på intervallet $[0, 1]$.

1. Find $E(Y^*)$ og $V(Y^*)$, hvor $Y^* = 2Y$.

$$\begin{aligned}
 V(Y^*) &= E(Y^{*2}) - E(Y^*)^2 \\
 E(Y^*) &= \int_{-\infty}^{\infty} 2y \cdot 1_{(0,1)}(y) dy & &= \int_0^1 4y^2 dy - 1 \\
 &= \int_0^1 2y dy & &= \left[\frac{4}{3}y^3 \right]_0^1 - 1 \\
 &= \left[y^2 \right]_0^1 & &= \frac{4}{3} - 1 \\
 &= \underline{\underline{1}} & &= \underline{\underline{\frac{1}{3}}}
 \end{aligned}$$

2. Opskriv tætheden $p^*(y^*)$ for Y^* .

Y^* er en (monoton) transformation af Y :

$$Y^* = t(Y) = 2Y \Leftrightarrow Y = t^{-1}(Y^*) = \frac{1}{2}Y^*$$

og tætheden for X er $p(x) = 1_{(0,1)}(x)$.

Vi kan derfor finde tætheden for Y^* , $p^*(y^*)$, ved (5.4.1):

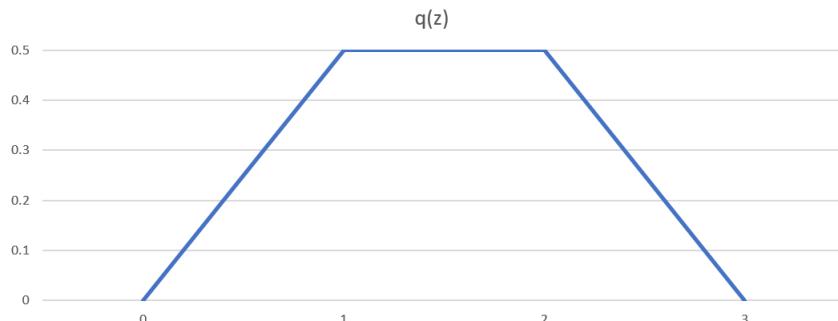
$$\begin{aligned}
 p^*(y^*) &= p\left(\frac{1}{2}y^*\right) \left| \frac{d}{dr}\left(\frac{1}{2}y^*\right) \right| \\
 &= 1_{(0,1)}\left(\frac{1}{2}y^*\right) \frac{1}{2} \\
 &= \frac{1_{(0,2)}(y^*)}{2}
 \end{aligned}$$

3. Find tætheden $q(z)$ for $Z = X + Y^*$.

Det er nemt at se, at Z kan antage værdier på intervallet $(0, 3)$. Men Z er ikke ligefordelt på dette interval da $Y^* = 2Y \in (0, 2)$ begrænses af i hvilket delinterval af Z vi befinder os:

- $(0 \leq Z < 1) : Vi \ kan \ bruge \ "den \ nedre \ halvdel" \ af \ Y^*, \ Y^* \in (0, 1)$
- $(1 \leq Z < 2) : Vi \ kan \ bruge \ hele \ Y^*, \ Y^* \in (0, 2)$
- $(2 \leq Z \leq 3) : Vi \ kan \ bruge \ "den \ øvre \ halvdel" \ af \ Y^*, \ Y^* \in (1, 2)$

Tætheden q af Z ser derfor således ud:



Da X og Y^* er uafhængige er sandsynlighedstætheden $q(z)$ ifølge (6.3.2) givet ved

$$q(z) = \int_{-\infty}^{\infty} p(x)p^*(z-x)dx$$

Idet vi deler Z og $Y^* = Z - X$ op i delintervallerne givet ovenfor får vi således:

$$\begin{aligned} q(z) &= \underbrace{\int_0^1 1_{(0,1)}(x) \frac{1_{(0,1)}(z-x)}{2} dx}_{z \in (0,1)} + \underbrace{\int_0^1 1_{(0,1)}(x) \frac{1_{(0,2)}(z-x)}{2} dx}_{z \in (1,2)} + \underbrace{\int_0^1 1_{(0,1)}(x) \frac{1_{(1,2)}(z-x)}{2} dx}_{z \in (2,3)} \\ &= \frac{1}{2} \int_0^z 1_{(0,1)}(z) dx + \frac{1}{2} \int_0^1 1_{(1,2)}(z) dx + \frac{1}{2} \int_{z-2}^1 1_{(2,3)}(z) dx \\ &= \underline{\underline{\frac{1}{2} (1_{(0,1)}(z)z + 1_{(1,2)}(z) + 1_{(2,3)}(z)(3-z))}} \end{aligned}$$

4. Hvad er kovariansen mellem Z og Y ?

$$\begin{aligned} Cov(Z, Y) &= E((X + 2Y)Y) - E(X + 2Y)E(Y) \\ &= E(XY + 2Y^2) - E(X)E(Y) - 2E(Y)E(Y) \\ &= \underbrace{E(XY) - E(X)E(Y)}_{Cov(X,Y)=0} + 2(E(Y^2) - (E(Y))^2) \\ &= 2V(Y) \\ &= 2V\left(\frac{1}{2}Y^*\right) \\ &= \frac{1}{2} \times \frac{1}{3} \\ &= \underline{\underline{\frac{1}{6}}} \end{aligned}$$

Opgave U43.1.3

Lad (X, Y) være en to-dimensional stokastisk vektor på $[5, 10] \times [3, 7]$. Tæthedsfunktionen er givet ved

$$p(x, y) = \frac{1}{20} 1_{[5,10] \times [3,7]}(x, y).$$

1. Forklar hvorfor p er en tæthedsfunktion.

Ligefordelingen på en begrænset delmængde B af \mathbb{R}^n har sandsynlighedstætheden

$$p(x_1, \dots, x_n) = \frac{1_B(x_1, \dots, x_n)}{|B|}$$

hvor B betegner det n -dimensionale volumen af B .

Hvis vi definerer $B \equiv (5, 10) \times (3, 7)$ og antager at (X, Y) er en ligefordelt stokastisk vektor, har vi følgende sandsylyghedstæthed p for (X, Y) :

$$\begin{aligned} p(x, y) &= \frac{1_B(x, y)}{|B|} \\ &= \frac{1_{(5,10) \times (3,7)}(x, y)}{(10-5) \cdot (7-3)} \\ &= \underline{\underline{\frac{1_{(5,10) \times (3,7)}(x, y)}{20}}} \end{aligned}$$

2. Udregn $P(6 \leq X < 10, 4 \leq Y < 6)$

$$\begin{aligned} P(6 \leq X < 10, 4 \leq Y < 6) &= \int_6^{10} \int_4^6 \frac{1}{20} dy dx \\ &= \frac{1}{20}(6-4)(10-6) \\ &= \frac{2}{5} \\ &\underline{\underline{=}} \end{aligned}$$

3. Find de marginale fordelinger af X og Y .

Den marginale fordeling for dne ene variabel findes ved at integrere i forhold til den anden variabel.

$$\begin{aligned} p(x) &= \frac{1}{20} \int_3^7 1_{(5,10)}(x) dy \\ &= \frac{1}{5} \underline{\underline{1_{(5,10)}(x)}} \\ p(y) &= \frac{1}{20} \int_5^{10} 1_{(3,7)}(y) dx \\ &= \frac{1}{4} \underline{\underline{1_{(3,7)}(y)}} \end{aligned}$$

4. Find Middelværdien for X .

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot \frac{1}{5} 1_{(5,10)}(x) dx \\ &= \frac{1}{5} \int_5^{10} x dx \\ &= \frac{1}{10} [x^2]_5^{10} \\ &= 10 - 2,5 \\ &= \underline{\underline{7,5}} \end{aligned}$$

Opgave U43.1.4

Lad (X, Y) være en kontinuert to-dimensonal stokastisk vektor med sandsynlighedstæthed

$$f(x, y) = 6 \exp(-2x - 3y), \quad x \in [0, \infty) \text{ og } y \in [0, \infty)$$

1. Udregn følgende sandsynligheder $P(X \leq 2, Y \leq 4)$ og $P(X > 1, Y \leq 3)$.

$$\begin{aligned}
 P(X \leq 2, Y \leq 4) &= \int_0^2 \int_0^4 6 \exp(-2x - 3y) dy dx \\
 &= \int_0^2 \left[-2 \exp(-2x - 3y) \right]_{y=0}^{y=4} dx \\
 &= \int_0^2 (2 \exp(-2x) - 2 \exp(-2x - 12)) dx \\
 &= \left[-\exp(-2x) + \exp(-2x - 12) \right]_{x=0}^{x=2} \\
 &= 1 - \exp(-12) - \exp(-4) + \exp(-16) \\
 &= \underline{\underline{(1 - \exp(-4))(1 - \exp(-12))}}
 \end{aligned}$$

$$\begin{aligned}
 P(X > 1, Y \leq 3) &= \int_1^\infty \int_0^3 6 \exp(-2x - 3y) dy dx \\
 &= \int_1^\infty \left[-2 \exp(-2x - 3y) \right]_{y=0}^{y=3} dx \\
 &= \int_1^\infty (2 \exp(-2x) - 2 \exp(-2x - 9)) dx \\
 &= \lim_{n \rightarrow \infty} \left[-\exp(-2x) + \exp(-2x - 9) \right]_{x=1}^{x=n} \\
 &= \underline{\underline{\exp(-2) - \exp(-11)}}
 \end{aligned}$$

2. Find de marginale sandsynligheder for X og Y og angiv deres fordeling.

$$\begin{aligned}
 q(x) &= \int_0^\infty 6 \exp(-2x - 3y) dy \\
 &= \lim_{n \rightarrow \infty} \left[-2 \exp(-2x - 3y) \right]_{y=0}^{y=n} \\
 &= \underline{\underline{2 \exp(-2x)}}
 \end{aligned}$$

X er eksponentialfordelt med parameter $\lambda = 2$.

$$X \sim \exp(2).$$

$$\begin{aligned}
 q(y) &= \int_0^\infty 6 \exp(-2x - 3y) dx \\
 &= \lim_{n \rightarrow \infty} \left[-3 \exp(-2x - 3y) \right]_{x=0}^{x=n} \\
 &= \underline{\underline{3 \exp(-3y)}}
 \end{aligned}$$

Y er eksponentialfordelt med parameter $\lambda = 3$.

$$Y \sim \exp(3).$$

3. Find fordelingsfunktionen for X og udregn medianen

50% – fraktilen er den værdi X der sikrer at præcis det halve af sandsynlighedsmassen er ”under” X , $P(X \leq x) = F_X(x) = 0,5$.

Fordelingsfunktionen for X :

$$\begin{aligned} P(X \leq x) = F_X(x) &= \int_0^x 2 \exp(-2x) dx \\ &= \left[-\exp(-2x) \right]_0^x \\ &= 1 - \exp(-2x) \end{aligned}$$

50% – fraktilen:

$$\begin{aligned} 1 - \exp(-2x) = 0,5 &\Leftrightarrow \\ -2x = \log(\frac{1}{2}) &\Leftrightarrow \\ 2x = \log(2) &\Leftrightarrow \\ x = \frac{\log(2)}{2} & \end{aligned}$$

4. Vis, at X og Y er uafhængige

$$\begin{aligned} q(x) \cdot q(y) &= 2 \exp(-2x) \cdot 3 \exp(-3y) \\ &= 6 \exp(-2x - 3y) \\ &= f(x, y) \end{aligned}$$

Da produktet af de marginale sandsynlighedstætheder for X og Y er lig med sandsynlighedstætheden er for (X, Y) er X og Y uafhængige.

Opgave U43.2.1

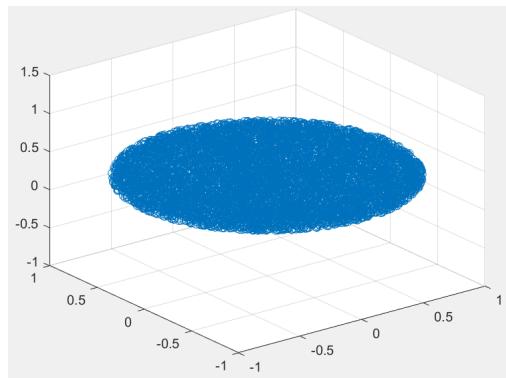
Antag, at X, Y er ligefordelte på enhedscirklen

$$A = \{x, y | x^2 + y^2 \leq 1\}$$

med tætheden

$$p(x, y) = \frac{1}{\pi} 1_A(x, y)$$

1. Prøv at tegne $p(x, y)$



2. Find den marginale tæthed for X , $p(x)$.

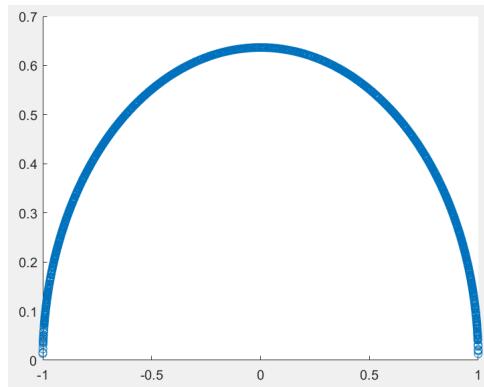
Vi kan omskrive tætheden $p(x, y)$ til

$$p(x, y) = \frac{1}{\pi} \left(1_{[-\sqrt{1-y^2}, \sqrt{1-y^2}]}(x) 1_{[-\sqrt{1-x^2}, \sqrt{1-x^2}]}(y) \right)$$

og integrere mht. y for at finde $p(x)$.

$$\begin{aligned}
 p(x) &= \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} 1_{[-\sqrt{1-y^2}, \sqrt{1-y^2}]}(x) dy \\
 &= \frac{1}{\pi} \left[y \right]_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \\
 &= \frac{2}{\pi} \sqrt{1-x^2}
 \end{aligned}$$

3. Tegn $p(x)$ og for tolk den ved at se på tegningen af $p(x, y)$



På tegningen af $p(x, y)$ ser vi, at jo tættere x er på 0, jo flere værdier af y kan den kobles med under betingelsen $x^2 + y^2 \leq 1$. Derfor er sadnsynligheden for x højest for $x = 0$, mens den går mod 0 for $x \rightarrow \pm 1$.

Opgave U43.2.2

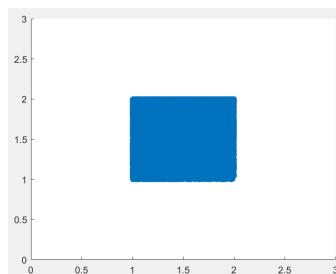
Antag, at X, Y er ligefordelte på mængden

$$A = \{x, y | x \in [1, 2], y \in [1, 2]\}$$

med tætheden

$$p(x, y) = 1_A(x, y)$$

1. Prøv at tegne $p(x, y)$.



2. Find tæthederne for X og Y . Fortolk.

Tæthed for X :

$$\begin{aligned} p(x) &= \int_1^2 1_A(x, y) dy \\ &= 1_{[1,2]}(x) \int_1^2 1_{[1,2]}(y) dy \\ &= 1_{[1,2]}(x) [y]_1^2 \\ &= \underline{\underline{1_{[1,2]}(x)}} \end{aligned}$$

Tæthed for Y :

$$\begin{aligned} p(y) &= \int_1^2 1_A(x, y) dx \\ &= 1_{[1,2]}(y) \int_1^2 1_{[1,2]}(x) dx \\ &= 1_{[1,2]}(y) [x]_1^2 \\ &= \underline{\underline{1_{[1,2]}(y)}} \end{aligned}$$

De marginale tætheder for X og Y er ligefordelinger på $[1, 2]$. Desuden er X og Y uafhængige da $p(x, y) = p(x) \cdot p(y)$.

3. Med $Z = X + Y$, find vha. spørgsmål 2, $E(Z)$ og $V(Z)$.

$$\begin{aligned} E(Z) &= E(X + Y) \\ &= E(X) + E(Y) \\ &= \int_{-\infty}^{\infty} x 1_{[1,2]}(x) dx + \int_{-\infty}^{\infty} y 1_{[1,2]}(y) dy \\ &= \left[\frac{1}{2} x^2 \right]_1^2 + \left[\frac{1}{2} y^2 \right]_1^2 \\ &= \underline{\underline{\frac{3}{2}}} \end{aligned}$$

Da X og Y er identisk fordelte og uafhængige, er $Var(Z) = Var(X + Y) = 2Var(X)$:

$$\begin{aligned} Var(Z) &= 2Var(X) \\ &= 2 \left(\int_{-\infty}^{\infty} x^2 1_{[1,2]}(x) dx - E(X)^2 \right) \\ &= 2 \left(\left[\frac{1}{3} x^3 \right]_1^2 - \left(\frac{3}{2} \right)^2 \right) \\ &= 2 \left(\frac{7}{3} - \left(\frac{3}{2} \right)^2 \right) \\ &= 2 \left(\frac{28 - 27}{12} \right) \\ &= \underline{\underline{\frac{1}{6}}} \end{aligned}$$

4. Find tætheden $q(z)$ for $Z = X + Y$.

Definér $\mathbf{Z} = \begin{pmatrix} Z \\ Z_2 \end{pmatrix}$ som en lineær transformation af vektoren $\begin{pmatrix} X \\ Y \end{pmatrix}$. Formuleres transformasjonen således, at $Z = X + Y$ og $Z_2 = Y$ er den marginale tæthed for Z lig med den tæthed vi

leder efter.

Således:

$$\begin{aligned}\mathbf{Z} &= \begin{pmatrix} Z \\ Z_2 \end{pmatrix} = A \begin{pmatrix} X \\ Y \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}\end{aligned}$$

Ifølge *Sætning 6.3.11* er tætheden for $\mathbf{Z} = \begin{pmatrix} Z \\ Z_2 \end{pmatrix}$ givet ved

$$\begin{aligned}q(z_1, z_2) &= \frac{p(A^{-1}(z, z_2)')}{|det(A)|} \\ &= \frac{p\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}(z, z_2)\right)}{1} \\ &= 1_A(z - z_2, z_2) \\ &= 1_{[1,2]}(z - z_2) \times 1_{[1,2]}(z_2) \\ &= 1_{[2,3]}(z) \times 1_{[1,z-1]}(z_2) + 1_{[3,4]}(z) \times 1_{[z-2,2]}(z_2)\end{aligned}$$

Da vi har fået adskilt Z og Z_2 i tætheden kan vi finde tætheden for $Z = X + Y$ ved at integrere mht. Z_2 .

$$\begin{aligned}q(z) &= \int_1^{z-1} 1_{[2,3]} dz_2 + \int_{z-2}^2 1_{[3,4]} dz_2 \\ &= 1_{[2,3]}\left[z_2\right]_1^{z-1} + 1_{[3,4]}\left[z_2\right]_{z-2}^2 \\ &= \underline{\underline{1_{[2,3]}(z-2) + 1_{[3,4]}(4-z)}}\end{aligned}$$

Prøv at tegne funktionen!

5. Benyt $q(z)$ til direkte at udregne $E(Z)$.

$$\begin{aligned}E(Z) &= \int_{-\infty}^{\infty} z(1_{[2,3]}(z-2) + 1_{[3,4]}(4-z)) dz \\ &= \int_2^3 (z^2 - 2z) dz + \int_3^4 (4z - z^2) dz \\ &= \left[\frac{1}{3}z^3 - z^2\right]_2^3 + \left[2z^2 - \frac{1}{3}z^3\right]_3^4 \\ &= \left(0 - \left(\frac{8}{3} - 4\right)\right) + \left(32 - \frac{64}{3} - (18 - 9)\right) \\ &= \frac{4}{3} + \frac{69 - 64}{3} \\ &= \underline{\underline{\frac{3}{2}}}\end{aligned}$$

6. Hvad er $Cov(X, Z)$?

$$\begin{aligned}
 Cov(X, Z) &= Cov(X, X + Y) \\
 &= E(X(X + Y)) - E(X)E(X + Y) \\
 &= E(X^2) + E(XY) - E(X)^2 - E(X)E(Y) \\
 &= Var(X) + Cov(X, Y) \\
 &= \frac{1}{2}Var(Z) \\
 &= \underline{\underline{\frac{1}{12}}}
 \end{aligned}$$

Opgave U43.2.3

Antag, at X, Y er uafhængige eksponentiaffordelte på $A = (0, \infty)^2$ med $p(x) = \exp(-x)$ og $q(y) = \exp(-y)$

1. Find tætheden $p(x, y)$ for (X, Y) .

Da X og Y er uafgængige er den simultane fordeling lig med produktet af de marginale fordelinger:

$$\begin{aligned}
 p(x, y) &= p(x)p(y) \\
 &= \exp(-x)\exp(-y) \\
 &= \underline{\underline{\exp(-(x+y))}}, \quad x, y \geq 0
 \end{aligned}$$

2. Find tætheden for $Z^+ = X + Y$.

Ifølge korollar 6.3.2 er tætheden r for Z^+ givet ved

$$r(z^+) = \int_{-\infty}^{\infty} p(x)q(z^+ - x)dx$$

Dermed:

$$\begin{aligned}
 r(z^+) &= \int_{-\infty}^{\infty} 1_{(0, z^+)}(x) \exp(-x) \exp(-(z^+ - x)) dx \\
 &= \int_0^{z^+} \exp(-x - z^+ + x) dx \\
 &= \int_0^{z^+} \exp(-z^+) dx \\
 &= \underline{\underline{z^+ \exp(-z^+)}}
 \end{aligned}$$

3. Find tætheden for $Z^- = X - Y$.

Vi definerer $Y^- = -Y$. Tætheden for Y^- er

$$\begin{aligned}
 q(y^-) &= \exp(-y^-) \left| \frac{d}{dy^-} (-y^-) \right| \\
 &= \exp(y^-)
 \end{aligned}$$

Med følgende omskrivning $Z^- = X - Y = Y^- + X$ har vi dermed:

$$\begin{aligned}
 r(z^-) &= \int_{-\infty}^{\infty} 1_{(-\infty,0)}(y^-) \exp(y^-) 1_{\mathbb{R}}(z^-) \exp(-(|z^- - y^-|)) dy^- \\
 &= \int_0^{\infty} \exp(-y) 1_{\mathbb{R}}(z^-) \exp(-(|z^- + y|)) dy \\
 &= \int_0^{\infty} \exp(-y) 1_{(-\infty,0)}(z^-) \exp(z^- - y) dy + \int_0^{\infty} \exp(-y) 1_{(0,\infty)}(z^-) \exp(-(z^- + y)) dy \\
 &= 1_{(-\infty,0)}(z^-) \exp(z^-) \int_0^{\infty} \exp(-2y) dy + 1_{(0,\infty)}(z^-) \exp(-z^-) \int_0^{\infty} \exp(-2y) dy \\
 &= 1_{(-\infty,0)}(z^-) \exp(z^-) \left[-\frac{1}{2} \exp(-2y) \right]_0^{\infty} + 1_{(0,\infty)}(z^-) \exp(-z^-) \left[-\frac{1}{2} \exp(-2y) \right]_0^{\infty} \\
 &= 1_{(-\infty,0)}(z^-) \frac{1}{2} \exp(z^-) + 1_{(0,\infty)}(z^-) \frac{1}{2} \exp(-z^-) \\
 &= \underline{\underline{\frac{1}{2} \exp(-|z^-|)}}
 \end{aligned}$$

Z^- er *Laplace-fordelt* med middelværdi 0 og varians 1.

Oppgave U43.2.4

Lad X_1, X_2, X_3 og X_4 være uafhængige identisk fordelte med middelværdi 5 og varians 9.
Sæt $Y = X_1 + 2X_2 - X_4$

1. Find $E(Y)$

$$\begin{aligned}
 E(Y) &= E(X_1 + 2X_2 - X_4) \\
 &= E(X_1) + 2E(X_2) - E(X_4) \\
 &= 5 + 2 \cdot 5 - 5 \\
 &= \underline{\underline{10}}
 \end{aligned}$$

2. Find $V(Y)$

$$\begin{aligned}
 V(Y) &= V(X_1 + 2X_2 - X_4) \\
 &= V(X_1) + 2^2 V(X_2) + V(X_4) \\
 &= 9 + 4 \cdot 9 + 9 \\
 &= \underline{\underline{54}}
 \end{aligned}$$

Husk, at

$$\begin{aligned}
 V(X + Y) &= E((X + Y)^2) - (E(X + Y))^2 \\
 &= E(X^2 + Y^2 + 2XY) - ((E(X))^2 + (E(Y))^2 + 2E(X)E(Y)) \\
 &= \underbrace{E(X^2)}_{V(X)} - (E(X))^2 + \underbrace{E(Y^2)}_{V(Y)} - (E(Y))^2 + 2 \underbrace{(E(XY) - E(X)E(Y))}_{Cov(X,Y)}
 \end{aligned}$$

$$\begin{aligned}
 V(X - Y) &= E((X - Y)^2) - (E(X - Y))^2 \\
 &= E(X^2 + Y^2 - 2XY) - ((E(X))^2 + (E(Y))^2 - 2E(X)E(Y)) \\
 &= \underbrace{E(X^2)}_{V(X)} - (E(X))^2 + \underbrace{E(Y^2)}_{V(Y)} - (E(Y))^2 - 2 \underbrace{(E(XY) - E(X)E(Y))}_{Cov(X,Y)}
 \end{aligned}$$

Opgave U43.2.5

Antag, at (X, Y) har tæthed $p(x, y) = 24xy$ på

$$A = \{x > 0, y > 0 | x + y < 1\}$$

1. Find $E(XY)$

$$\begin{aligned} E(X, Y) &= \int_0^1 \int_0^{1-y} xy \cdot 24xy \, dx \, dy \\ &= \int_0^1 \int_0^{1-y} 24x^2y^2 \, dx \, dy \\ &= \int_0^1 [8x^3y^2]_{x=0}^{x=1-y} \, dy \\ &= 8 \int_0^1 ((1-y)^3y^2) \, dy \\ &= 8 \int_0^1 ((1-y)^2(1-y)y^2) \, dy \\ &= 8 \int_0^1 ((1+y^2-2y)(y^2-y^3)) \, dy \\ &= 8 \int_0^1 ((y^2-y^3+y^4-y^5-2y^3+2y^4)) \, dy \\ &= 8 \int_0^1 ((-y^5+3y^4-3y^3+y^2)) \, dy \\ &= 8 \left[\left(-\frac{1}{6}y^6 + \frac{3}{5}y^5 - \frac{3}{4}y^4 + \frac{1}{3}y^3 \right) \right]_0^1 \\ &= 8 \left(\frac{-10 + 36 - 45 + 20}{60} \right) \\ &= \frac{8}{60} \\ &= \frac{2}{15} \end{aligned}$$

2. Find $Cov(X, Y)$

For at finde kovariansen skal vi kende $E(X)$ og $E(Y)$ som vi finder ved hjælp af tæthederne

$p(x)$ og $p(y)$ for X og Y .

$$\begin{aligned}
 E(X) &= \int_0^1 x \cdot p(x) dx \\
 &= \int_0^1 x \left(\int_0^{1-x} f(x, y) dy \right) dx \\
 &= \int_0^1 x \left(\int_0^{1-x} 24xy dy \right) dx \\
 &= \int_0^1 x \left[12xy^2 \right]_{y=0}^{y=1-x} dx \\
 &= \int_0^1 x (12x(1+x^2 - 2x)) dx \\
 &= \int_0^1 (12x^2 + 12x^4 - 24x^3) dx \\
 &= \left[4x^3 + \frac{12}{5}x^5 - 6x^4 \right]_0^1 \\
 &= \frac{20 + 12 - 30}{5} \\
 &= \underline{\underline{\frac{2}{5}}}
 \end{aligned}$$

Da x og y indgår på samme måde i $f(x, y)$ er $p(x) = p(y)$ og $E(X) = E(Y)$:

$$\begin{aligned}
 Cov(X, Y) &= E(XY) - E(X)E(Y) \\
 &= \frac{2}{15} - \frac{2}{5} \cdot \frac{2}{5} \\
 &= \frac{10 - 12}{75} \\
 &= \underline{\underline{-\frac{2}{75}}}
 \end{aligned}$$

3. Udregn korrelationen mellem X og Y

For at beregne korrelationen skal vi kende variansen af X og Y :

$$\begin{aligned}
 V(X) = V(Y) &= \int_0^1 x^2 \left(\int_0^{1-x} f(x, y) dy \right) dx - (E(X))^2 \\
 &= \int_0^1 x^2 \left(\int_0^{1-x} 24xy dy \right) dx - \left(\frac{2}{5}\right)^2 \\
 &= \int_0^1 x^2 \left[12xy^2 \right]_{y=0}^{y=1-x} dx - \left(\frac{2}{5}\right)^2 \\
 &= \int_0^1 x^2 (12x(1+x^2-2x)) dx - \left(\frac{2}{5}\right)^2 \\
 &= \int_0^1 (12x^3 + 12x^5 - 24x^4) dx - \left(\frac{2}{5}\right)^2 \\
 &= \left[3x^4 + 2x^6 - \frac{24}{5}x^5 \right]_0^1 - \left(\frac{2}{5}\right)^2 \\
 &= \left(\frac{15}{5} + \frac{10}{5} - \frac{24}{5} \right) - \left(\frac{2}{5}\right)^2 \\
 &= \frac{5-4}{25} \\
 &= \frac{1}{25}
 \end{aligned}$$

$$\begin{aligned}
 Corr(X, Y) &= \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}} \\
 &= \frac{\frac{-2}{75}}{\frac{1}{25}} \\
 &= \frac{-50}{75} \\
 &= -\frac{2}{3}
 \end{aligned}$$

4. Udregn $Cov(3X, 5Y)$

$$\begin{aligned}
 Cov(3X, 5Y) &= 3 \cdot 5 Cov(X, Y) \\
 &= 15 \cdot \left(-\frac{2}{75} \right) \\
 &= -\frac{2}{5}
 \end{aligned}$$

5. Udregn $Cov(X + 1, Y - 2)$

$$\begin{aligned}
 Cov(X + 1, Y - 2) &= 1 \cdot 1 Cov(X, Y) \\
 &= -\frac{2}{75}
 \end{aligned}$$

6. Udregn $Cov(X + 1, 5Y - 4)$

$$\begin{aligned}
 Cov(X + 1, 5Y - 4) &= 1 \cdot 5 Cov(X, Y) \\
 &= 5 \cdot \left(-\frac{2}{75} \right) \\
 &= -\frac{2}{15} \\
 &\underline{\underline{-}}
 \end{aligned}$$

7. Udregn $Cov(3X + 5, X)$

$$\begin{aligned}
 Cov(3X + 5, X) &= 3 \cdot 1 Cov(X, X) \\
 &= 3 Var(X) \\
 &= 3 \cdot \frac{1}{25} \\
 &= \frac{3}{25} \\
 &\underline{\underline{-}}
 \end{aligned}$$

Opgave 6.4

Lad (X, Y) være en kontinuert to-dimensonal stokastisk vektor med sandsynlighedstæthed

$$p(x, y) = \begin{cases} 3xy^{-2} & \text{for } x \in (0, 1) \text{ og } y \in (1, 3) \\ 0 & \text{ellers.} \end{cases}$$

Find de marginale sandsynlighedstæther for X og Y , og vis, at X og Y er uafhængige.

Marginal sandsynlighedstæthed p for X :

$$\begin{aligned}
 p(x) &= \mathbb{I}_{x \in (0, 1)} \int_1^3 3xy^{-2} dy \\
 &= \mathbb{I}_{x \in (0, 1)} \left[-3xy^{-1} \right]_{y=1}^{y=3} \\
 &= \mathbb{I}_{x \in (0, 1)} (3x - x) \\
 &= \begin{cases} 2x & \text{for } x \in (0, 1) \\ 0 & \text{ellers} \end{cases} \\
 &\underline{\underline{-}}
 \end{aligned}$$

Marginal sandsynlighedstæthed p for Y :

$$\begin{aligned}
 p(y) &= \mathbb{I}_{y \in (1, 3)} \int_0^1 3xy^{-2} dx \\
 &= \mathbb{I}_{y \in (1, 3)} \left[\frac{3}{2}x^2y^{-2} \right]_{x=0}^{x=1} \\
 &= \mathbb{I}_{y \in (1, 3)} \left(\frac{3}{2}y^{-2} \right) \\
 &= \begin{cases} \frac{3}{2}y^{-2} & \text{for } y \in (1, 3) \\ 0 & \text{ellers} \end{cases} \\
 &\underline{\underline{-}}
 \end{aligned}$$

Da det gælder, at

$$p(x) \cdot p(y) = \begin{cases} 2x \cdot \frac{3}{2}y^{-2} & \text{for } x \in (0, 1) \text{ og } y \in (1, 3) \\ 0 & \text{ellers} \end{cases} = \begin{cases} 3xy^{-2} & \text{for } x \in (0, 1) \text{ og } y \in (1, 3) \\ 0 & \text{ellers} \end{cases} = p(x, y)$$

er det vist, at X og Y er uafhængige.

Opgave 6.21

Lad X være ligefordelt i intervallet $(-1, 1)$, og definer $Y = X^2$. Vis, at X og Y er ukorrelerede. Er de uafhængige?

Sandsynlighedstætheden for X er $p(x) = \frac{1_{(-1,1)}(x)}{2}$. Da $f : x \rightarrow x^2$ ikke er monoton på $(-1, 1)$ kan vi ikke bruge (5.4.1) eller (5.4.1) til at udlede sandsynlighedstætheden for Y .

Vi udleder istedet sandsynlighedstætheden q for Y ved hjælp af fordelingsfunktionerne for X og Y , F_X og F_Y :

$$\begin{aligned} F_Y &= P(Y \leq y) = P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= P(X \leq \sqrt{y}) - P(X \leq -\sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

Og idet p angiver sandsynlighedstætheden for X , kan vi ved at differentiere finde sandsynlighedstætheden q for Y :

$$\begin{aligned} q(y) &= F'_Y(y) = \frac{d}{dy} F_X(\sqrt{y}) - \frac{d}{dy} F_X(-\sqrt{y}) \\ &= \frac{p(\sqrt{y})}{2\sqrt{y}} + \frac{p(-\sqrt{y})}{2\sqrt{y}} \\ &= \frac{p(\sqrt{y}) + p(-\sqrt{y})}{2\sqrt{y}} \\ &= \frac{1_{(-1,1)}(\sqrt{y}) + 1_{(-1,1)}(-\sqrt{y})}{4\sqrt{y}} \\ &= \frac{2 \times 1_{(0,1)}(\sqrt{y})}{4\sqrt{y}} \\ &= \frac{1_{(0,1)}(y)}{2\sqrt{y}} \end{aligned}$$

X og Y er ukorrelerede hvis, og kun hvis, $Cov(X, Y) = 0$:

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ &= \int_0^1 \int_{-1}^1 \left(\frac{x}{2}\right) \left(\frac{y}{2\sqrt{y}}\right) dx dy - \int_{-1}^1 \frac{x}{2} dx \int_0^1 \frac{y}{2\sqrt{y}} dy \\ &= \int_0^1 \left(\frac{y}{2\sqrt{y}}\right) \underbrace{\left[\frac{x^2}{4}\right]_{-1}^1}_{=0} dy - \underbrace{\left[\frac{x^2}{4}\right]_{-1}^1}_{=0} \frac{1}{3} \left[y^{\frac{3}{2}}\right]_0 \\ &= \int_0^1 0 dy \\ &= 0 \end{aligned}$$

X og Y er altså ukorrelerede, men tydeligvis ikke uafhængige da Y er en funktion af X .

Afleveringsopgave til uge 44

Vi skal benytte Korollar 6.3.2 til at vise at med X og Y uafhængige standard normalfordelte ($N(0, 1)$) så er $Z = X + Y$ normalfordelt med middelværdi 0 og varians 2, $N(0, 2)$.

1. Vis, at tæthedden for (X, Y) kan skrives som,

$$p(x, y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2)\right)$$

Benyt, at X og Y er uafhængige til at skrive tæthedden for (X, Y) som produktet af tæthedsfunktionerne for X og Y .

$$\begin{aligned} p(x, y) &= p(x) \cdot p(y) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{x^2}{2}\right) \exp\left(-\frac{y^2}{2}\right) \\ &= \underline{\underline{\frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2)\right)}} \end{aligned}$$

2. Vis, at tæthedden for Z kan skrives som,

$$q(z) = \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}((z-w)^2 + w^2)\right) dw$$

Benyt Korollar 6.3.2 og sæt $X = W$:

$$\begin{aligned} q(z) &= \int_{-\infty}^{\infty} p_X(w) \cdot p_Y(z-w) dw \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-w)^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw \\ &= \underline{\underline{\frac{1}{2\pi} \exp\left(-\frac{1}{2}((z-w)^2 + w^2)\right)}} \end{aligned}$$

Bemærk at samme resultat opnås ved at sætte $Y = W$.

3. Vis, at $q(z)$ kan opskrives som

$$q(z) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{1}{4}z^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw$$

Lav omregninger af $q(z)$ og benyt, at z kan betragtes som en konstant når der integreres med hensyn til w :

$$\begin{aligned} q(z) &= \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}((z-w)^2 + w^2)\right) dw \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}(z^2 + w^2 - 2zw + w^2)\right) dw \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi}} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw \\ &= \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{4}z^2 + \frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw \\ &= \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{4}z^2\right) \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw \\ &= \underline{\underline{\frac{1}{\sqrt{4\pi}} \exp\left(-\frac{1}{4}z^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw}} \end{aligned}$$

4. Vis, at

$$\frac{1}{\sqrt{4\pi}} \exp\left(-\frac{1}{4}z^2\right)$$

er tæthed for $N(0, 2)$ fordelingen

Hvis vi kan vise, at

$$\frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right)$$

er en tæthed for w , så gælder det, at

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw = 1.$$

Vi omregner udtrykket:

$$\begin{aligned} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) &= \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{4}z^2 + zw - w^2\right) \\ &= \frac{1}{\sqrt{\pi}} \exp\left(-\left(\frac{1}{2}z - w\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi(\frac{1}{2})}} \exp\left(-\frac{(w - \frac{1}{2}z)^2}{2 \cdot \frac{1}{2}}\right) \end{aligned}$$

og ser, at det kan udtrykkes som en tæthed for $w \sim N\left(\frac{1}{2}z, \frac{1}{2}\right)$.

Dermed gælder det for $Z \sim N(0, 2)$:

$$\begin{aligned} q(z) &= \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{1}{4}z^2\right) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw}_{=1} \\ &= \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{1}{4}z^2\right) \end{aligned}$$

Dette kunne også vises nemt ved at sætte $\mu = 0$ og $\sigma^2 = 2$ ind i formlen for normalfordelingens tæthed:

$$\begin{aligned} q(z) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{1}{4}z^2\right) \end{aligned}$$

5. Givet 4. skal vi nu vise $m = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw = 1$.

Omskriv udtrykket til

$$m = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{2}\left(\sqrt{2}w - \frac{1}{\sqrt{2}}z\right)^2\right) dw.$$

Vis, at

$$m = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}v^2\right) dv$$

og argumentér for, at $m=1$.

$$\begin{aligned} m &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(\frac{1}{4}z^2 - \frac{1}{2}(z^2 - 2zw + 2w^2)\right) dw \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{2}(\frac{1}{2}z^2 - 2zw + 2w^2)\right) dw \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{2}\left(\sqrt{2}w - \frac{1}{\sqrt{2}}z\right)^2\right) dw \end{aligned}$$

Nu definerer vi

$$\begin{aligned} v &= \sqrt{2}w - \frac{1}{\sqrt{2}}z \Leftrightarrow \\ \frac{dv}{dw} &= \sqrt{2} \Leftrightarrow \\ dw &= \frac{1}{\sqrt{2}}dv \end{aligned}$$

og substituerer ind i udtrykket for m :

$$\begin{aligned} m &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\frac{1}{2}v^2\right) \frac{1}{\sqrt{2}} dv \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}v^2\right) dv \text{ (standard normalfordelingen integreret på } \mathbb{R}) \\ &= 1 \end{aligned}$$

6. Mere generelt kan man vise, at med X_1, \dots, X_M uafhængige $N(0, 1)$ fordelte så er $Z = \sum_{i=1}^M X_i$ normalfordelt med middelværdi 0 og varians M , $N(0, M)$. Vis, at

$$\frac{Z}{M} = \frac{1}{M} \sum_{i=1}^M X_i \sim N\left(0, \frac{1}{M}\right).$$

Hvad sker der når $M \rightarrow \infty$?

Da X_1, \dots, X_M er uafhængige $N(0, 1)$ fordelte gælder

$$\begin{aligned} E\left(\sum_{i=1}^M X_i\right) &= \sum_{i=1}^M E(X_i) = 0 \\ Var\left(\sum_{i=1}^M X_i\right) &= \sum_{i=1}^M Var(X_i) = M \\ Z &= \sum_{i=1}^M X_i \sim N(0, M) \end{aligned}$$

For $\frac{Z}{M}$ gælder

$$\begin{aligned} E\left(\frac{Z}{M}\right) &= \frac{E(Z)}{M} = 0 \\ Var\left(\frac{Z}{M}\right) &= \frac{Var(Z)}{M^2} = \frac{1}{M} \\ \frac{Z}{M} &\sim N\left(0, \frac{1}{M}\right) \end{aligned}$$

$\frac{Z}{M} = \frac{1}{M} \sum_{i=1}^M X_i$ er gennemsnittet, \bar{X}_M af M uafhængige realisationer af en stokastisk variabel X . Vores udregninger viser os, at dette gennemsnit har fordelingen

$$\frac{Z}{M} = \bar{X}_M \sim N\left(E(X), \frac{Var(X)}{M}\right)$$

således, at $Var\left(\frac{Z}{M}\right) \rightarrow 0$ for $M \rightarrow \infty$.

Et gennemsnit af en stokastisk variabel går altså mod en normalfordeling med den sande middelværdi og en varians lig 0 når antallet af realistationer i gennemsnittet går mod uendeligt.

Dette er indholdet i *Den centrale grænseværdisætning* som også viser, at gennemsnittet af stokastiske variable der ikke er normalfordelt går mod en sådan normalfordeling.

Opgave 6.14

Lad X_1 og X_2 være uafhængige stokastiske variable, som begge er eksponentiaffordelte med parameter 1.

- (a) Find tætheden for den stokastiske vektor $(X_1 + X_2, X_1)$.

Tætheden p for (X_1, X_2) er

$$\begin{aligned} p(x_1, x_2) &= 1_{(0, \infty)}(x_1) \exp(-x_1) \cdot 1_{(0, \infty)}(x_2) \exp(-x_2) \\ &= 1_{(0, \infty)^2}(x_1, x_2) \exp(-(x_1 + x_2)) \end{aligned}$$

Den stokastiske vektor $(X_1 + X_2, X_1)$ er dannet ved følgende lineære transformation af (X_1, X_2) :

$$\begin{pmatrix} X_1 + X_2 \\ X_1 \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$$

Ifølge *Sætning 6.3.11* er tætheden q for $(X_1 + X_2, X_1)$, hvor p er tætheden for (X_1, X_2) , givet ved

$$\begin{aligned} q(x_1 + x_2, x_1) &= \frac{p\left(A^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)}{|det(A)|} \\ &= \frac{p\left(\begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)}{1} \\ &= p(x_2, x_1 - x_2) \\ &= 1_{(0, \infty)^2}(x_1, x_2) \exp(-(x_2 + x_1 - x_2)) \\ &= \underline{\underline{1_{(0, \infty)}(x_1) \exp(-x_1)}} \end{aligned}$$

- (b) Vis, at

$$\frac{X_1}{X_1 + X_2}$$

er ligefordelt på $(0, 1)$. Hvad er fordelingen af $\frac{X_1}{X_1 + X_2}$, hvis de to eksponentiaffordelinger har parameter λ ?

Ifølge *Korollar 6.3.6* har $Z = \frac{X_1}{X_1 + X_2}$ tætheden

$$q(z) = \int_0^\infty p(x_1 + x_2, z(x_1 + x_2))(x_1 + x_2) d(x_1 + x_2),$$

hvor p her betegner tætheden for $(X_1 + X_2, X_1)$.

Da $z(x_1 + x_2) = x_1$ og $p(x_1 + x_2, x_1) = 1_{(0, \infty)}(x_1) \exp(-x_1)$ får vi

$$\begin{aligned} q(z) &= \int_0^\infty \exp(-x_1)(x_1 + x_2) d(x_1 + x_2) \\ &= \int_0^\infty \exp(-(x_1 + x_2 - x_2))(x_1 + x_2) d(x_1 + x_2) \\ &= \lim_{n \rightarrow \infty} \left[-\exp(-(x_1 + x_2 - x_2))(x_1 + x_2) + \int \exp(-(x_1 + x_2 - x_2)) d(x_1 + x_2) \right]_{x_1 + x_2 = 0}^{x_1 + x_2 = n} \\ &= \lim_{n \rightarrow \infty} \left[-\exp(-(x_1 + x_2 - x_2))(x_1 + x_2) - \exp(-(x_1 + x_2 - x_2)) \right]_{x_1 + x_2 = 0}^{x_1 + x_2 = n} \\ &= 1 \quad (\text{da } (x_1, x_2) \in (0, \infty)^2 \text{ gælder } x_1 + x_2 = 0 \Leftrightarrow x_1 = 0 \wedge x_2 = 0) \end{aligned}$$

Da $q\left(\frac{x_1}{x_1+x_2}\right)$ er lig 1 på $(0, 1)$ er det hermed vist at den er ligefordelt på dette interval.

Hvis de to eksponentialfordelinger har parameter λ er tætheden p for (X_1, X_2)

$$p(x_1, x_2) = \lambda^2 \exp(-\lambda(x_1 + x_2))$$

og tætheden q for $(x_1 + x_2, x_1)$ er

$$q(x_1 + x_2, x_1) = 1_{(0, \infty)}(x_1) \lambda^2 \exp(-x_1)$$

Udledningen af fordelingen for $Z = \frac{X_1}{X_1+X_2}$ bliver da

$$\begin{aligned} q(z) &= \int_0^\infty \lambda^2 \exp(-x_1) (x_1 + x_2) d(x_1 + x_2) \\ q(z) &= \lambda^2 \int_0^\infty \exp(-x_1) (x_1 + x_2) d(x_1 + x_2) \\ &= \underline{\underline{\lambda^2}} \end{aligned}$$

Centrale begreber

- Ligefordelingen på $B \in \mathbb{R}^n$ $\left(p(x_1, \dots, x_n) = \frac{1_B(x_1, \dots, x_n)}{|B|}\right)$ (6.1.2)

- Middelværdi af en kontinuert stokastisk variabel X :

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx \quad (5.2.2)$$

- Lad (X, Y) være en to-dimensional kontinuert stokastisk vektor med sandsynlighedstæthed $p(x, y)$.

Den marginale sandsynlighed q for X er:

$$q(x) = \int_{\mathbb{R}} p(x, y) dy \quad (6.1.5)$$

- Lad (X, Y) være to uafhængige kontinuerte stokastiske variable med marginale tætheder p_1 og p_2 .

$Z = X + Y$ er da en kontinuert stokastisk variabel med sandsynlighedstæthed:

$$q(z) = \int_{-\infty}^{\infty} p_1(x) p_2(z - x) dx \quad (6.3.2)$$

- Regneregler for kovarianser:

$$Cov(X, Y) = E(XY) - E(X)E(Y) \quad (3.8.2)$$

$$Cov(a + bX, c + dY) = bd Cov(X, Y) \quad (3.8.3)$$

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \quad (3.8.4)$$

$$Cov(X, Y) = Cov(Y, X) \quad (3.8.5)$$

$$Cov(X, X) = Var(X)$$

Opgave U44.1.1

Betrægt kast med to uafhængige terninger (så ikke en kontinuert fordeling). Lad de to kast være repræsenteret ved de stokastiske variable X_1 og X_2 , hvor

$$(X_1, X_2) \in \{1, 2, \dots, 6\}^2 = \{x_{1,i}\} \times \{x_{2,j}\}_{i,j=1,2,\dots,6}$$

1. Lad $Z = X_1 + X_2$. Find $P(X_1 = i|Z \geq 4)$ for $i = 1, 2, \dots, 6$.

Husk at sandsynlighed for et givet udfald med de to terninger er

$$P(X_1 = i, X_2 = j) = \frac{1}{36}, \quad i, j = 1, 2, \dots, 6.$$

Desuden er sandsynligheden for $Z \geq 4$ givet ved

$$P(Z \geq 4) = 1 - P(X_1 + X_2 \leq 3) = \frac{33}{36}.$$

Således (see også *Property D.1* i *Rahbek: "On conditional expectations"*):

$$\begin{aligned} P(X_1 = 1|Z \geq 4) &= \frac{P(X_1 = 1, Z \geq 4)}{P(Z \geq 4)} = \frac{4/36}{33/36} = \frac{4}{33} \\ P(X_1 = 2|Z \geq 4) &= \frac{P(X_1 = 2, Z \geq 4)}{P(Z \geq 4)} = \frac{5/36}{33/36} = \frac{5}{33} \\ P(X_1 = k|Z \geq 4) &= \frac{P(X_1 = k, Z \geq 4)}{P(Z \geq 4)} = \frac{6/36}{33/36} = \frac{6}{33}, \quad k = 3, 4, 5, 6. \end{aligned}$$

Fordelingen af $X_1|Z \geq 4$ er altså:

$$P(X_1 = i|Z \geq 4) = \begin{cases} \frac{4}{33}, & i = 1 \\ \frac{5}{33}, & i = 2 \\ \frac{6}{33}, & i \geq 3 \end{cases}$$

2. Brug dette til at udregne $E(X_1|z \geq 4)$.

Se *Definition 1* i *Rahbek: "n conditional expectations"*:

$$\begin{aligned} E(X_1|Z \geq 4) &= \sum_{i=1}^6 x_{1,i} P(X_1 = x_{1,i}|Z \geq 4) \\ &= \frac{1 \cdot 4 + 2 \cdot 5 + (3 + 4 + 5 + 6) \cdot 6}{33} \\ &= \frac{122}{33} \approx 3,7 \end{aligned}$$

Opgave U44.1.2

Betrægt to kommuner hver med deres fordeling af velhavere hhv. ikke-velhavere. Lad $Z \in \{1, 2\}$ angive kommunen og $V \in \{0, 1\}$ angive om man er velhavende i en given kommune. Sandsynligheden for at være velhavende i kommune 1 angives som

$$P(V = 1|Z = 1) = 0,8 = 1 - P(V = 0|Z = 1),$$

mens tilsvarende $P(V = 1|Z = 2) = 0,1$. I den ene kommune er der altså større sandsynlighed for at være velhavende end i den anden.

1. *Udregn*

$$E(V|Z = 1) \text{ og } E(V|Z = 2).$$

Forklar hvad disse udtryk betyder.

$$\begin{aligned} E(V|Z = 1) &= 0 \cdot P(V = 0|Z = 1) + 1 \cdot P(V = 1|Z = 1) = \underline{\underline{0,8}} \\ E(V|Z = 2) &= 0 \cdot P(V = 0|Z = 2) + 1 \cdot P(V = 1|Z = 2) = \underline{\underline{0,2}}. \end{aligned}$$

For en binær variabel $V \in \{0, 1\}$ er sandsynligheden for $V = 1$ lig med den forventede værdi af V .

2. *Vis, at man kan skrive*

$$E(V|Z = z) = f(z) = 0,8 \cdot \mathbf{1}(z = 1) + 0,1 \cdot \mathbf{1}(z = 2),$$

hvor $\mathbf{1}(z = 1) = 1$, hvis $z = 1$ og ellers lig med nul.

Det følger af definitionen på indikatorvariablen at $E(V|Z = z)$ kun kan være enten 0,8 eller 0,1.

3. *Hvad udtrykker $E(V|Z = z)$, og hvad betyder det at udtrykket afhænger af z som angivet ved funktionen $f(z)$?*

$E(V|Z = z)$ udtrykker den forventede værdi af den stokastiske variabel V , givet at vi allerede kender værdien af Z . Dermed er den betingede forventede værdi af V en funktion af z , som er realisationen af Z . Hvis V og Z ikke er uafhængige, så vil den forventede værdi af V ændre sig med z .

4. *Man definerer nu den stokastiske variabel, kaldet den "betingede middelværdi af V givet Z "*

$$E(V|Z) = f(Z).$$

Vis, at

$$E(f(Z)) = E(E(V|Z)) = 0,8 \cdot P(Z = 1) + 0,1 \cdot P(Z = 2)$$

$f(Z)$ antager værdien 0,8, hvis $Z = 1$. Dette sker med sandsynlighed $P(Z = 1)$. På samme måde antager $f(Z)$ værdien 0,1, hvis $Z = 2$, hvilket sker med sandynlighed $P(Z = 2)$. Dermed er den forventede værdi af $f(Z)$ givet ved

$$E(f(Z)) = E(E(V|Z)) = 0,8 \cdot P(Z = 1) + 0,1 \cdot P(Z = 2) = E(V).$$

At $(E(V|Z)) = E(V)$ kaldes "Law of iterated expectations" og er et vigtigt resultat i sandsynligedsteori.

Opgave U44.1.3

Antag, at X er ligefordelt på $A = [0, 10]$.

- Opskriv tætheden $p(x)$ for X , og vis at $P(X > 5) = \frac{1}{2}$.

Tætheden er ifølge (6.1.2) i bogen:

$$p(x) = \frac{\mathbf{1}_{(0,10)}(x)}{10}.$$

Dermed:

$$\begin{aligned} P(X > 5) &= \int_5^{10} \frac{\mathbf{1}_{(0,10)}(x)}{10} dx \\ &= \frac{1}{10} [x]_5^{10} \\ &= \frac{1}{2} \end{aligned}$$

- Vis, at $E(X) = 5$ ved formelt at udregne det som integral og ved at se på tegningen af $p(x)$.

Middelværdien beregnes ved (5.2.2) i bogen:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \frac{\mathbf{1}_{(0,10)}(x)}{10} dx \\ &= \int_0^{10} x \frac{\mathbf{1}_{(0,10)}(x)}{10} dx \\ &= \frac{1}{10} \left[\frac{1}{2} x^2 \right]_0^{10} \\ &= 5 \end{aligned}$$

- Vis, at tætheden for X givet at $X > 5$ kan skrives som $q(x) = \frac{2}{10} \mathbf{1}_{(0,10)}(x)$. Vis det formelt og vis det ved at se på tegningen af $p(x)$.

Benyt Definition 9 i Rahbek: "On conditional expectations":

$$\begin{aligned} q(x) &= \frac{\frac{\mathbf{1}_{(0,10)}(x)}{10}}{\frac{1}{2}} \\ &= \frac{2}{10} \mathbf{1}_{(0,10)}(x) \end{aligned}$$

- Er $E(X|X > 5) = 7,5$?

$$\begin{aligned} E(X|X > 5) &= \frac{1}{5} \int_5^{10} x dx \\ &= \frac{1}{10} [x^2]_5^{10} \\ &= \frac{75}{10} \\ &= 7,5 \end{aligned}$$

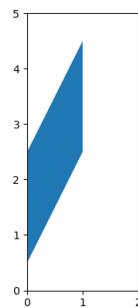
Opgave U44.1.4

På et marked for obligationer antager vi, at ”ratingen” (her et indeks mellem 0 og 1) X og prisen (i 1000\$) Y kan beskrives ved en simultan fordeling. Vi antager, at (X, Y) er ligefordelt på følgende mængde:

$$B = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x \leq 1; 0,5 + 2x \leq y \leq 2,5 + 2x\}$$

- Opskriv tæthedsfunktionen $f_{X,Y}(x, y)$ for den simultane fordeling af (X, Y) .

(X, Y) er ligefordelt på B :



som har arealet $A = 1 \cdot 2 = 2$. Ifølge (6.1.2) i bogen har (X, Y) dermed tætheden

$$\begin{aligned} p(x, y) &= \frac{1_B(x, y)}{|B|} \\ &= \underline{\underline{\frac{1_B(x, y)}{2}}} \end{aligned}$$

hvor $|B|$ betegner det n -dimensionale volumen af B . Da $n = 2$ i dette tilfælde er $|B| = A$ arealet af B .

- Udregn sandsynligheden for, at obligations-prisen er højere end 2000\$, dvs. $P(Y > 2)$.

På tegningen af B er det nemt at aflæse $P(Y \leq 2)$ som den proportion af arealet hvor $Y \leq 2$:

$$\begin{aligned} P(Y \leq 2) &= \frac{\frac{3}{2} \cdot \frac{3}{4} \cdot \frac{1}{2}}{2} \\ &= \frac{9}{32} \Leftrightarrow \\ P(Y > 2) &= 1 - \frac{9}{32} \\ &\approx \underline{\underline{0,7188}} \end{aligned}$$

Sandsynligheden kan også udregnes ved at integrere med de relevante grænser. Regner vi først $P(Y \leq 2)$ ved vi at følgende skal gælde:

$$\begin{aligned} 0,5 + 2x \leq y \leq 2 &\Leftrightarrow \\ x \leq 0,75 \end{aligned}$$

Dermed får vi

$$\begin{aligned}
 P(Y \leq 2) &= \int_0^{\frac{3}{4}} \int_{0,5+2x}^2 \frac{1}{2} dy dx \\
 &= \frac{1}{2} \int_0^{\frac{3}{4}} \left[y \right]_{y=0,5+2x}^{y=2} dx \\
 &= \frac{1}{2} \int_0^{\frac{3}{4}} \left(\frac{3}{2} - 2x \right) dx \\
 &= \frac{1}{2} \left[\frac{3}{2}x - x^2 \right]_{x=0}^{x=\frac{3}{4}} \\
 &= \frac{1}{2} \left(\frac{9}{8} - \frac{9}{16} \right) \\
 &= \frac{9}{32} \Leftrightarrow \\
 P(Y > 2) &= 1 - \frac{9}{32} \\
 &= \frac{23}{32}
 \end{aligned}$$

3. Udregn den marginale fordeling af prisen Y og angiv tæthedsfunktionen $f_Y(y)$ for Y .

For at kunne integrere X ud af den simultane sandsynlighed for (X, Y) og dermed udregne den marginale fordeling af Y skal vi definere grænserne for X . På tegningen af B kan vi se at den funktionen der udspænder parallelogrammet ændrer sig ved $y = 2,5$. Vi er derfor nødt til at definere tætheden for Y for $0,5 \leq Y \leq 2,5$ og $2,5 \leq Y \leq 4,5$.

Når $0,5 \leq Y \leq 2,5$ gælder

$$\begin{aligned}
 0,5 + 2x &\leq y \Leftrightarrow \\
 x &\leq \frac{y - 0,5}{2}
 \end{aligned}$$

Dermed er tætheden p for Y

$$\begin{aligned}
 p(y) &= \int_0^{\frac{y-0,5}{2}} \frac{1}{2} dx \\
 &= \left[\frac{1}{2}x \right]_{x=0}^{x=\frac{y-0,5}{2}} \\
 &= \frac{y - 0,5}{4}, \quad \text{for } 0,5 \leq y \leq 2,5
 \end{aligned}$$

Når $2,5 \leq Y \leq 4,5$ gælder

$$\begin{aligned}
 y &\leq 2,5 + 2x \Leftrightarrow \\
 x &\geq \frac{y - 2,5}{2}
 \end{aligned}$$

Dermed er tætheden p for Y

$$\begin{aligned}
 p(y) &= \int_{\frac{y-2,5}{2}}^1 \frac{1}{2} dx \\
 &= \left[\frac{1}{2}x \right]_{x=\frac{y-2,5}{2}}^{x=1} \\
 &= \frac{1}{2} - \frac{y - 2,5}{4} \\
 &= \frac{4,5 - y}{4}, \quad \text{for } 2,5 \leq y \leq 4,5
 \end{aligned}$$

Tæthedsfunktion for Y :

$$f_Y(y) = \begin{cases} \frac{y-0,5}{4}, & \text{for } 0,5 \leq y \leq 2,5 \\ \frac{4,5-y}{4}, & \text{for } 2,5 \leq y \leq 4,5 \end{cases}$$

For at tjekke at vi har fundet det rigtige udtryk for $f_Y(y)$ sikrer vi at integralet af $f_Y(y)$ er lig 1 for $0,5 \leq y \leq 4,5$:

$$\begin{aligned} \int_{0,5}^{4,5} f_Y(y) dy &= \int_{0,5}^{2,5} \frac{y-0,5}{4} dy + \int_{2,5}^{4,5} \frac{4,5-y}{4} dy \\ &= \left[\frac{1}{8}y^2 - \frac{1}{8}y \right]_{0,5}^{2,5} + \left[\frac{9}{8}y - \frac{1}{8}y^2 \right]_{2,5}^{4,5} \\ &= \left(\frac{25}{32} - \frac{5}{16} - \frac{1}{32} + \frac{1}{16} \right) + \left(\frac{81}{16} - \frac{81}{32} - \frac{45}{16} + \frac{25}{32} \right) \\ &= \frac{16}{32} + \frac{16}{32} \\ &= 1 \end{aligned}$$

4. Angiv den betingede fordeling af X givet $Y = 1$, ved at angive tæthedsfunktionen $f_{X|Y=1}(x)$

Hvis $Y = 1$ gælder $X \leq 0,25$. Betingelsen $Y = 1$ omdanner med andre ord X til at være en ligefordelt variabel på $(0; 0,25)$.

Dermed

$$\begin{aligned} f_{X|Y=1}(x) &= \frac{1_{(0;0,25)}(x)}{0,25 - 0} \\ &= \underline{\underline{4 \cdot 1_{(0;0,25)}(x)}} \end{aligned}$$

5. Udregn den forventede rating af en obligation der koster 1000\$ og en som koster 2000\$. Dvs. udregn de to betingede middelværdier $E(X|Y = 1)$ og $E(X|Y = 2)$.

Definition 8 i Rahbek: "On conditional expectations" lyder:

Den betingede forventning $E(X|Y = y)$ er en funktion af y og er givet ved

$$E(X|Y = y) = \int_{\mathbb{R}} x f_{X|Y=y}(x|y) dx$$

Dermed

$$\begin{aligned} E(X|Y = 1) &= \int_{\mathbb{R}} x f_{X|Y=1}(x) dx \\ &= \int_{-\infty}^{\infty} x \cdot 4 \cdot 1_{(0;0,25)}(x) dx \\ &= \int_0^{0,25} 4x dx \\ &= [2x^2]_0^{0,25} \\ &= \underline{\underline{0,125}} \end{aligned}$$

For at beregne $E(X|Y = 2)$ skal vi kende $f_{X|Y=2}(x)$. Hvis $Y = 2$ er X ligefordelt på $(0; 0, 75)$.
Dermed

$$\begin{aligned} f_{X|Y=2}(x) &= \frac{1_{(0;0,75)}(x)}{0,75 - 0} \\ &= \frac{4}{3} \cdot 1_{(0;0,75)}(x) \\ E(X|Y = 2) &= \int_{\mathbb{R}} x f_{X|Y=2}(x) dx \\ &= \int_{-\infty}^{\infty} x \cdot \frac{4}{3} \cdot 1_{(0;0,75)}(x) dx \\ &= \int_0^{0,75} \frac{4}{3} x dx \\ &= \left[\frac{2}{3} x^2 \right]_0^{0,75} \\ &= \frac{9}{16} \cdot \frac{2}{3} \\ &= \underline{\underline{0,375}} \end{aligned}$$

6. Udregn variansen af ratingen på en obligation som koster hhv. 1000\$ og 2000\$, dvs. udregn $Var(X|Y = 1)$ og $Var(X|Y = 2)$.

Den betingede varians er givet ved

$$Var(X|Y = y) = E(X^2|Y = y) - E(X|Y = y)^2$$

Dermed

$$\begin{aligned} Var(X|Y = 1) &= E(X^2|Y = 1) - E(X|Y = 1)^2 \\ &= \int_0^{0,25} 4x^2 - 0,125^2 \\ &= \left[\frac{4}{3} x^3 \right]_0^{0,25} - 0,125^2 \\ &= \frac{4}{3} \cdot \frac{1}{64} - \frac{1}{64} \\ &= \underline{\underline{\frac{1}{192}}} \end{aligned}$$

$$\begin{aligned} Var(X|Y = 2) &= E(X^2|Y = 2) - E(X|Y = 2)^2 \\ &= \int_0^{0,75} \frac{4}{3} x^2 - 0,375^2 \\ &= \left[\frac{4}{9} x^3 \right]_0^{0,75} - 0,375^2 \\ &= \frac{4}{9} \cdot \frac{27}{64} - \frac{9}{64} \\ &= \underline{\underline{\frac{3}{64}}} \end{aligned}$$

7. For hvilken pris er der størst variation i ratingen?

For at udlede et generelt udtryk for den betingede varians af $X|Y = y$ er vi nødt til at udlede

$f_{X|Y=y}(x)$.

For $0,5 \leq Y \leq 2,5$ gælder

$$0,5 + 2x \leq y \Leftrightarrow \\ x \leq \frac{y - 0,5}{2}$$

så X er ligefordelt på $\left(0; \frac{y-0,5}{2}\right)$.

For $2,5 \leq Y \leq 4,5$ gælder

$$2,5 + 2x \geq y \Leftrightarrow \\ x \geq \frac{y - 2,5}{2}$$

så X er ligefordelt på $\left(\frac{y-2,5}{2}; 1\right)$.

Dermed er den tæthed for $X|Y = y$

$$f_{X|Y=y}(x) = \begin{cases} 2 \cdot \frac{1_{(0, \frac{y-0,5}{2})}(x)}{y-0,5} \\ 2 \cdot \frac{1_{(\frac{y-2,5}{2}, 1)}(x)}{2-(y-2,5)} \end{cases}$$

Variansen af X givet Y er dermed

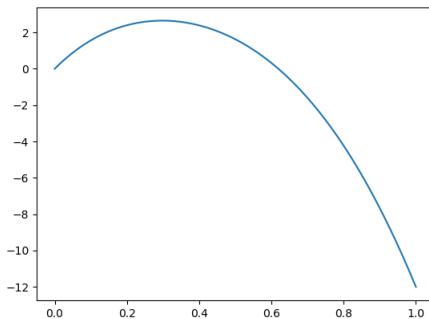
$$0 \leq X \leq \frac{Y-0,5}{2} : \text{Var}(X|Y = y) = E(X^2|Y = y) - E(X|Y = y)^2 \\ = \int_{-\infty}^{\infty} x^2 f_{X|Y=y}(x) dx - \left(\int_{-\infty}^{\infty} x f_{X|Y}(x) dx \right)^2 \\ = \int_0^{\frac{y-0,5}{2}} x^2 \frac{2}{y-0,5} dx - \left(\int_0^{\frac{y-0,5}{2}} x \frac{2}{y-0,5} dx \right)^2 \\ = \frac{2}{y-0,5} \left[\frac{1}{3} x^3 \right]_0^{\frac{y-0,5}{2}} - \left(\frac{2}{y-0,5} \left[\frac{1}{2} x^2 \right]_0^{\frac{y-0,5}{2}} \right)^2 \\ = \frac{1}{12} \frac{(y-0,5)^3}{y-0,5} - \left(\frac{1}{4} \frac{(y-0,5)^2}{y-0,5} \right)^2 \\ = \frac{1}{12} (y-0,5)^2 - \frac{1}{16} (y-0,5)^2 \\ = \frac{1}{48} (y-0,5)^2 \quad (\text{stregt voksende i } y)$$

$$\begin{aligned}
 \frac{Y - 2,5}{2} \leq X \leq 1 : \text{Var}(X|Y = y) &= \int_{\frac{y-2,5}{2}}^1 x^2 \frac{2}{2 - (y - 2,5)} dx - \left(\int_{\frac{y-2,5}{2}}^1 x \frac{2}{2 - (y - 2,5)} dx \right)^2 \\
 &= \frac{2}{2 - (y - 2,5)} \left[\frac{1}{3} x^3 \right]_{\frac{y-2,5}{2}}^1 - \left(\frac{2}{2 - (y - 2,5)} \left[\frac{1}{2} x^2 \right]_{\frac{y-2,5}{2}}^1 \right)^2 \\
 &= \frac{2}{3} \left(\frac{1 - \left(\frac{y-2,5}{2} \right)^3}{2 - (y - 2,5)} \right) - \left(\frac{1 - \left(\frac{y-2,5}{2} \right)^2}{2 - (y - 2,5)} \right)^2 \\
 &= \frac{2}{3} \left(\frac{1 - z^3}{2 - 2z} \right) - \left(\frac{1 - z^2}{2 - 2z} \right)^2, \quad \left(z = \frac{y - 2,5}{2} \in [0, 1] \right) \\
 &= \frac{2}{3} \left(\frac{1 - z^3}{2 - 2z} \right) - \left(\frac{1 + z^4 - 2z^2}{(2 - 2z)^2} \right) \\
 &= \frac{2}{3} \left(\frac{(2 - 2z)(1 - z^3)}{(2 - 2z)^2} \right) - \left(\frac{1 + z^4 - 2z^2}{(2 - 2z)^2} \right) \\
 &= \frac{2}{3} \left(\frac{2 - 2z^3 - 2z + 2z^4 - 1 - z^4 + 2z^2}{(1 - 2z)^2} \right) \\
 &= \frac{2}{3} \left(\frac{z^4 - 2z^3 + 2z^2 - 2z + 1}{(2 - 2z)^2} \right)
 \end{aligned}$$

$$\begin{aligned}
 \frac{\delta Var(X|Y=y)}{\delta z} &= \frac{2}{3} \left(\frac{(4z^3 - 6z^2 + 4z - 2)(2 - 2z)^2 + (z^4 - 2z^3 + 2z^2 - 2z + 1)(4 - 4z)}{(2 - 2z)^4} \right) \\
 &= \frac{2}{3} \left(\frac{(4z^3 - 6z^2 + 4z - 2)(4 + 4z^2 - 8z)}{(2 - 2z)^4} \right) \\
 &+ \frac{2}{3} \left(\frac{4z^4 - 4z^5 - 8z^3 + 8z^4 + 8z^2 - 8z^3 - 8z + 8z^2 + 4 - 4z}{(2 - 2z)^4} \right) \\
 &= \frac{2}{3} \left(\frac{16z^3 + 16z^5 - 32z^4 - 24z^2 - 24z^4 + 48z^3 + 16z + 16z^3 - 32z^2}{(2 - 2z)^4} \right) \\
 &+ \frac{2}{3} \left(\frac{-8 - 8z^2 + 16z - 4z^5 + 12z^4 - 16z^3 + 16z^2 - 12z + 4}{(2 - 2z)^4} \right) \\
 &= \frac{2}{3} \left(\frac{12z^5 - 44z^4 + 48z^3 - 48z^2 + 20z - 4}{(2 - 2z)^4} \right)
 \end{aligned}$$

Hvis vi kan vise, at polynomiet $g(z) = 12z^5 - 44z^4 + 48z^3 - 48z^2 + 20z - 4 < 4$ for $z \in [0, 1]$ så er det vist at $Var(X|Y=y)$ er strengt aftagende for $y \in [2, 5; 4, 5]$.

Plot af $g(z)$:



Plottet af $g(z)$ viser, at $g(z) < 4$ for $z \in (0, 1]$. Dermed er $\frac{\delta Var(X|Y=y)}{\delta z} < 0$ for $y \in [2, 5; 4, 5]$ og variansen er størst ved $y = 2, 5$. Da $Var(X|Y=y)$ er strengt voksende for $y \in [0, 5; 2, 5]$ kan vi konkludere at variationen i ratingen er størst ved prisen 2.500\$.

Opgave U44.2.1

Lad U og V være uafhængige standard normalfordelte variable.

Definér følgende variable:

$$\begin{aligned}
 X &= \alpha_1 + \beta_1 U \\
 Y &= \alpha_2 + \beta_2 U + \delta_2 V
 \end{aligned}$$

- Udregn $P(0, 1 < U < 0, 5)$.

Tæthedsfunktionen for en standard normalfordelt parameter, U , er

$$f_U(u) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{u^2}{2} \right\}$$

Og $P(0, 1 < U < 0, 5)$ er dermed

$$P(0, 1 < U < 0, 5) = \int_{0,1}^{0,5} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{u^2}{2} \right\} du$$

hvilket vi ikke kan løse analytisk.

Funktionen `normal()` i STATA returnerer den akkumulerede sandsynlighed i standard normalfordelingen, $\Phi()$. Vi kan derfor få sandsynligheden ved følgende kode:

$$\begin{aligned} P(0, 1 < U < 0, 5) &= \Phi(0, 5) - \Phi(0, 1) \\ &= \text{normal}(0.5) - \text{normal}(0.1) \\ &\approx \underline{\underline{0, 1516}} \end{aligned}$$

2. Udregn $E(X)$, $E(Y)$, $V(X)$ og $V(Y)$.

Husk, at $E(U) = E(V) = 0$, $V(U) = V(V) = 1$ og V og U er uafhængige.

$$\begin{aligned} E(X) &= E(\alpha_1 + \beta_1 U) = \alpha_1 + \beta_1 E(U) = \underline{\underline{\alpha_1}} \\ E(Y) &= E(\alpha_2 + \beta_2 U + \delta_2 V) = \alpha_2 + \beta_2 E(U) + \delta_2 E(V) = \underline{\underline{\alpha_2}} \\ V(X) &= V(\alpha_1 + \beta_1 U) = \beta_1^2 V(U) = \underline{\underline{\beta_1^2}} \\ V(Y) &= V(\alpha_2 + \beta_2 U + \delta_2 V) = \beta_2^2 V(U) + \delta_2^2 V(V) = \underline{\underline{\beta_2^2 + \delta_2^2}} \end{aligned}$$

3. Udregn $E(X \cdot Y)$ og $Cov(X, Y)$.

$$\begin{aligned} E(X \cdot Y) &= E((\alpha_1 + \beta_1 U) \cdot (\alpha_2 + \beta_2 U + \delta_2 V)) \\ &= E(\alpha_1 \alpha_2 + \alpha_1 \beta_2 U + \alpha_1 \delta_2 V + \alpha_2 \beta_1 U + \beta_1 \beta_2 U^2 + \beta_1 \delta_2 U V) \\ &= \alpha_1 \alpha_2 + \alpha_1 \beta_2 E(U) + \alpha_1 \delta_2 E(V) + \alpha_2 \beta_1 E(U) + \beta_1 \beta_2 E(U^2) + \beta_1 \delta_2 E(U) E(V) \\ &= \alpha_1 \alpha_2 + \beta_1 \beta_2 V(U) \\ &= \underline{\underline{\alpha_1 \alpha_2 + \beta_1 \beta_2}} \\ Cov(X, Y) &= E(X \cdot Y) - E(X) \cdot E(Y) \\ &= \alpha_1 \alpha_2 + \beta_1 \beta_2 - \alpha_1 \alpha_2 \\ &= \underline{\underline{\beta_1 \beta_2}} \end{aligned}$$

4. Hvad skal α_1 og β_1 sættes til for at $E(X) = 10$ og $V(X) = 4$?

Ovenfor har vi udledt $E(X) = \alpha_1$ og $V(X) = \beta_1^2$.

$$\begin{aligned} E(X) = 10 &\Leftrightarrow \\ \alpha_1 &= \underline{\underline{10}} \\ V(X) = 4 &\Leftrightarrow \\ \beta_1 &= \underline{\underline{2}} \vee \beta_1 = \underline{\underline{-2}} \end{aligned}$$

Opgave U44.2.2

I analyser af indkomstmobilitet over tid er man blandt andet interesseret i at undersøge, om det er sådan, at hvis man har lav timeløn eller indkomst i en periode, er der stor sandsynlighed for, at det også er tilfældet i den efterfølgende periode (lav mobilitet). I denne type af analyse anvender man ofte en bestemt model til at beskrive sammenhængen mellem indkomst eller timeløn i 2 perioder. Dette vil vi også gøre i det følgende.

Vi definerer stokastiske variable Y_1 og Y_2 på følgende måde:

$$\begin{aligned} Y_1 &: \text{Timeløn periode 1 (målt i periode 1 kroner)} \\ Y_2 &: \text{Timeløn periode 2 (målt i periode 1 kroner)} \end{aligned}$$

Timelønnen i periode 1 er normalfordelt med middelværdi μ og varians σ^2 , dvs.

$$Y_1 \sim N(\mu, \sigma^2)$$

Timelønnen i periode 2 er givet ved:

$$Y_2 = \alpha + \beta Y_1 + U$$

hvor Y_1 og U er uafhængige og $U \sim N(0, v^2)$.

1. Lad $\mu = 350$ og $\sigma^2 = 12365$.

- (a) Udregn sandsynligheden for at timelønnen i periode 1 er højest 275.

Benyt, at $X \sim N(\mu, \sigma^2) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

$$\begin{aligned} P(Y_1 \leq 275) &= \Phi\left(\frac{275 - 350}{\sqrt{12365}}\right) \\ &= \text{normal}\left(\frac{275 - 350}{\sqrt{12365}}\right) \quad (\text{i STATA}) \\ &\approx \underline{\underline{0,2492}} \end{aligned}$$

- (b) Udregn sandsynligheden for at timelønnen i periode 1 er mindst 425.

$$\begin{aligned} P(Y_1 \geq 425) &= 1 - \Phi\left(\frac{425 - 350}{\sqrt{12365}}\right) \\ &= P(Y_1 \leq 275) \\ &\approx \underline{\underline{0,2492}} \end{aligned}$$

Da Y_1 er symmetrisk om sin middelværdi $\mu = 350$.

2. Udregn middelværdi og varians af Y_2 og angiv fordelingen af Y_2 .

$$\begin{aligned} E(Y_2) &= E(\alpha + \beta Y_1 + U) \\ &= \alpha + \beta E(Y_1) + E(U) \\ &= \underline{\underline{\alpha + \beta \mu}} \\ V(Y_2) &= V(\alpha + \beta Y_1 + U) \\ &= \beta^2 V(Y_1) + V(U) \quad (\text{Husk, at } Y_1 \text{ og } U \text{ er uafhængige}) \\ &= \underline{\underline{\beta^2 \sigma^2 + v^2}} \end{aligned}$$

Da summen af to normalfordelte variable også er normalfordelt er fordelingenq af Y_2 :

$$Y_2 \sim N(\alpha + \beta \mu, \beta^2 \sigma^2 + v^2)$$

3. Er Y_1 og Y_2 uafhængige?

Da Y_1 og Y_2 begge er normalfordelte (og kun fordi de er normalfordelte) gælder, at hvis $Cov(Y_1, Y_2) = 0$ er Y_1 og Y_2 uafhængige.

I det følgende benytter vi, at $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$ ((3.8.4) i bogen).

$$\begin{aligned} Cov(Y_1, Y_2) &= Cov(Y_1, \alpha + \beta Y_1 + U) \\ &= \beta Cov(Y_1, Y_1) + Cov(Y_1, U) \\ &= \beta V(Y_1) \\ &= \beta \sigma^2 \end{aligned}$$

Dermed er Y_1 og Y_2 uafhængige hvis $\beta = 0$ eller $\sigma^2 = 0$.

4. Angiv den betingede middelværdi og varians af Y_2 givet $Y_1 = y_1$.

$$\begin{aligned} E(Y_2|Y_1 = y_1) &= E[(\alpha + \beta Y_1 + U)|Y_1 = y_1] \\ &= \alpha + \beta E(Y_1|Y_1 = y_1) + E(U|Y_1 = y_1) \\ &= \underline{\underline{\alpha + \beta y_1}} \\ V(Y_2|Y_1 = y_1) &= V[(\alpha + \beta Y_1 + U)|Y_1 = y_1] \\ &= \beta^2 V(Y_1|Y_1 = y_1) + V(U|Y_1 = y_1) \\ &= V(U) \\ &= \underline{\underline{v^2}} \end{aligned}$$

5. Antag at $\beta \neq 0$. Lad $\mu = 350$ og $\sigma^2 = 12.365$. Lad også $\alpha = 350(1 - \beta)$ og $v^2 = 12.365(1 - \beta^2)$.

(a) Angiv de marginale fordelinger af Y_1 og Y_2 .

$$\begin{aligned} E(Y_1) &= 350 \\ V(Y_1) &= 12365 \\ Y_1 &\sim \underline{\underline{N(350, 12.365)}} \\ E(Y_2) &= \alpha + \beta \mu = 350(1 - \beta) + \beta \cdot 350 = 350 \\ V(Y_2) &= \beta^2 \sigma^2 + v^2 = \beta^2 \cdot 12.365 + (1 - \beta^2) \cdot 12.365 = 12.365 \\ Y_2 &\sim \underline{\underline{N(350, 12.365)}} \end{aligned}$$

(b) Vis at for $\beta > 0$ gælder $E(Y_2|Y_1 = 275) < E(Y_2)$

Fra tidligere udregninger ved vi, at

$$\begin{aligned} E(Y_2) &= \alpha + \beta \mu \\ E(Y_2|Y_1 = y_1) &= \alpha + \beta y_1 \end{aligned}$$

og dermed

$$E(Y_2|Y_1 = 275) = \alpha + \beta \cdot 275 < E(Y_2) = \alpha + \beta \cdot 350$$

(c) Lad nu $\beta = 0,90$. Udregn $P(Y_2 \leq 275|Y_1 \leq 275)$ og $P(Y_2 \geq 425|Y_1 \leq 275)$.

Du kan bruge følgende resultater, (som gælder når $\beta = 0,90$):

$$\begin{aligned} P(Y_1 \leq 275, Y_2 \leq 275) &= 0,193 \\ P(Y_1 \leq 275, Y_2 \leq 425) &= 0,250 \end{aligned}$$

$$\begin{aligned}
 P(Y_2 \leq 275 | Y_1 \leq 275) &= \frac{P(Y_1 \leq 275, Y_2 \leq 275)}{P(Y_1 \leq 275)} \\
 &= \frac{0,193}{\Phi\left(\frac{275-350}{\sqrt{12.365}}\right)} \\
 &= \frac{0,193}{\text{normal}\left(\frac{275-350}{\sqrt{12.365}}\right)} \\
 &\approx 0.7720
 \end{aligned}$$

For at beregne $P(Y_2 \geq 425 | Y_1 \leq 275)$ skal vi kende $P(Y_1 \leq 275, Y_2 \geq 425)$. Da vi har givet $P(Y_1 \leq 275, Y_2 \leq 425) = 0,250$ kan vi udnytte, at

$$\begin{aligned}
 P(Y_1 \leq 275, Y_2 \leq 425) + P(Y_1 \leq 275, Y_2 \geq 425) &= P(Y_1 \leq 275) \Leftrightarrow \\
 \Phi\left(\frac{275-350}{\sqrt{12.365}}\right) - 0,250 &= P(Y_1 \leq 275, Y_2 \geq 425) \Leftrightarrow \\
 P(Y_2 \geq 425 | Y_1 \leq 275) &= \frac{\text{normal}\left(\frac{275-350}{\sqrt{12.365}}\right) - 0,250}{\text{normal}\left(\frac{275-350}{\sqrt{12.365}}\right)} \\
 &\approx 0
 \end{aligned}$$

(d) *Giv en kort fortolkning af resultaterne i spørgsmål 5(b)-5(c).*

I 5(b) ser vi, at den forventede værdi af Y_2 ændres, når vi har viden om udfaldet af Y_1 .

I 5(c) ser vi, at sandsynlighederne for $Y_2 \leq 275$ og $Y_2 \geq 425$ som ubetinget er de samme ændres markant når der betinges på et givet udfald af Y_1 .

Opgave U44.2.3

Lad (X, Y) være en kontinuert to-dimensional stokastisk vektor med sandsynlighedstæthed

$$f_{X,Y}(x, y) = 6 \exp(-2x - 3y), \quad x \in [0, \infty[\text{ og } y \in [0, \infty[$$

Definér nu mængden

$$A = \{(x, y) : 0 \leq x + y < 1, \quad x > 0, y > 0\}$$

1. Argumentér for hvorfor X og Y er uafhængige.

De marginale tætheder for X og Y er

$$\begin{aligned}
 f_X(x) &= 1_{[0, \infty[}(x) \int_0^\infty 6 \exp(-2x - 3y) dy \\
 &= 1_{[0, \infty[}(x) \exp(-2x) \lim_{n \rightarrow \infty} \left[-2 \exp(-3y) \right]_0^n \\
 &= 1_{[0, \infty[}(x) 2 \exp(-2x) \\
 f_Y(y) &= 1_{[0, \infty[}(y) \int_0^\infty 6 \exp(-2x - 3y) dx \\
 &= 1_{[0, \infty[}(y) \exp(-3y) \lim_{n \rightarrow \infty} \left[-3 \exp(-2x) \right]_0^n \\
 &= 1_{[0, \infty[}(y) 3 \exp(-3y)
 \end{aligned}$$

Da $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ er X og Y uafhængige.

2. Udregn $P((X, Y) \in A)$.

Hvis $(X, Y) \in A$ er X opad begrænset af $1 - Y$.

Dermed

$$\begin{aligned}
 P((X, Y) \in A) &= \int_0^1 \int_0^{1-y} 6 \exp(-2x - 3y) dx dy \\
 &= \int_0^1 \left[-3 \exp(-2x - 3y) \right]_{x=0}^{x=1-y} dy \\
 &= \int_0^1 (3 \exp(-3y) - 3 \exp(-2(1-y) - 3y)) dy \\
 &= \int_0^1 (3 \exp(-3y) - 3 \exp(-y - 2)) dy \\
 &= \left[3 \exp(-y - 2) - \exp(-3y) \right]_0^1 \\
 &= 3 \exp(-3) - \exp(-3) - 3 \exp(-2) + 1 \\
 &= 1 + 2 \exp(-3) - 3 \exp(-2) \\
 &\approx \underline{\underline{0,6935}}
 \end{aligned}$$

3. Find tætheden for (X, Y) givet $(X, Y) \in A$.

$$\text{Definition 11: } f_{X,Y|A}(x, y) = \frac{f_{X,Y}(x,y)}{P((X,Y) \in A)}.$$

Dermed

$$f_{X,Y|A}(x, y) = \frac{6 \exp(-2x - 3y)}{1 + 2 \exp(-3) - 3 \exp(-2)} \approx \frac{6 \exp(-2x - 3y)}{0,6935}$$

4. Er X og Y givet A uafhængige?

De marginale fordelinger af X og Y givet A er

$$\begin{aligned}
 f_{X|A}(x) &= \int_0^{1-x} \frac{6 \exp(-2x - 3y)}{0,6935} dy \\
 &= \frac{1}{0,6935} \left[-2 \exp(-2x - 3y) \right]_{y=0}^{y=1-x} \\
 &= \frac{1}{0,6935} (2 \exp(-2x) - 2 \exp(-2x - 3(1-x))) \\
 &= \frac{1}{0,6935} (2 \exp(-2x) - 2 \exp(x-3)) \\
 f_{Y|A}(y) &= \int_0^{1-y} \frac{6 \exp(-2x - 3y)}{0,6935} dx \\
 &= \frac{1}{0,6935} \left[-3 \exp(-2x - 3y) \right]_{x=0}^{x=1-y} \\
 &= \frac{1}{0,6935} (3 \exp(-3y) - 3 \exp(-2(1-y) - 3y)) \\
 &= \frac{1}{0,6935} (3 \exp(-3y) - 3 \exp(-y - 2))
 \end{aligned}$$

Da $f_{X,Y|A}(x, y) \neq f_{X|A}(x) \cdot f_{Y|A}(y)$ er X og Y givet A ikke uafhængige.

5. Find $E(X|(X, Y) \in A)$, $E(Y|(X, Y) \in A)$ og $E(XY|(X, Y) \in A)$.

$$\begin{aligned}
E(X|(X, Y) \in A) &= \int_0^1 xf_{X|A}(x)dx \\
&= \int_0^1 \frac{2x}{0,6935} (\exp(-2x) - \exp(x-3)) \\
&= \left[\frac{2x}{0,6935} \left(-\frac{1}{2} \exp(-2x) - \exp(x-3) \right) \right]_0^1 - \frac{2}{0,6935} \int_0^1 \left(-\frac{1}{2} \exp(-2x) - \exp(x-3) \right) \\
&= \frac{2}{0,6935} \left(-\frac{1}{2} \exp(-2) - \exp(-2) \right) - \frac{2}{0,6935} \left[\frac{1}{4} \exp(-2x) - \exp(x-3) \right]_0^1 \\
&= \frac{2}{0,6935} \left(-\frac{3}{2} \exp(-2) \right) - \frac{2}{0,6935} \left(\frac{1}{4} \exp(-2) - \exp(-2) - \frac{1}{4} + \exp(-3) \right) \\
&= \frac{2}{0,6935} \left(\frac{1}{4} + \left(-\frac{6}{4} + \frac{3}{4} \right) \exp(-2) - \exp(-3) \right) \\
&= \underline{\underline{\frac{2}{0,6935} \left(\frac{1}{4} - \frac{3}{4} \exp(-2) - \exp(-3) \right) \approx 0,2861}}
\end{aligned}$$

$$E(Y|(X, Y) \in A) = \int_0^1 yf_{Y|A}(y)dy$$

$$E(XY|(X, Y) \in A) = \int_0^1 \int_0^1 xyf_{X,Y|A}(x, y) dx dy$$

Centrale begreber

- Ligefordelingen på $B \in \mathbb{R}^n$ $\left(p(x_1, \dots, x_n) = \frac{1_B(x_1, \dots, x_n)}{|B|} \right)$ (6.1.2)
- Lad (X, Y) være en to-dimensional kontinuert stokastisk vektor med sandsynlighedstæthed $p(x, y)$.
Den marginale sandsynlighed q for X er:

$$q(x) = \int_{\mathbb{R}} p(x, y) dy \quad (6.1.5)$$

- Normalfordelingen $\left(p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \right)$
- Fordelingsfunktion for en kontinuert fordeling:

$$F(x) = \int_{-\infty}^x p(y) dy$$

- Fordeling af $Y = t(X)$:

$$q(y) = \begin{cases} p(t^{-1}(y)) \left| \frac{d}{dy} t^{-1}(y) \right| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.1)$$

eller

$$q(y) = \begin{cases} p(t^{-1}(y)) / |t'(t^{-1}(y))| & \text{hvis } y \in (v, h) \\ 0 & \text{hvis } y \notin (v, h) \end{cases} \quad (5.4.2)$$

Opgave U45.1

Denne opgave skal give jer en forståelse for OLS regression. Lad X være $N(\mu, \sigma^2)$ fordelt.

- Vis at $Y = \frac{1}{\sqrt{\sigma^2}}(X - \mu)$ er $N(0, 1)$ fordelt.

Y er en transformation af X :

$$\begin{aligned} Y = t(X) &= \frac{1}{\sqrt{\sigma^2}}(X - \mu) \Leftrightarrow \\ X = t^{-1}(Y) &= \sqrt{\sigma^2}Y + \mu \Leftrightarrow \\ \frac{d}{dy}t^{-1}(y) &= \sqrt{\sigma^2} \end{aligned}$$

Tæthedsfunktionen for X er

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}.$$

Tæthedsfunktionen for $Y = t(X)$ er

$$\begin{aligned} f_Y(y) &= f_X(t^{-1}(y)) \left| \frac{d}{dy}t^{-1}(y) \right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\sqrt{\sigma^2}y + \mu - \mu)^2}{2\sigma^2}\right\} |\sqrt{\sigma^2}| \\ &= \frac{\sqrt{\sigma^2}}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left\{-\frac{\sigma^2 y^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \Leftrightarrow \\ Y &\sim N(0, 1) \end{aligned}$$

- Lad nu (Y, X) være to-dimensional normalfordelt $N(\mu, \Omega)$ med

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} \quad \Omega = \begin{pmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{pmatrix}$$

Vis at $Z = \frac{1}{\sqrt{\sigma_X^2}}(X - \mu_X)$ er $N(0, 1)$ fordelt.

Ifølge Egenskab G.1 i Rahbek: "Lidt regneregler" er de marginale fordelinger for Y og X

$$\begin{aligned} Y &\sim N(\mu_Y, \sigma_Y^2) \\ X &\sim N(\mu_X, \sigma_X^2). \end{aligned}$$

Dermed følger af resultatet i forrige opgave, at

$$Z = \frac{1}{\sqrt{\sigma_X^2}}(X - \mu_X) \sim N(0, 1)$$

- Lad (Y, X) være som i spørgsmål 2, men med $\mu = (0, 0)'$. Vis at

$$Y - \beta X$$

med $\beta = \sigma_{XY}/\sigma_X^2$ er $N(0, \sigma^2)$ fordelt med

$$\sigma^2 = \sigma_Y^2 - \beta\sigma_{XY}$$

Ifølge *Sørensen: Sætning 6.3.12* er summen af n normalfordelinger også normalfordelt. Dermed ved vi fra vores regneregler for middelværdi og varians, at

$$\begin{aligned} E(Y - \beta X) &= E(Y) - \beta E(X) = 0 - \beta 0 = 0 \\ Var(Y - \beta X) &= Var(Y) + \beta^2 Var(X) - 2\beta Cov(Y, X) \\ &= \sigma_Y^2 + \left(\frac{\sigma_{XY}}{\sigma_X^2} \right)^2 \sigma_X^2 - 2 \frac{\sigma_{XY}}{\sigma_X^2} \sigma_{XY} \\ &= \sigma_Y^2 + \frac{\sigma_{XY}}{\sigma_X^2} \sigma_{XY} - 2 \frac{\sigma_{XY}}{\sigma_X^2} \sigma_{XY} \\ &= \sigma_Y^2 - \frac{\sigma_{XY}}{\sigma_X^2} \sigma_{XY} \\ &= \sigma_Y^2 - \beta \sigma_{XY} \end{aligned}$$

4. Lad (Y, X) være som i spørgsmål 3. Vis at

$$E((Y - \beta X)X) = 0$$

Dette betyder at $Y - \beta X$ og X er uafhængige. Forklar hvorfor.

Vi ved at $\mu = (\mu_Y, \mu_X)' = (0, 0)'$:

$$\begin{aligned} E((Y - \beta X)X) &= E(YX) - \beta E(X^2) \\ &= \left(E(YX) - \underbrace{E(Y)E(X)}_{= 0} \right) - \beta \left(E(X^2) - \underbrace{E(X)^2}_{= 0} \right) \\ &= \sigma_{XY} - \beta \sigma_X^2 \\ &= \sigma_{XY} - \frac{\sigma_{XY}}{\sigma_X^2} \sigma_X^2 \\ &= \sigma_{XY} - \sigma_{XY} \\ &= 0 \end{aligned}$$

Ifølge *Egenskab G.2 i Rahbek: "Lidt regneregler"* er to normalfordelte variable, X og Y , uafhængige hvis og kun hvis $\sigma_{XY} = 0$.

Da vi ved, at $E(X) = 0$ er

$$E((Y - \beta X)X) = Cov(Y - \beta X, X)$$

og da denne konvarians er lig nul er $Y - \beta X$ og X altså uafhængige.

Opgave U45.2

Denne opgave skal træne jer i at bruge formlerne for den bi-variate normalfordeling. Lad (X, Y) være normalfordelt $N(m, \Omega)$ med

$$m = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

1. Hvad er $E(X)$ og $V(X)$?

Fra ovenstående ved vi at de marginale fordelinger for X og Y er

$$X \sim N(1, 1)$$

$$Y \sim N(0, 1)$$

$$\sigma_{XY} = \rho$$

Dermed

$$E(X) = \underline{\underline{1}}$$

$$V(X) = \underline{\underline{1}}$$

2. Hvordan er Y fordelt?

$$Y \sim \underline{\underline{N(0, 1)}}$$

3. Hvad er $Cov(Y, X)$?

$$Cov(Y, X) = \sigma_{XY} = \underline{\underline{\rho}}$$

4. Hvad er $E(Y|X = x)$?

Ifølge Egenskab G.3 i Rahbek: "Lidt regneregler" er den betingede fordeling af Y givet $X = x$ $N(\mu_{Y|X}, \sigma_{Y|X}^2)$ fordelt med

$$\mu_{Y|X} = E(Y|X = x) = \mu_Y + \omega(x - \mu_X), \quad \omega = \frac{\sigma_{XY}}{\sigma_X^2}$$

$$\sigma_{Y|X}^2 = V(Y|X = x) = \sigma_Y^2 - \omega\sigma_{XY} = \sigma_Y^2 - \frac{(\sigma_{XY})^2}{\sigma_X^2}.$$

Da vi har $\mu_Y = 0$, $\mu_X = 1$, $\sigma_X^2 = 1$ og $\sigma_{XY} = \rho$ er

$$E(Y|X = x) = 0 + \frac{\rho}{1}(x - 1) = \underline{\underline{\rho x - \rho}}$$

5. Hvad er $E(X|Y = y)$?

Da vi har $\mu_x = 1$, $\mu_Y = 0$, $\sigma_Y^2 = 1$ og $\sigma_{XY} = \rho$ er

$$E(X|Y = y) = \mu_{X|Y} = \mu_X + \omega(y - \mu_Y), \quad \omega = \frac{\sigma_{XY}}{\sigma_Y^2}$$

$$= 1 + \frac{\rho}{1}(y - 0)$$

$$= \underline{\underline{1 + \rho y}}$$

6. Hvad er $V(Y|X)$?

Da vi har $\sigma_Y^2 = \sigma_X^2 = 1$ og $\sigma_{XY} = \rho$ er

$$\sigma_{Y|X}^2 = V(Y|X) = \sigma_Y^2 - \omega\sigma_{XY}$$

$$= 1 - \frac{\rho}{1}\rho$$

$$= \underline{\underline{1 - \rho^2}}$$

7. Hvad gælder der for (X, Y) hvis $\rho = 0, 9$? og hvis $\rho = 0$?

På udtrykket $\sigma_{Y|X}^2 = 1 - \rho^2$ ses det at variansen af Y givet X falder jo mere korreleerde X og Y er (da $\sigma_X^2 = \sigma_Y^2 = 1$ er $Corr(X, Y) = Cov(X, Y) = \rho$). Dette er logisk da X forklarer mere af variationen i Y hvis korrelationen er højere. Dermed falder variationen i Y markant hvis X er givet, og X og Y er stærkt korreleerde.

Hvis $\rho = 0$ og X og Y dermed er uafhængige ændres variansen af Y ikke af at X er givet.

Opgave U45.Ekstra

Opskriv tætheden $p(y, x)$ for (Y, X) som er normalfordelte $N(\mu, \Omega)$ med

$$\mu = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \quad \Omega = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$$

Fra Rahbek: "Lidt regneregler" ved vi, at hvis (X, Y) er normalfordelt med middelværdi

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$$

og varians

$$\Omega = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix},$$

hvor $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$, $\sigma_{XY} = \sigma_{YX}$ og $\det(\Omega) > 0$, da er tætheden for (X, Y) givet ved

$$f_{X,Y}(x, y) = \left(\frac{1}{\sqrt{2\pi}} \right)^2 \frac{1}{\sqrt{\det(\Omega)}} \exp \left\{ -\frac{1}{2} (x - \mu_X, y - \mu_Y) \Omega^{-1} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \right\}.$$

For den givne stokastiske vektor (Y, X) har vi

$$\begin{aligned} \mu_Y &= 0 \\ \mu_X &= 2 \\ \det(\Omega) &= 1^2 - 0,5^2 = 0,75 \\ \Omega^{-1} &= \frac{1}{\det(\Omega)} \begin{pmatrix} 1 & -0,5 \\ -0,5 & 1 \end{pmatrix} = \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}. \end{aligned}$$

Dermed

$$\begin{aligned} f_{Y,X}(y, x) &= \left(\frac{1}{\sqrt{2\pi}} \right)^2 \frac{1}{\sqrt{\frac{3}{4}}} \exp \left\{ -\frac{1}{2} (y - 0, x - 2) \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} y - 0 \\ x - 2 \end{pmatrix} \right\} \\ &= \frac{2}{2\sqrt{3}\pi} \exp \left\{ -\frac{1}{2} (y, x - 2) \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} y \\ x - 2 \end{pmatrix} \right\} \end{aligned}$$

Opgave U45.3

I en empirisk undersøgelse af afkast på en amerikansk aktie Y (f.eks. Microsoft) og aktieindekset X (f.eks S&P500) blev følgende model opstillet:

$$Y = \beta X + \varepsilon,$$

med $\varepsilon \sim N(0, \sigma^2)$.

- Parameteren β kaldes slet og ret "beta" i finansiering. Hvis (Y, X) er simultant normalfordelte, angiv en fortolkning af β .

Den forventede værdi af Y er

$$\begin{aligned} E(Y) &= E(\beta X + \varepsilon) \\ &= \beta E(X) \end{aligned}$$

Den forventede værdi af Y givet $X = x$ er

$$\begin{aligned}
 E(Y|X = x) &= E(\beta X + \varepsilon|X = x) \\
 &= \beta E(X|X = x) + E(\varepsilon|X = x) \\
 &= \beta x \\
 &= \mu_Y + \omega(x - \mu_X) \quad (\text{Egenskab G.3 i Rahbek: "Lidt regneregler"}) \\
 &= \beta\mu_X + \omega x - \omega\mu_X \Leftrightarrow \\
 \beta x - \beta\mu_X &= \omega x - \omega\mu_X \Leftrightarrow \\
 \beta &= \omega.
 \end{aligned}$$

Vi har altså at $\beta = \omega = \frac{\sigma_{XY}}{\sigma_X^2}$. β er altså et mål for kovariansen mellem X og Y normeret med variansen af X . En større kovarians betyder større marginal effekt af X på Y . Større variation i X betyder mindre marginal effekt af X på Y .

2. I en anden undersøgelse blev der opstillet en model for Y ,

$$Y = \varepsilon_Y$$

med $\varepsilon_Y \sim N(0, \sigma_Y^2)$, og en model for X ,

$$X = \varepsilon_X$$

med $\varepsilon_X \sim N(0, \sigma_X^2)$. Disse modeller for hhv. X og Y hænger fint sammen med $Y = \beta X + \varepsilon$. Forklar hvorfor.

Hvis vi igen benytter Egenskab G.3 i Rahbek: "Lidt regneregler", har vi

$$\begin{aligned}
 \mu_{Y|X} &= \mu_Y + \omega(x - \mu_X) \\
 &= 0 + \omega x - 0 \\
 &= \omega x.
 \end{aligned}$$

Sætter vi igen $\beta = \omega = \frac{\sigma_{XY}}{\sigma_X^2}$ og $Y = \beta X + \varepsilon$ har vi

$$E(Y|X = x) = \beta x$$

og dermed ser vi sammenhængen mellem de to modeller.

Opgave U45.4

Denne opgave skal træne jeres brug af property G.3 i noten. Antag at (Y, X) er to-dimensionalt $N(\mu, \Omega)$ fordelte. Det oplyses at

$$E(Y|X) = X \text{ og } V(Y|X) = 1$$

samt at $E(X) = 0$ og $V(X) = 1$. Find μ og Ω .

Ifølge Egenskab G.3 i Rahbek: "Lidt regneregler" er den betingede fordeling af Y givet $X = x$ $N(\mu_{Y|X}, \sigma_{Y|X}^2)$ fordelt med

$$\begin{aligned}\mu_{Y|X} &= E(Y|X = x) = \mu_Y + \omega(x - \mu_X), \quad \omega = \frac{\sigma_{XY}}{\sigma_X^2} \\ \sigma_{Y|X}^2 &= V(Y|X = x) = \sigma_Y^2 - \omega\sigma_{XY} = \sigma_Y^2 - \frac{(\sigma_{XY})^2}{\sigma_X^2}.\end{aligned}$$

Vi vil finde

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} \text{ og } \Omega = \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{pmatrix}$$

og vi ved, at $\mu_{Y|X} = X$, $\sigma_{Y|X}^2 = 1$, $\mu_X = 0$ og $\sigma_X^2 = 1$.

To regneregler skal bruges til udledningen.

- BM.2 i Rahbek: "Lidt regneregler": $E(X|X = x) = x$ og $E(X|X) = X$.
- Law of Iterated Expectations: $E(E(Y|X)) = Y$.

$$\begin{aligned}E(Y|X) &= X \Leftrightarrow \\ E(E(Y|X)) &= E(X) \Leftrightarrow \\ E(Y) &= 0\end{aligned}$$

og

$$\begin{aligned}\mu_{Y|X} &= \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X) \Leftrightarrow \\ X &= 0 + \frac{\sigma_{XY}}{1}(X - 0) \Leftrightarrow \\ X &= \sigma_{XY}X \Leftrightarrow \\ \sigma_{XY} &= 1\end{aligned}$$

og

$$\begin{aligned}\sigma_{Y|X}^2 &= \sigma_Y^2 - \frac{(\sigma_{XY})^2}{\sigma_X^2} \Leftrightarrow \\ 1 &= \sigma_Y^2 - \frac{1^2}{1} \Leftrightarrow \\ \sigma_Y^2 &= 2.\end{aligned}$$

Endelig har vi

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N(\mu, \Omega), \quad \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ og } \Omega = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

Bevis for *Law of Iterated Expectations*: *Definition 8 i Rahbek: "On conditional expectations":*
Den betingede forventning $E(X|Y = y)$ er en funktion af y og er givet ved

$$E(X|Y = y) = \int_{\mathbb{R}} x f_{X|Y}(x|y) dx$$

I vores tilfælde er der byttet om på X og Y . Vi sætter $\psi(x) = \int_{\mathbb{R}} y f_{Y|X}(y|x) dy$:

$$\begin{aligned} E(E(Y|X = x)) &= E(\psi(x)) \\ &= \int_{\mathbb{R}} \psi(x) f_X(x) dx \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} y f_{Y|X}(y|x) f_X(x) dy dx \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} y f_{X,Y}(x,y) dx dy \\ &= \int_{\mathbb{R}} y \int_{\mathbb{R}} f_{X,Y}(x,y) dx dy \\ &= \int_{\mathbb{R}} y f_Y(y) dy \\ &= E(Y) \end{aligned}$$

Opgave U45.5

Lad Z_1 og Z_2 være uafhængige standard $N(0, 1)$ fordelte. Definér

$$\begin{aligned} Y &= 2Z_1 + Z_2 \\ X &= 3Z_1 \end{aligned}$$

- Hvordan er (Y, X) fordelt?

(Y, X) er dannet ved en lineær transformation af (Z_1, Z_2) :

$$\begin{pmatrix} Y \\ X \end{pmatrix} = A \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \quad A = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix}$$

hvor $(Z_1, Z_2)'$ er $N(\mu, \Omega)$ fordelt med

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ og } \Omega = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Ifølge Egenskab G.5 i Rahbek: "Lidt regneregler" er $A(Y, X)'$ to-dimensionalt fordelt som $N(A\mu, A\Omega A')$.

Dermed

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N(\mu, \Omega), \quad \mu = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ og } \Omega = \begin{pmatrix} 2 & 1 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 6 & 9 \end{pmatrix}$$

- Find $E(Y|Z_1)$

$$\begin{aligned} E(Y|Z_1) &= E((2Z_1 + Z_2)|Z_1) \\ &= 2E(Z_1|Z_1) + E(Z_2|Z_1) \\ &= 2 \cdot Z_1 + 0 \\ &= \underline{\underline{2Z_1}} \end{aligned}$$

3. Find $E(X|Z_2)$

$$\begin{aligned} E(X|Z_2) &= E(3Z_1|Z_2) \\ &= 3E(Z_1) \\ &= \underline{\underline{0}} \end{aligned}$$

4. Find $E(Y|X)$

$$\begin{aligned} \mu_{Y|X} &= \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X) \\ &= 0 + \frac{6}{9}(X - 0) \\ &= \frac{2}{3}\underline{\underline{X}} \end{aligned}$$

Opgave U45.6

Denne opgave genopfrisker bl.a. diskrete fordelinger. lad X være en diskret ligefordelt variabel på mængden $\{-1, 0, 1\}$.

1. Find $E(X)$ og $V(X)$ samt $P(X \geq 0)$.

$$\begin{aligned} E(X) &= \frac{1}{3}(-1 + 0 + 1) = \underline{\underline{0}} \\ V(X) &= \frac{1}{3}((-1)^2 + 0^2 + 1^2) - 0^2 = \frac{2}{3} \\ P(X \geq 0) &= 1 - P(X = -1) = \frac{2}{3} \end{aligned}$$

2. Find $E(X|X \geq 0)$.

Benyt Definition 4 i Rahbek: "On conditional expectations":

$$\begin{aligned} E(X|X \geq 0) &= \sum_{x_i \geq 0} x_i P(X = x_i | X \geq 0) \\ &= \frac{\sum_{x_i \geq 0} x_i p(x_i)}{P(X \geq 0)} \\ &= \frac{\frac{1}{3}(0+1)}{\frac{2}{3}} \\ &= \frac{1}{2} \end{aligned}$$

3. Lad nu Y være en kontinuert ligefordelt variabel på intervallet $(-1, 1)$.
Find $E(Y)$ og $V(Y)$ samt $P(Y \geq 0)$.

Y har tætheden

$$f_Y(y) = \frac{1_{(-1,1)}(y)}{2}.$$

$$\begin{aligned}
 E(Y) &= \int_{-1}^1 \frac{y}{2} dy \\
 &= \left[\frac{1}{4}y^2 \right]_{-1}^1 \\
 &= \underline{\underline{0}} \\
 V(Y) &= \int_{-1}^1 \frac{y^2}{2} dy - 0^2 \\
 &= \left[\frac{1}{6}y^3 \right]_{-1}^1 \\
 &= \frac{1}{6} - \left(-\frac{1}{6} \right) \\
 &= \underline{\underline{\frac{1}{3}}} \\
 P(Y \geq 0) &= \int_0^1 \frac{1}{2} dy \\
 &= \left[\frac{1}{2}y \right]_0^1 \\
 &= \underline{\underline{\frac{1}{2}}}
 \end{aligned}$$

4. Find $E(Y|Y > 0)$.

Benyt *Property C.3* i *Rahbek*: "On conditional expectations":

$$\begin{aligned}
 E(Y|Y > 0) &= \int_0^1 y \frac{f_Y(y)}{P(Y \geq 0)} dy \\
 &= 2 \int_0^1 \frac{y}{2} dy \\
 &= 2 \left[\frac{1}{4}y^2 \right]_0^1 \\
 &= \underline{\underline{\frac{1}{2}}}
 \end{aligned}$$

Opgave U45.7

Denne opgave omhandler OLS regression og viser hvordan vi vha. observeret data kan estimere parametre i en nyttefunktion.

Forestil dig en økonomisk model, hvor agenter maksimerer deres nytte ved at vælge mellem to goder, C_1 og C_2 . Vi antager, at forbrugerne kun kan forbruge deres formue, w .

Nyttemaksimerings-problemet er da

$$\max_{c_1, c_2} C_1^\alpha C_2^{1-\alpha}, \text{ u.b.b. } P_1 C_1 + P_2 C_2 = w,$$

hvor vi antager, at agenterne har identiske Cobb-Douglas nyttefunktioner.

Man kan vise, at optimalt forbrug er givet ved

$$\begin{aligned} C_1 &= \alpha \frac{w}{P_1} \\ C_2 &= (1 - \alpha) \frac{w}{P_2}, \end{aligned}$$

sådan, at hvis vi tager ratioen får vi

$$C = \frac{\alpha}{1 - \alpha} P,$$

hvor $P = P_1/P_2$ er den relative pris og $C = C_1/C_2$ er forbrugsratioen.

Lad os nu forestille os, at vi har T perioder hvor prisen har varieret, sådan at vi har et sæt af stokastiske variable $\{C_t, P_t\}_{t=1}^T$. Vi antager, at vi har observeret forbrug med uafhængige målefejl, ε_t , med $E(\varepsilon_t) = 0$ for alle t .

En lineær regressionmodel for forbrugsratioen er således

$$C_t = \tilde{\alpha} P_t + \varepsilon_t,$$

hvor det antages, at $E(\varepsilon_t | P_t) = E(\varepsilon_t) = 0$.

- Vis, at den betingede middelværdi af forbrugsratioen er $E(C_t | P_t) = \tilde{\alpha} P_t$. Hvad er fortolkningen af $\tilde{\alpha}$ i forhold til vores model?

$$\begin{aligned} E(C_t | P_t) &= E((\tilde{\alpha}_t P_t + \varepsilon_t) | P_t) \\ &= \tilde{\alpha} E(P_t | P_t) + E(\varepsilon_t | P_t) \\ &= \tilde{\alpha} P_t. \end{aligned}$$

Vi har $\frac{\partial E(C_t | P_t)}{\partial P_t} = \tilde{\alpha}$. $\tilde{\alpha} = \frac{\alpha}{1-\alpha}$ udtrykker altså hvor meget den forventede værdi af forbrugsratioen ændrer sig, når den relative pris ændrer sig med én enhed.

- Udled OLS-estimatoren af $\tilde{\alpha} = \frac{\alpha}{1-\alpha}$ under antagelsen at, at vi har T sæt af stokastiske variable, $\{C_t, P_t\}_{t=1}^T$.

Vi benytter Law of Iterated Expectations til at vise, at $E(\varepsilon_t P_t) = 0$:

$$\begin{aligned} E(\varepsilon_t P_t) &= E(E(\varepsilon_t P_t | P_t)) \\ &= E(P_t E(\varepsilon_t)) \\ &= E(P_t \cdot 0) \\ &= 0 \end{aligned}$$

Indsætter vi $\varepsilon_t = C_t - \tilde{\alpha}P_t$ får vi:

$$\begin{aligned} E((C_t - \tilde{\alpha}_{OLS}P_t)P_t) &= 0 \Leftrightarrow \\ E(C_t P_t) &= \tilde{\alpha}_{OLS}E(P_t^2) \Leftrightarrow \\ \tilde{\alpha}_{OLS} &= \frac{E(C_t P_t)}{E(P_t^2)} \end{aligned}$$

OLS-estimatoren $\tilde{\alpha}$ minimerer den kvadrerede afstand mellem observationen C_t og den lineære prediktion af C_t givet P_t , $E(C_t|P_t) = \tilde{\alpha}P_t$. Afstanden mellem observeret og predikteret værdi af C_t er $\varepsilon_T = C_t - \tilde{\alpha}P_t$ or dermed har vi:

$$\tilde{\alpha}_{OLS} = \arg \min_{\tilde{\alpha}} E((C_t - \tilde{\alpha}P_t)^2).$$

Da funktionen vi minimerer er strengt konveks ved vi, at $\tilde{\alpha}_{OLS}$ opfylder førsteordensbetingelsen:

$$\begin{aligned} \left. \frac{\partial E((C_t - \tilde{\alpha}P_t)^2)}{\partial \tilde{\alpha}} \right|_{\tilde{\alpha}=\tilde{\alpha}_{OLS}} &= 0 \Leftrightarrow \\ -2E(P_t(C_t - \tilde{\alpha}_{OLS}P_t)) &= 0 \Leftrightarrow \\ E(C_t P_t) &= \tilde{\alpha}_{OLS}E(P_t^2) \Leftrightarrow \\ \tilde{\alpha}_{OLS} &= \frac{E(C_t P_t)}{E(P_t^2)} \end{aligned}$$

3. Udled *OLS*-estimatet af $\tilde{\alpha} = \frac{\alpha}{1-\alpha}$ under antagelsen af, at vi har T sæt af observationer af de underliggende stokastiske variable, $\{c_t, p_t\}_{t=1}^T$.

Når vi kun har observationer af realisationer af de stokastiske variable C_t og P_t så udnytter vi, at vi kan opstille en *unbiased*, eller *middlelret* estimator for den forventede værdi af en stokastisk variabel X , nemlig gennemsnittet af X :

$$E\left(\frac{1}{T} \sum_{t=1}^N X_t\right) = \frac{1}{T} \sum_{t=1}^N E(X_t) = E(X_t).$$

Har vi en række observationer af X , $\{x_t\}_{t=1}^T$, og antager vi, at alle observationer er realisationer fra den samme fordeling så er gennemsnittet af observationerne, $\frac{1}{T} \sum_{t=1}^T x_i$ et unbiased estimat af middelværdien af den stokastiske variabel X .

Dermed kan vi på baggrund af den udledte estimator

$$\tilde{\alpha}_{OLS} = \frac{E(C_t P_t)}{E(P_t^2)},$$

opstille estimatet:

$$\hat{\tilde{\alpha}}_{OLS} = \frac{\frac{1}{T} \sum_{t=1}^T c_t p_t}{\frac{1}{T} \sum_{t=1}^T p_t^2} = \frac{\sum_{t=1}^T c_t p_t}{\sum_{t=1}^T p_t^2}.$$

Estimatet af $\tilde{\alpha}$ kan også udledes som det argument der minimerer gennemsnittet af de kvadrerede afstande mellem observationen c_t og den lineære prediktion $\hat{c}_t = \tilde{\alpha}p_t$:

$$\hat{\tilde{\alpha}}_{OLS} = \arg \min_{\tilde{\alpha}} \frac{1}{T} \sum_{t=1}^T (c_t - \tilde{\alpha}p_t)^2.$$

Førsteordensbetingelsen er:

$$\begin{aligned}
 & \frac{\partial \frac{1}{T} \sum_{t=1}^T (c_t - \tilde{\alpha} p_t)^2}{\partial \tilde{\alpha}} \Bigg|_{\tilde{\alpha} = \hat{\alpha}_{OLS}} = 0 \Leftrightarrow \\
 & -\frac{2}{T} \sum_{t=1}^T p_t (c_t - \hat{\alpha}_{OLS} p_t) = 0 \Leftrightarrow \\
 & \frac{1}{T} \sum_{t=1}^T c_t p_t = \hat{\alpha}_{OLS} \frac{1}{T} \sum_{t=1}^T p_t^2 \Leftrightarrow \\
 & \hat{\alpha}_{OLS} = \frac{\frac{1}{T} \sum_{t=1}^T c_t p_t}{\frac{1}{T} \sum_{t=1}^T p_t^2} = \frac{\sum_{t=1}^T c_t p_t}{\sum_{t=1}^T p_t^2}.
 \end{aligned}$$

Centrale begreber

- Teoretiske og empiriske momenter:

Teoretisk middelværdi:

$$E(Y) = \sum_{i=1}^n y_i \cdot p(y_i)$$

Empirisk middelværdi:

$$m_y = \frac{1}{n} \sum_{i=1}^n y_i$$

Teoretisk varians:

$$V(Y) = E\left(Y - E(Y)\right)^2$$

Empirisk varians:

$$v_y = \frac{1}{n} \sum_{i=1}^n (y_i - m_y)^2$$

Opgave 1

Vi betragter et finansielt marked og observerer at n handler initieres. Vi lader y_i betegne vores observation af, om handlen går i orden og gennemføres, eller om der opstår problemer undervejs, så handlen annuleres. Variablen er kodet som

$$y_i = \begin{cases} 1 & \text{hvis handlen gennemføres} \\ 0 & \text{ellers} \end{cases}$$

Betrægt et datasæt $\{y_i\}_{i=1}^n$ med $n = 17$ observationer:

$$\{1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0\}.$$

1. Beskriv datasættet $\{y_i\}_{i=1}^n$ ved de relative frekvenser, dvs.

$$f_{y=1} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i = 1) \quad \text{og} \quad f_{y=0} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i = 0)$$

hvor $\mathbb{I}(\cdot)$ er indikator-funktionen. Find også de kumulerede relative frekvenser.

y_i antager værdien 0 syv gange, og værdien 1 ti gange. Derfor har vi:

$$f_{y=1} = \frac{1}{17} \cdot 7 \approx 0,41$$

$$f_{y=0} = \frac{1}{17} \cdot 10 \approx 0,59$$

	Frekvens	Kumuleret frekvens	Fraktion	Kumuleret fraktion
$y = 0$	Ingen handel	7	7	0,41
$y = 1$	Handel	10	17	0,59

2. Udregn det empiriske gennemsnit, m_y , for $\{y_i\}_{i=1}^n$, og vis at det er lig den relative frekvens, $m_y = f_{y=1}$.
 Vis også at dette sammenfald ikke er tilfældet, hvis $\{y_i\}_{i=1}^n$ i stedet er kodet som $\{2, 2, 1, 1, \dots, 2, 1\}$.

Empirisk gennemsnit:

$$m_y = \frac{0 * 7 + 1 * 10}{17} \approx 0,59 = f_{y=1}$$

Hvis sample space er $Y = \{1, 2\}$

$$m_y = \frac{1 * 7 + 10 * 20}{17} \approx 1,59 \neq f_{y=1}$$

3. Udregn den empiriske standard-afvigelse, empirisk skewness (skævhed) og empirisk kurtosis (topstøjhed) for det observerede datasæt, $\{y_i\}_{i=1}^n$. Se afsnit 2.2 i Nielsen (2017).

Beregnet i Stata:

Variabel	Standardafvigelse	Skewness	Kurtosis
y	0,51	-0,36	1,13

4. Vi vil nu tænke på den binære observation, y_i , som en realisation af en stokastisk variabel, Y_i . Den stokastiske variabel Y_i er karakteriseret ved en såkaldt Bernoulli fordeling

$$Y_i \sim \text{Bernoulli}(\theta),$$

hvor θ er en parameter, $0 < \theta < 1$. Bemærk, at Bernoulli fordelingen er en binomial-fordeling med antals-parameter lig én og sandsynlighedsfunktionen er givet ved:

$$f_{Y_i}(y|\theta) = \begin{cases} P(y=1) = \theta & \text{hvis } y=1 \\ P(y=0) = 1 - \theta & \text{hvis } y=0 \end{cases}$$

Find de fire første standardiserede momenter for Bernoulli fordelingen som funktion af θ , dvs. mean, variance, skewness og kurtosis.

Udregn de teoretiske momenter for tilfældet $\theta = m_y$, med m_y lig det empiriske gennemsnit udregnet ovenfor, og sammenligne med de empiriske momenter for data.

Diskuter betydningen af dette resultat.

De fire første momenter for Bernoulli-fordelingen med parameter θ . Tredje kolonne angiver de teoretiske momenter når $\theta = \frac{10}{17}$, og fjerde kolonne angiver de empiriske momenter beregnet i OxMetrics:

Moment	Formel	$\theta = \frac{10}{17}$	Stata
Gennemsnit:	θ	0,5882	0,5882
Varians:	$\theta(1 - \theta)$	0,2422	0,2574
Skewness:	$\frac{1-2\theta}{\sqrt{\theta(1-\theta)}}$	-0,3586	-0,3586
Kurtosis:	$\frac{1-6\theta(1-\theta)}{\theta(1-\theta)} + 3$	1,1286	1,1286

Den teoretiske og empiriske varians er forskellig, da Stata beregner den empiriske varians som

$$v_y = \frac{1}{n-1} \sum_{i=1}^n (y_i - m_y)^2.$$

Hvis der divideres med n is stedet for $n - 1$ fås det teoretiske resultat (se do-file).

Det ses, at for en given række af udfald af en binær variabel vil vi kunne lade gennemsnittet betegne sandsynlighedsparameteren i en Bernoulli fordeling. De teoretiske momenter i Bernoullifordelingen vil da være identiske med de empiriske momenter i data.

5. Find den empiriske median, altså 50% fraktilen for datasættet $\{y_i\}_{i=1}^n$. Overvej hvor nyttig medianen (eller andre fraktile) er til beskrivelse af en binær variabel.

Vi ser at medianen er 1, hvilket ikke er overraskende da den midterste observation på en ordnet række af udfaldene er 1.

Empiriske fraktile:

$$\alpha_i = \frac{i-1}{n-1}$$

hvor n er antallet af forskellige observationer og i er placeringen på den ordnede liste af observationer af den observation som du vil beregne fraktil af.

Med binære data er $n = 2$ og vi får så 2 fraktile:

$$\alpha_1 = \frac{1-1}{2-1} = 0 \text{ (0\% - fraktil)}$$

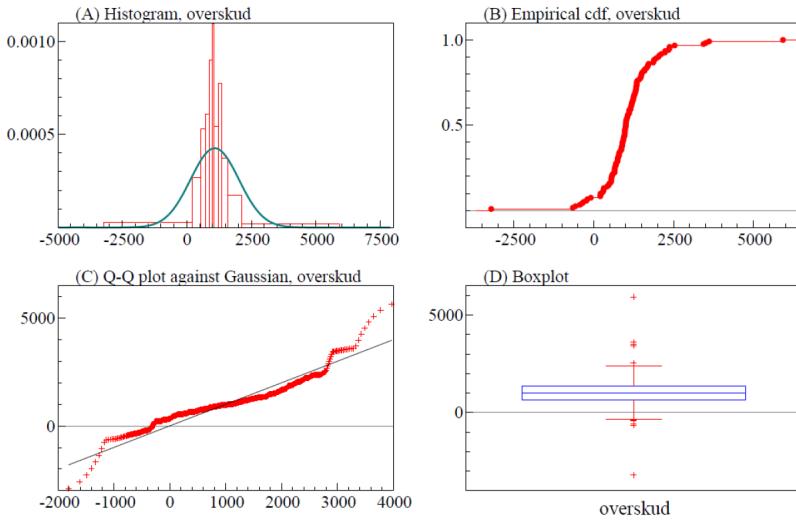
og

$$\alpha_2 = \frac{2-1}{2-1} = 1 \text{ (100\% - fraktil)}$$

Fraktile er ikke nyttige til beskrivelse af binære variable

Opgave 2

Figuren nedenfor viser beskrivende statistik for $n = 121$ enkeltmandsvirksomheders overskud et givet år, variablen overskud $_i$ $i = 1, 2, \dots, 121$.



1. Forklar hvad figurerne viser og hvordan de er konstrueret baseret på beskrivende statistik udregnet som funktion af $\{\text{overskud}_i\}_{i=1}^{121}$.

- Øverste figur til venstre er et *histogram*, hvor højden af de enkelte søjler angiver densiteten af observationer i det interval som søjlen dækker. Søjlerne er kodet til at indeholde det samme antal observationer. Derfor er de forskellig i bredden, men har samme areal.
- Øverste figur til højre er den *empiriske fordelingsfunktion (CDF)*. Den empiriske sandsynlighed stiger med $1/n$ for hvert ordnet datapunkt. *CDF*-grafen stiger hurtigst omkring gennemsnittet hvor der er flest observationer.
- Nederste figur til venstre er et *Q-Q plot*. Den lige linje indikerer hvor perfekt normalfordelte realisationer ville befinde sig. Det ses at de laveste observationer er lavere, og de højeste observationer er højere end perfekt normalfordelte realisationer.
- Nederste figur til højre er et *boxplot*, hvor boksen i midten dækker intervallet fra 25%- til 75%-fraktilen, benævnt *inter-quartile range (iqr)*. Den vandrette linje i boksen angiver medianen af observationerne. Whiskers over og under boksen angiver 1,5 gange iqr. Observationerne uden for disse whiskers betragtes som outliers.

Opgave 3

Vi vil nu se på karakterfordelingen fra jeres gymnasiale uddannelse sammen med jeres nuværende tidsforbrug på studiet. Load spørgeskema-svarene ind i STATA.

Lad x_i være jeres gymnasiale gennemsnit (variabel: `gns_gym`) og y_i være gennemsnitligt antal ugentlige timer brugt på studiet (variabel: `studietimer`). Vi har disse data for $i = 1, \dots, N$.

1. *Er disse variable diskrete eller kontinuerte (vi vil antage at udfaldsrummet er kontinuert)?*

Antal studietimer er angivet i hele tal mens gymnasiegennemsnittet er angivet med én decimal. Begge variable er strengt taget diskrete, da de er begrænsede i deres mulige værdier.

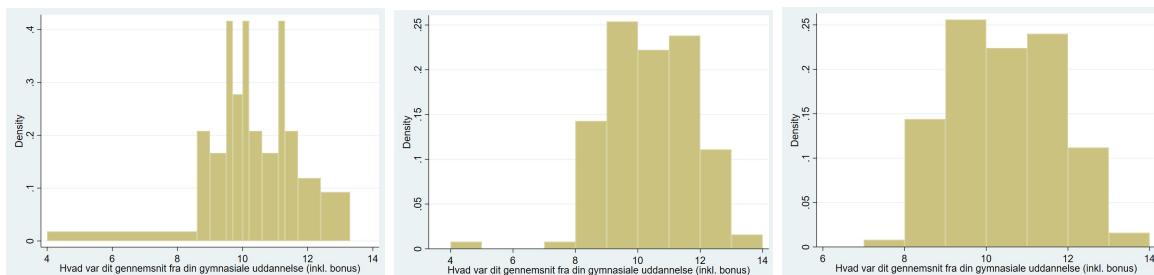
Især for gymnasiegennemsnittet gælder dog, at antallet af mulige værdier (over 100) er så højt at variablen i praksis vil blive betragtet som kontinuert.

2. *Undersøg karakterer og studietimer med histogram, empiriske momenter og fraktiler.*

Undersøg i opstillingen af histogrammet om valg af antal intervaller og interval-endepunkter har betydning for histogrammets udseende.

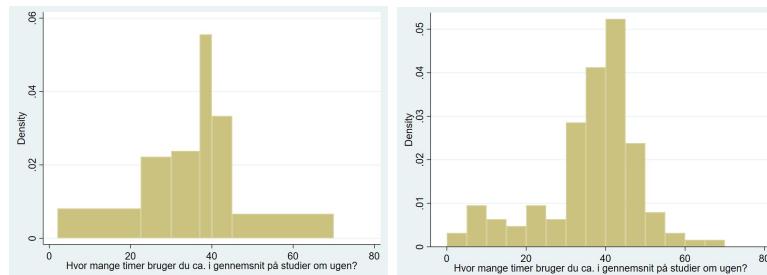
Histogrammer kan enten konstruerws med et fast antal observationer i hver søjle, eller med fast bredde af søjlerne.

Histogram over karakterer med 1/12 sample i hver søjle (venstre) og én karakter i hver søjle (højre):



Histogrammet længst til højre er uden den ekstreme observation af karakterer (=4). Når denne outlier fjernes er fordelingen mere symmetrisk.

Histogrammer for studietimer:



Deskriptiv statistik for de to variable:

. su gns_gym, d				. su studietimer, d			
Hvad var dit gennemsnit fra din gymnasiale uddannelse (inkl. bonus)				Hvor mange timer bruger du ca. i gennemsnit på studier om ugen?			
Percentiles	Smallest			Percentiles	Smallest		
1%	7.6	4		1%	4	2	
5%	8.3	7.6		5%	8	4	
10%	8.6	8	Obs	10%	15	5	Obs
25%	9.5	8.1	Sum of Wgt.	25%	30	5	Sum of Wgt.
50%	10.2			50%	37		Mean
		Mean	10.32857				34.34524
		Largest	1.402219				Std. Dev.
75%	11.3	12.7		75%	40	55	
90%	12.3	12.7	Variance	98%	45	55	Variance
95%	12.6	13.2	Skewness	95%	50	64	Skewness
99%	13.2	13.3	Kurtosis	99%	64	70	Kurtosis
			4.975115				- .6329457
							3.730545

3. Definér den empiriske korrelation mellem y_i og x_i .
Hvordan skal korrelationen fortolkes?

Den empiriske korrelation er givet ved:

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})} \sqrt{\sum_{i=1}^n (y_i - \bar{y})}},$$

hvor \bar{x} og \bar{y} er gennemsnittet af x of y .

Korrelationen er normeret til intervallet $[-1, 1]$, hvor $\rho = -1$ er perfekt negativ lineær sammenhæng, $\rho = 0$ er ingen sammenhæng og $\rho = 1$ er perfekt lineær sammenhæng.

4. Undersøg om der i datasættet er en klar korrelation mellem tidsforbrug og karakterer i gymnasiet.

Korrelation beregnet i STATA:

. pwcorr(gns_gym studietimer), sig

		gns_gym	studie~r
		gns_gym	1.0000
		studietimer	0.1874 1.0000 0.0356

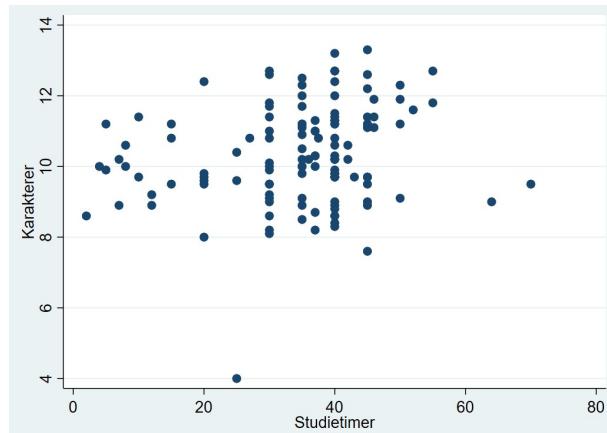
Korrelationen er på 0,19 og er signifikant forskellig fra nul på et 5% signifikansniveau.

5. Tegn også et krydsplot mellem y_i og x_i og for tolk resultatet i sammenhæng med den udregnede korrelation.

Hvad kan man lære af dette?

Hvad er den økonometriske fortolkning af sammenhængen mellem tidsforbrug og karakterer i gymnasiet i jeres spørgeskemasvar?

Krydsplot mellem karakterer og studietimer:



En grafisk inspektion af sammenhængen mellem karakterer og studietimer indikerer ikke en klar sammenhæng mellem de to variable.

På udregningen af den empiriske korrelation ses det dog, at der er en klar signifikant sammenhæng. Grafisk og matematisk analyse bør altid kombineres.

Den økonometriske fortolkning er, at der lader til at være en positiv sammenhæng mellem karakterer i gymnasiet og nuværende tidsforbrug på studiet.

Opgave 4

Vi vil nu se på jeres alkoholforbrug i forhold til jeres boligsituation. Load spørgeskemasvarene ind i stata.

Vi skal buge variablene genstande og bolig, som angiver hhv. gennemsnitligt antal ugentlige genstande og boligsituationen, som er defineret som

$$bolig_i = \begin{cases} 1 & \text{hvis bor alene eller sammen med kæreste/ægtefælle} \\ 2 & \text{hvis bor hos forældre} \\ 3 & \text{hvis bor med andre i delelejlighed} \\ 4 & \text{hvis bor på kollegie} \\ 5 & \text{hvis ingen af ovenstående} \end{cases}$$

1. Beskriv de to variable ved fordelingerne (pdf) og kumulerede fordelinger (cdf)
Hvor stor en andel indtager mindst 7 genstande om ugen?

Frekvenser, andele og kumulerede andele fra STATA:

-> tabulation of genstande				
Hvor mange genstande alkohol drikker du ca. i gennemsnit om ugen?	Freq.	Percent	Cum.	
0	6	4.76	4.76	
.1	1	0.79	5.56	
.25	1	0.79	6.35	
1	6	4.76	11.11	
2	12	9.52	20.63	
3	9	7.14	27.78	
4	7	5.56	33.33	
5	12	9.52	42.86	
6	5	3.97	46.83	
7	12	9.52	56.35	
8	5	3.97	60.32	
10	19	15.08	75.40	
12	4	3.17	78.57	
14	5	3.97	82.54	
15	8	6.35	88.89	
16	1	0.79	89.68	
Total	126	100.00	30	2
				1.59
				100.00

-> tabulation of bolig				
Hvordan er dine boligforhold?	Freq.	Percent	Cum.	
1: Bor alene eller sammen med kæreste/æ	40	31.75	31.75	
2: Bor hos forældre	23	18.25	50.00	
3: Bor med andre i delelejlighed	52	41.27	91.27	
4: Bor på kollegie	10	7.94	99.21	
5: Ingen af de ovenstående	1	0.79	100.00	
Total	126	100.00		
				2
				1.59
				100.00

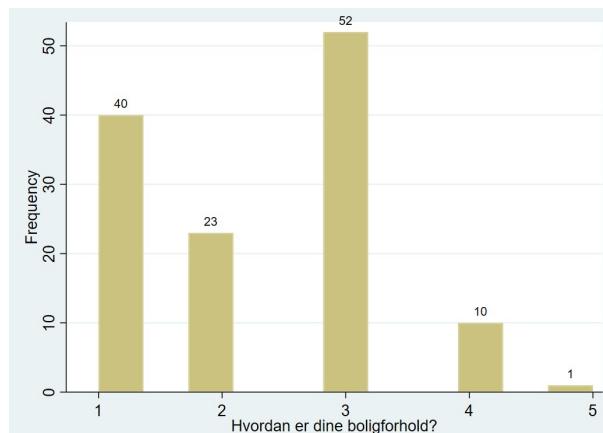
Andelen der indtager mindst 7 genstande om ugen:

$$\frac{100 - 46,83}{100} = \underline{\underline{0,5317}}$$

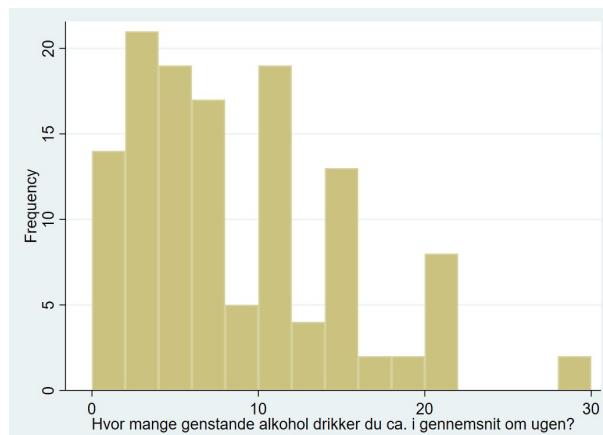
2. Udregn empiriske momenter og kommentér på resultaterne.

Udregn median og kvartiler og kommentér på resultaterne.

Da bolig er en kategorisk variabel giver det ikke mening at beregne deskriptiv statistik på denne. Fordelingen af besvarelser på de fem boligkategorier kan dog illustreres i et histogram:



Histogram for *genstande*:



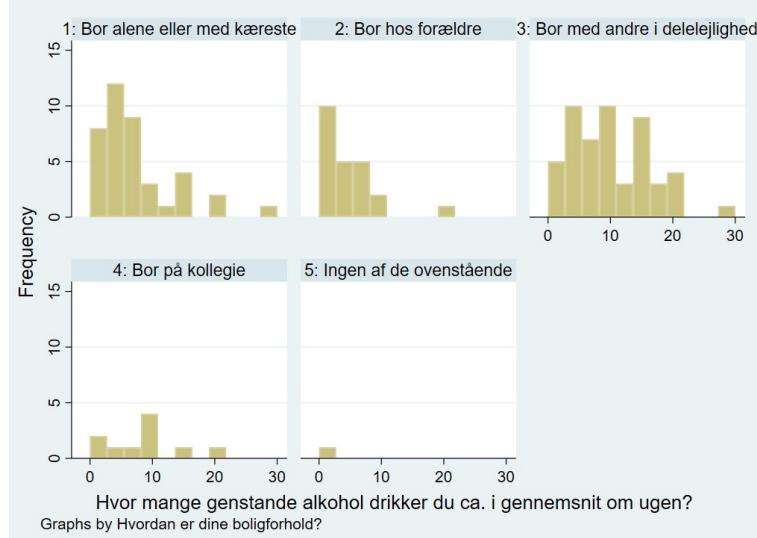
Momenter, median og kvartiler for *genstande*:

. su genstande, d			
Hvor mange genstande alkohol drikker du ca. i gennemsnit om ugen?			
Percentiles	Smallest		
1%	0	0	
5%	.1	0	
10%	1	0	Obs 126
25%	3	0	Sum of Wgt. 126
50%	7		Mean 8.074206
		Largest	Std. Dev. 6.267873
75%	10	20	
90%	17	21	Variance 39.28623
95%	20	30	Skewness 1.022227
99%	30	30	Kurtosis 3.975818

Histogrammet viser en højreskæv fordeling. Dette bekræftes af, at gennemsnittet 8,07 er højere end medianen 7 samt at skævheden er højere end nul. Afstanden mellem kvartilerne (*iqr*) er 7 genstande. De højeste observationer ligge altså 2-3 gange *iqr* over 75%-fraktilen.

3. Beregn det gennemsnitlige alkoholforbrug for hver af de fem grupper. Er der en systematisk forskel? Fortolk resultaterne.

Histogrammer for alkoholforbrug opdelt på boligtype:



På baggrund af histogrammerne ser det ud til, at studerende der bor i delelejlighed eller på kollegie har et højere alkoholforbrug.

Gennemsnit for genstande betinget på bolig:

-> bolig = 1: Bor alene eller sammen			-> bolig = 2: Bor hos forældre		
variable	mean	N	variable	mean	N
genstande	7.45	40	genstande	4.276087	23
-> bolig = 3: Bor med andre i delel			-> bolig = 4: Bor på kollegie		
variable	mean	N	variable	mean	N
genstande	10.21154	52	genstande	8.9	10
-> bolig = 5: Ingen af de ovenståen					
variable	mean	N	variable	mean	N
genstande	1	1			

Studerende som bor hos forældre har et markant lavere gennemsnitligt alkoholforbrug. Studerende som bor alene har et lidt lavere gennemsnitligt forbrug end studerende som bor med andre i delelejlighed eller på kollegie.

4. Beskriv også fordelingen(pdf og cdf) af genstande for de to grupper: bolig $\in\{1,2\}$ versus bolig $\in\{3,4,5\}$.

pdf og cdf af genstande for bolig $\in\{1,2\}$ og bolig $\in\{3,4,5\}$:

. tab genstande if bolig<3				. tab genstande if bolig>=3			
Hvor mange genstande alkohol drikker du ca. i gennemsnit om ugen?	Freq.	Percent	Cum.	Hvor mange genstande alkohol drikker du ca. i gennemsnit om ugen?	Freq.	Percent	Cum.
0	6	9.52	9.52	1	4	6.35	6.35
.1	1	1.59	11.11	2	4	6.35	12.70
.25	1	1.59	12.70	3	4	6.35	19.05
1	2	3.17	15.87	4	3	4.76	23.81
2	8	12.70	28.57	5	4	6.35	30.16
3	5	7.94	36.51	7	5	7.94	38.10
4	4	6.35	42.86	8	3	4.76	42.86
5	8	12.70	55.56	10	14	22.22	65.08
6	5	7.94	63.49	12	3	4.76	69.84
7	7	11.11	74.60	14	4	6.35	76.19
8	2	3.17	77.78	15	6	9.52	85.71
10	5	7.94	85.71	17	1	1.59	87.30
12	1	1.59	87.30	18	2	3.17	90.48
14	1	1.59	88.89	20	4	6.35	96.83
15	2	3.17	92.06	21	1	1.59	98.41
16	1	1.59	93.65	30	1	1.59	100.00
20	3	4.76	98.41	Total	63	100.00	
30	1	1.59	100.00				
Total	63	100.00					

Opgave 5

Datasættet i filen `titanic.xls` indeholder oplysninger om 2201 passagerer på RMS Titanic, som sank i vinteren 1912. Der er information om passagerernes overlevelsessstatus,

$$survived_i = \begin{cases} 1 & \text{hvis person } i \text{ overlevede} \\ 0 & \text{ellers} \end{cases}$$

og deres køn,

$$female_i = \begin{cases} 1 & \text{hvis person } i \text{ var kvinde} \\ 0 & \text{ellers} \end{cases}$$

deres aldersklasse,

$$child_i = \begin{cases} 1 & \text{hvis person } i \text{ var et barn} \\ 0 & \text{ellers} \end{cases}$$

og hvilken klasse de rejste på

$$class_i = \begin{cases} 1 & \text{hvis person } i \text{ rejste på første klasse} \\ 2 & \text{hvis person } i \text{ rejste på anden klasse} \\ 3 & \text{hvis person } i \text{ rejste på tredje klasse} \\ 0 & \text{hvis person } i \text{ var besætningsmedlem} \end{cases}$$

1. Lav en beskrivende statistik af de fire variable i datasættet.

Overvej hvordan man mest hensigsmæssigt sammenfatter informationen i hver enkelt variabel.

Som vi har set tidligere giver det ikke mening at beregne momenter og fraktiler for binære variable. For kategorivariable som den der inddeler klasserne er momenter og fraktiler heller ikke nyttige.

De relative frekvenser og kumulerede relative frekvenser giver til gengæld et godt billede af fordelingen.

survived	Freq.	Percent	Cum.
0	1,490	67.70	67.70
1	711	32.30	100.00
Total	2,201	100.00	
female	Freq.	Percent	Cum.
0	1,731	78.65	78.65
1	470	21.35	100.00
Total	2,201	100.00	
child	Freq.	Percent	Cum.
0	2,092	95.05	95.05
1	109	4.95	100.00
Total	2,201	100.00	

class	Freq.	Percent	Cum.
0	885	40.21	40.21
1	325	14.77	54.98
2	285	12.95	67.92
3	706	32.08	100.00
Total	2,201	100.00	

For de binære variable ved vi fra tidligere, at den relative frekvens for realisationen 1 af den stokastiske variabel er lig med gennemsnittet, og desuden kan fortolkes som sandsynligheden for at være henholdsvis *overlevet, kvinde eller barn*.

2. *Det er velkendt, at der var for få redningsbåde om bord på Titanic. Selskabet bag Titanic havde en politik om at redde kvinder og børn først i tilfælde af ulykke, parolen "Women and children first!".*



Brug beskrivende statistik til at undersøge om datasættet indikerer at kvinder faktisk overlevede i større omfang end mænd.

Gør dette ved først at beskrive den empiriske fordeling af overlevelsen hos mænd og kvinder separat. Konstruer dernæst en 2×2 tabel til at belyse spørgsmålet.

Overvej om selskabets parole blev overholdt og om der er belæg for overskriften i The New York Herald.

Overlevelse betinget på køn.

Kvinder:

-> female = 1

survived	Freq.	Percent	Cum.
0	126	26.81	26.81
1	344	73.19	100.00
Total	470	100.00	

Mænd:

-> female = 0

survived	Freq.	Percent	Cum.
0	1,364	78.80	78.80
1	367	21.20	100.00
Total	1,731	100.00	

Vi ser, at der var 2201 passagerer ombord på Titanic. 470 kvinder hvoraf 344 overlevede, og 1731 mænd hvoraf 367 overlevede.

Andelen af overlevende mænd og kvinder er angivet i nedenstående 2×2 tabel

	Kvinde	Mand
Andel overlevet	0,73	0,21
Andel ikke-overlevet	0,27	0,79

Overskriften i *The New York Herald* er forkert i og med at hovedparten af overlevende ikke var kvinder og børn. Der var faktisk flere mænd end kvinder der overlevede i absolutte tal. Dog var overlevelseschancen langt højere for kvinder (73%) end for mænd 21%.

Overlevelseschancen for børn var 52% hvilket altså var højere end for mænd, men lavere end for kvinder.

-> child = 1			
survived	Freq.	Percent	Cum.
0	52	47.71	47.71
1	57	52.29	100.00
Total	109	100.00	

3. Undersøg med en ny antals-tabel om overlevelsrateen varierede med rejsekasse og om besætningens overlevesrate er højere eller lavere end de rejsendes.

Fordeling for passagerer på første klasse:

-> class = 1			
survived	Freq.	Percent	Cum.
0	122	37.54	37.54
1	203	62.46	100.00
Total	325	100.00	

Fordeling for passagerer på anden klasse:

-> class = 2			
survived	Freq.	Percent	Cum.
0	167	58.60	58.60
1	118	41.40	100.00
Total	285	100.00	

Fordeling for passagerer på tredje klasse:

-> class = 3			
survived	Freq.	Percent	Cum.
0	528	74.79	74.79
1	178	25.21	100.00
Total	706	100.00	

Fordeling for besætningsmedlemmer:

-> class = 0				
survived	Freq.	Percent	Cum.	
0	673	76.05	76.05	
1	212	23.95	100.00	
Total	885	100.00		

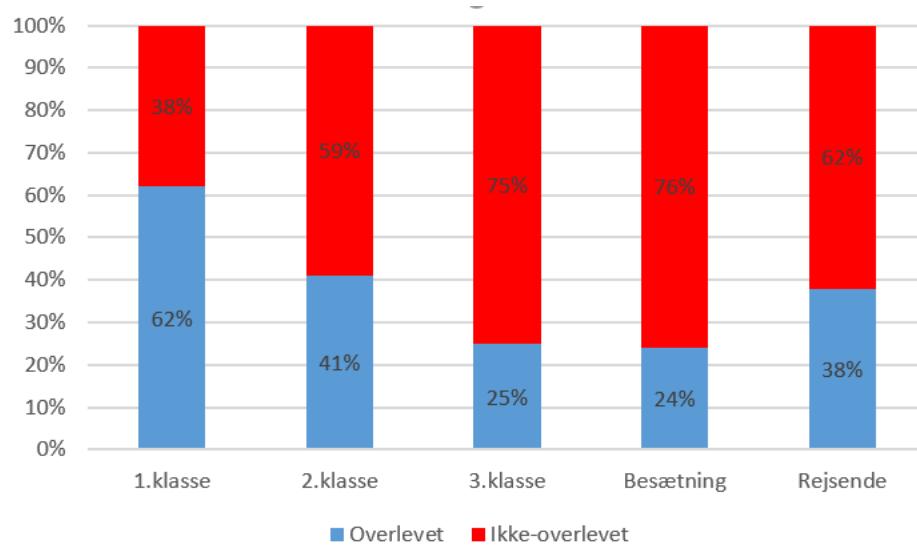
Fordeling for rejsende generelt:

. tab survived if class>0				
survived	Freq.	Percent	Cum.	
0	817	62.08	62.08	
1	499	37.92	100.00	
Total	1,316	100.00		

Andelen af overlevende i de forskellige grupper angivet i tabel:

	1.klasse	2.klasse	3.klasse	Besætning	Rejsende
Andel overlevet	0,62	0,41	0,25	0,24	0,38
Andel ikke-overlevet	0,38	0,59	0,75	0,76	0,62

Som nedenstående figur viser er andelen af overlevende højere jo dyrere klasse passagerene rejste på. Desuden var der ca. 50% flere overlevende blandt rejsende i forhold til besætningsmedlemmer.



Opgave 6

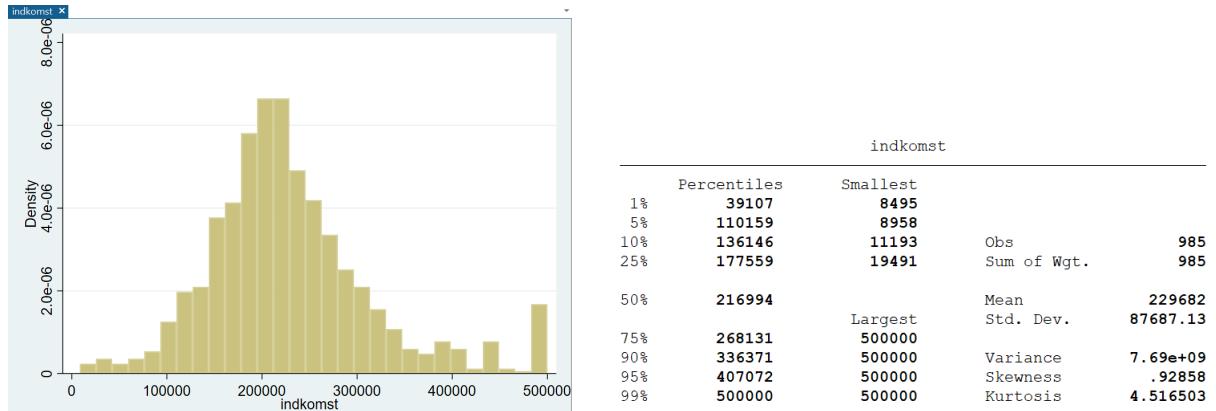
Datasættet i filen PolitData1994.xls indeholder en tilfældigt udvalgt stikprøve bestående af 985 danskere for året 1994. Observationerne i datasættet vedrører personer mellem 26 og 60 år der alle har fuldtidsbeskæftigelse, og datasættet indeholder følgende informationer:

- indkomst_i Registreret brutto-indkomst for individ i i kr.
- alder_i Alder af individ i ultimo 1994.
- kvinde_i Dummy. Antager værdien 1 hvis individ i er kvinde.

1. Beskriv bruttoindkomsten for hele stikprøven, $\{y_i\}_{i=1}^n$.

Kommentér på fordelingens karakteristika. Er den empiriske fordeling af indkomsten fx. symmetrisk eller skæv?

Histogram og deskriptiv statistik af indkomst:

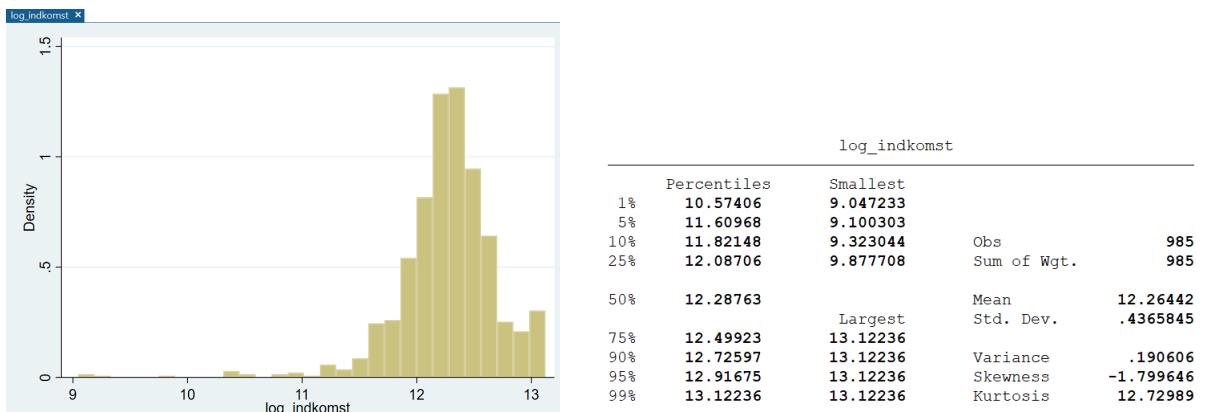


Det ses på histogrammet at fordelingen er højreskæv, hvilket bekræftes af positiv skævhed og et gennemsnit over medianen. Kurtosis er 4,52 og dermed højere end i normalfordelingen. Der er altså flere ekstreme observationer end i en normalfordeling.

2. Betragt nu den log-transformerede indkomst, $x_i = \log(y_i)$. Denne transformation kan laves i STATA ved at skrive gen $x=\log(y)$

Beskriv den transformerede variabel, $\{x_i\}_{i=1}^n$, og forklar hvordan fordelingen ændres af transformasjonen.

Histogram og deskriptiv statistik af log indkomst:

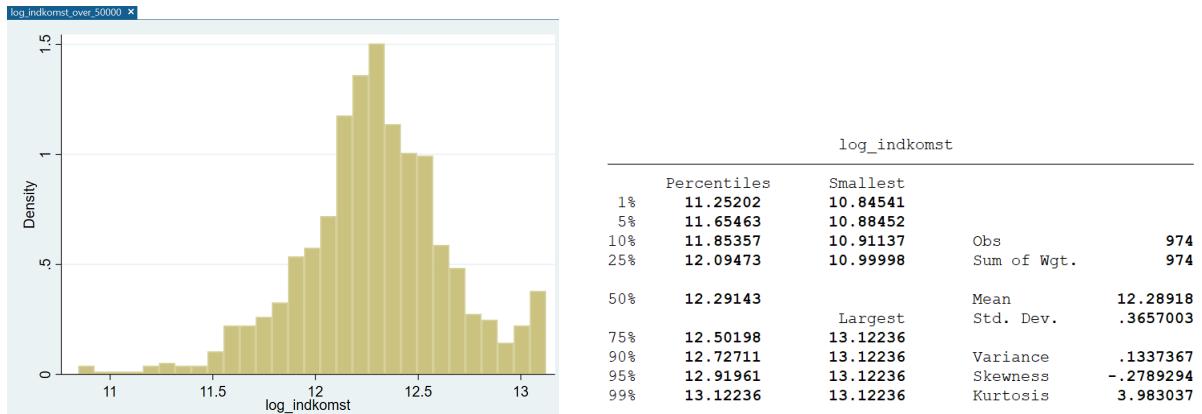


Fordelingen er nu ventreskæv, hvilket bekræftes af negativ skævhed og et gennemsnit lige under medianen. Kurtosis er meget høj, 12,73.

3. Nogle personer har bemærkelsesværdigt lave indkomster givet at de er registreret med fuldtidsbeskæftigelse.

Beskriv nu fordelingen af log-indkomsterne $\{x_i\}_{i=1}^n$ for de personer der har en indkomst større end 50.000kr. Hvordan ændrer fordelingen sig?

Histogram og deskriptiv statistik af log indkomst for indkomster over 50.000:



Fordelingen er nu mere symmetrisk, hvilket bekræftes af skævhed tættere på nul og at gennemsnit og median næsten er identiske.

4. Beskriv nu indkomstfordelingen for mænd og kvinder og sammenlign. Du kan enten bruge de rå data eller data i logaritmer efter hvad du synes giver mest mening.

Kommentér på resultaterne.

Indkomst betinget på køn.

Kvinder:

-> kvinde = 1		indkomst			
Percentiles	Smallest	Percentiles	Smallest		
1%	64193	51298	1%	109482	80204
5%	104639	53344	5%	159611	82294
10%	119383	54796	10%	184167	102693
25%	150604	59873	25%	207180	103646
50%	188411		50%	244357	
		Largest	Mean		264383.7
			Std. Dev.		85758.49
75%	230753	495746	75%	301156	500000
90%	275264	500000	90%	390191	500000
95%	314072	500000	95%	450219	500000
99%	487363	500000	Kurtosis	6.402371	500000

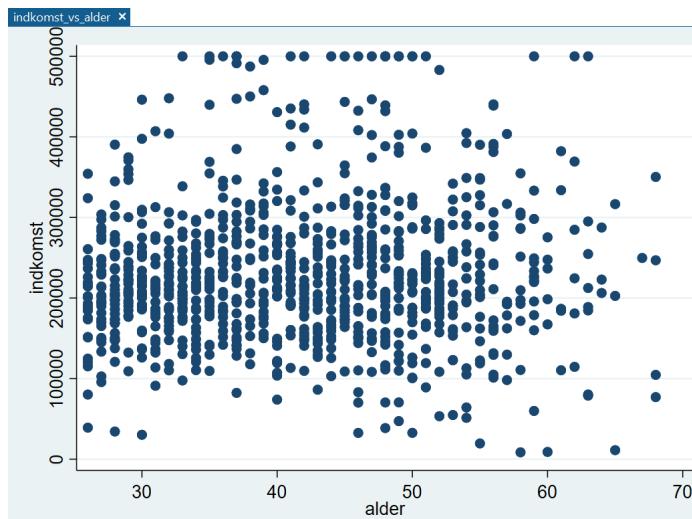
Mænd:

-> kvinde = 0		indkomst			
Percentiles	Smallest	Percentiles	Smallest		
1%	109482	80204	1%	109482	80204
5%	159611	82294	5%	159611	82294
10%	184167	102693	10%	184167	102693
25%	207180	103646	25%	207180	103646
50%	244357		50%	244357	
		Largest	Mean		264383.7
			Std. Dev.		85758.49
75%	301156	500000	75%	301156	500000
90%	390191	500000	90%	390191	500000
95%	450219	500000	95%	450219	500000
99%	500000	500000	Kurtosis	500000	500000

Den gennemsnitlige indkomst er markant højere for mænd. Kvinders indkomst er mere skævt fordelt end mænds.

5. Undersøg ved anvendelse af beskrivende statistik om det ser ud til at være en sammenhæng mellem indkomst og alder.

Korrelationen mellem alder og indkomst er tæt på nul, 0,0494. Der lader altså ikke til at være korrelation mellem alder og indkomst. Dette bekræftes af et scatterplot af observationerne af invigernes alder og indkomst:



Likelihood funktionen og Maximum Likelihood Estimation

For at kunne lave estimation ved Maximim Likelihood skal følgende gælde:

Assumption 3.1

- i) Den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y|\theta_i), \quad i = 1, 2, \dots, n,$$

hvor $f_{Y_i}(\cdot|\cdot)$ angiver sandsynlighedsfunktionen for Y_i . y_i er en realisering af Y_i og θ_i er sandsynlighedsfunktionens parameter.

- ii) De stokastiske variable er identisk fordelt således, at

$$\theta_i = \theta_j = \theta \quad \text{og} \quad \forall i, j \quad | \quad f_{Y_i}(y|\theta) = f_{Y_j}(y|\theta)$$

parameteren $\theta = (\theta_1, \dots, \theta_k)'$ er et element i parameterrummet $\theta \in \Theta \subseteq \mathbb{R}^k$

- iii) De stokastiske variable Y_i og Y_j er uafhængige for alle $i \neq j$ således, at den simultane sandsynlighed er givet ved

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n|\theta) &= f_{Y_1}(y_1|\theta) \cdot f_{Y_2}(y_2|\theta) \cdot \dots \cdot f_{Y_n}(y_n|\theta) \\ &= \prod_{i=1}^n f_{Y_i}(y_i|\theta) \end{aligned}$$

Opgave 1

Denne opgave gennemgår trinnen i opstillingen af likelihood funktionen og udledning af maksimum likelihood estimatoren. Nedenfor bliver I bedt om at gøre det samme for andre eksempler, men fremgangsmåden er altid den samme.

Vi vil opstille en model for antallet af patenter et firma udtager, og vi har observeret antallet af patenter for $n = 21$ virksomheder $\{y_i\}_{i=1}^n$, nemlig

$$\{y_i\}_{i=1}^n = \{3, 1, 4, 4, 0, 2, 4, 8, 3, 5, 5, 2, 2, 2, 1, 3, 5, 3, 2, 0, 4\}$$

sådan at $\sum_{i=1}^{21} y_i = 63$.

Vi tænker på observationerne som realisationer af stokastiske variable $\{Y_i\}_{i=1}^n$ og vi antager, at:

$$Y_i \stackrel{d}{=} \text{Poisson}(\lambda_i) \quad , \quad i = 1, 2, \dots, n$$

sådan at sandsynlighedsfunktionen er givet ved:

$$f_{Y_i}(y|\lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^y}{y!}$$

- Opskriv udfaldsrummet for de stokastiske variable, \mathbb{Y} , så $Y_i \in \mathbb{Y}$.

Opskriv parameter-rummet, Θ , så $\lambda_i \in \Theta$.

Y_i kan teoretisk antage alle positive hele tal samt 0:

$$\underline{Y_i \in \mathbb{Y} = \mathbb{N}_0 = \{0, 1, 2, \dots\}}$$

Parameteren λ_i kan antage alle positive reelle tal:

$$\underline{\lambda_i \in \Theta = \{\lambda \in \mathbb{R} : \lambda > 0\}}$$

- Antag, at de stokastiske variable er identisk fordele, sådan at $\lambda_i = \lambda$ for alle i .

Skriv sandsynlighedsfunktionen for dette tilfælde og diskutér om det er en realistisk forudsætning. Ændrer diskussionen sig hvis du får at vide at nogle af virksomhederne er fra IT branchen og nogle er fra service-branchen.

Da parameteren λ er den samme for alle stokastiske processor kan vi fjerne fodtegnet. Sandsynligheden for udfaldet y_i :

$$\underline{f_{Y_i}(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}}$$

For at opstille en likelihood model er det nødvendigt at antage at alle virksomheder, på tværs af brancher, har samme sandsynlighed for at udtage et givet antal patenter. Dette er nok ikke en realistisk forudsætning da virksomheder inden for højteknologi-brancher som *IT* eller *medico* må forventes at udtage flere patenter end virksomheder indenfor eksempelvis *bygge og anlæg* eller *servicesektoren*.

3. Giv definitionen på, at Y_i og Y_j er uafhængige stokastiske variable for $i \neq j$.

Antag at de stokastiske variable $\{Y_i\}_{i=1}^n$ er uafhængige og opskriv den samlede sandsynlighedsfunktion, $f_{Y_1, \dots, Y_n}(y_1, \dots, y_n | \lambda)$, for dette tilfælde.

Diskutér om det er en realistisk forudsætning i dette tilfælde.

Hvis Y_i og Y_j er uafhængige gælder:

$$f_{Y_1, Y_2}(y_1, y_2 | \lambda) = f_{Y_1}(y_1 | \lambda) \cdot f_{Y_2}(y_2 | \lambda)$$

og hvis vi antager, at de stokastiske variable $\{Y_i\}_{i=1}^n$ er uafhængige er den samlede sandsynlighedsfunktion givet ved:

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n | \lambda) &= f_{Y_1}(y_1 | \lambda) \cdot f_{Y_2}(y_2 | \lambda) \cdot \dots \cdot f_{Y_n}(y_n | \lambda) \\ &= \prod_{i=1}^n f_{Y_i}(y_i | \lambda) \end{aligned}$$

Igen kan man argumentere for, at den nødvendige uafhængighedsantagelse ikke er realistisk. Indenfor en given branche kan antallet af udtagne patenter være en konkurrencefaktor, hvilket vil betyde at et firmas patent vil anspore konkurrerende firmaer til også at udtagte patenter.

4. Sample likelihood funktionen, $L(\lambda | y_1, \dots, y_n)$, er defineret som sandsynlighedsfunktionen, men nu som funktion af λ ,

$$L(\lambda | y_1, \dots, y_n) = f_{Y_1, \dots, Y_n}(y_1, \dots, y_n | \lambda).$$

Opskriv sample likelihood funktionen og forklar hvordan værdien af likelihood funktionen kan fortolkes.

Opskriv også log-likelihood funktionen.

Opskriv også log-likelihood funktionen som funktion af de stokastiske variable $\{Y_i\}_{i=1}^n$. Forklar forskellen på $\log L(\lambda | y_1, \dots, y_n)$ og $\log L(\lambda | Y_1, \dots, Y_n)$.

Når vi opstiller Sample likelihood funktionen ”bytter vi om” på det der betinges på, så vi får sandsynligheden for parameteren betinget på observationerne:

$$L(\lambda | y_1, \dots, y_n) = f_{Y_1, \dots, Y_n}(y_1, \dots, y_n | \lambda).$$

Hver enkelt realisation y_i af den stokastiske variabel Y_i bidrager til den samlede likelihood med værdien af sandsynlighedsfunktionen evalueret i netop denne realisation:

$$\ell(\lambda | y_i) = f_{Y_i}(y_i)$$

hvor $\ell(\cdot | \cdot)$ betegner likelihood bidraget fra den enkelte observation.

Da vi har antaget identiske og uafhængige antigelser kan vi udtrykke den samlede likelihood for en given række af realisationer af en stokastisk variabel som produktet af likelihood bidragene med samme parameter for hver realisation:

$$L(\lambda | y_1, \dots, y_n) = \prod_{i=1}^n \ell(\lambda | y_i)$$

For observationerne af patenter er sample likelihood funktionen givet ved:

$$L(\lambda|y_1, \dots, y_n) = \prod_{i=1}^n \ell(\lambda|y_i) = \prod_{i=1}^n \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}$$

Hvor $n = 21$ i vores model for antal patenter.

I en model med diskret sandsynlighedsfunktion kan værdien af likelihood funktionen fortolkes som sandsynligheden for de observerede data som funktion af parameteren λ .

Log-likelihood funktionen som funktion af observationer:

$$\begin{aligned} \log L(\lambda|y_1, \dots, y_n) &= \log \left(\prod_{i=1}^n \ell(\lambda|y_i) \right) \\ &= \sum_{i=1}^n \log \ell(\lambda|y_i) \\ &= \sum_{i=1}^n \log \left(\frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \right) \\ &= \sum_{i=1}^n \left(y_i \log(\lambda) - \lambda - \log(y_i!) \right) \end{aligned}$$

Log-likelihood funktionen som funktion af de stokastiske variable:

$$\begin{aligned} \log L(\lambda|Y_1, \dots, Y_n) &= \log \left(\prod_{i=1}^n \ell(\lambda|Y_i) \right) \\ &= \sum_{i=1}^n \log \ell(\lambda|Y_i) \\ &= \sum_{i=1}^n \log \left(\frac{\exp(-\lambda)\lambda^{Y_i}}{Y_i!} \right) \\ &= \sum_{i=1}^n \left(Y_i \log(\lambda) - \lambda - \log(Y_i!) \right) \end{aligned}$$

Hvis log-likelihood funktionen angives som funktion af observationer, y_i (realiserede værdier af en stokastisk variabel) så er funktionsværdien et givet tal, et *estimat*.

Hvis log-likelihood funktionen derimod angives som funktion af stokastiske variable, Y_i , så er funktionsværdien også en stokastisk variabel, en *estimator*.

5. *Maximum likelihood estimatoren er defineret som:*

$$\hat{\lambda}(Y_1, \dots, Y_n) = \arg \max_{\lambda \in \Theta} \log L(\lambda|Y_1, \dots, Y_n).$$

Find førsteordens betingelsen for et maksimum.

$$FOC : \frac{\partial \log L(\hat{\lambda}|Y_1, \dots, Y_n)}{\partial \hat{\lambda}} = 0$$

6. Løs førsteordensbetingelsen og find maximum likelihood estimatoren, $\hat{\lambda}(Y_1, \dots, Y_n)$

Da vi har antaget at alle udfald af den stokastiske variabel Y er identisk fordelt gælder:

$$\begin{aligned}\frac{\partial \log L(\lambda|Y_1, \dots, Y_n)}{\partial \lambda} &= \frac{\partial \sum_{i=1}^n \log \ell(\lambda|Y_i)}{\partial \lambda} \\ &= \sum_{i=1}^n \frac{\partial \log \ell(\lambda|Y_i)}{\partial \lambda} \\ &= \sum_{i=1}^n s_i(\lambda)\end{aligned}$$

hvor $s_i(\lambda)$, den første afledte af likelihood bidraget, betegner *score-bidraget* fra hver enkelt stokastiske variabel.

Vi finder altså den første afledte, *den samlede score*, af likelihood funktionen som summen af score bidrag:

$$\begin{aligned}S(\lambda) &= \sum_{i=1}^n s_i(\lambda) = \sum_{i=1}^n \frac{\partial \log \ell(\lambda|Y_i)}{\partial \lambda} \\ &= \sum_{i=1}^n \left(\frac{Y_i}{\lambda} - 1 \right) \\ &= \frac{\sum_{i=1}^n Y_i}{\lambda} - n\end{aligned}$$

Når vi sætter dette lig 0 kan vi udlede maximum likelihood estimatoren $\hat{\lambda}(Y_1, \dots, Y_n)$.

$$\begin{aligned}\frac{\sum_{i=1}^n Y_i}{\hat{\lambda}} - n &= 0 \Leftrightarrow \\ \frac{\sum_{i=1}^n Y_i}{\hat{\lambda}} &= n \Leftrightarrow \\ \hat{\lambda} &= \underline{\underline{\frac{1}{n} \sum_{i=1}^n Y_i}}\end{aligned}$$

7. Find andenordensbettingelsen og argumentér for, at løsningen er et maksimum

Den anden afledte af hver stokastisk variabel udgør *Hessian-bidraget*:

$$H_i(\lambda) = \frac{\partial^2 \log \ell(\lambda|Y_i)}{\partial \lambda \partial \lambda'} = \frac{\partial}{\partial \lambda} s_i(\lambda) = -\frac{Y_i}{\lambda^2}$$

og hvis vi summerer over alle observationer får vi den anden afledte af log-likelihood funktionen, den samlede *Hessian*:

$$H(\lambda) = \sum_{i=1}^{21} H_i(\lambda) = -\frac{\sum_{i=1}^{21} Y_i}{\lambda^2}$$

Hvis vi udelukker tilfældet hvor $y_i = 0$ for alle i (se delopgave 9), så er $\sum_{i=1}^{21} Y_i > 0$. Den anden afledte er dermed negativ. Likelihood funktionen og log-likelihood funktionen evalueret i $\hat{\lambda}$ er et

globalt maksimum og giver dermed den maksimale log-likelihood værdi.

8. Indsæt informationen fra de observerede data og find maximum likelihood estimatet, $\hat{\lambda}(y_1, \dots, y_{21})$:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{21} \times 63 = \underline{\underline{3}}$$

9. Overvej, hvad der ville ske hvis ingen firmaer har udtaget patenter, $\sum_{i=1}^{21} y_i = 0$

Da vil estimatet være lig nul:

$$\hat{\lambda} = E(Y_i) = \text{Var}(Y_i) = 0.$$

0 er ikke del af parameterrummet for en Poissonfordelt stokastisk variabel, så i tilfældet $\sum_{i=1}^n y_i = 0$ kan Poissonfordelingen ikke bruges. Vores antagelse om at vi kender sandsynlighedsfunktionen for den stokastiske variabel Y_i er derfor ikke opfyldt, og modellen er fejlspecifieret.

Opgave 2

Vi er ansat til at undersøge den forventede levetid af nogle nye elsparepærer. Lad den stokastiske variabel, $Y_i \in \mathbb{Y} = \{y \in \mathbb{Y} | y > 0\}$ betegne levetiden af elsparepære i , $i = 1, 2, \dots, n$. Efter tilbagemelding fra forbrugerne har vi n observerede levetider,

$$\{y_i\}_{i=1}^n = \{y_1, y_2, \dots, y_n\}$$

Fra tidligere undersøgelser ved vi, at levetiden for en enkelt elsparepære approksimativt følger en eksponentialfordeling med parameter θ , dvs. for $i = 1, 2, \dots, n$ gælder, at

$$Y_i \stackrel{d}{=} \text{exponential}(\theta), \quad 0 < \theta < \infty$$

Eksponential-fordelingen er beskrevet ved tæthedsfunktionen,

$$f_{Y_i}(y|\theta) = \theta \exp\{-\theta y\}, \quad y \in \mathbb{Y}.$$

Parameteren, θ , kan tolkes sådan at $E(Y_i) = \theta^{-1}$.

1. Hvad er forskellen på Y_i og y_i i præsentationen ovenfor?

Hvad er y i tæthedsfunktionen?

Skitsér tæthedsfunktionen som en funktion af y eller find et billede af den på Wikipedia. Alternativt kan STATA bruges.

Y_i er den stokastiske variabel i den statistiske model vi bruger til at beskrive data.
 y_i er den egentlige observation som er en realisering af Y_i .

For at få en idé om, hvad y udtrykker i tæthedsfunktionen kan man tænke på eksponentialfordelingen for $y_i = n$ som $n - 1$ uafhængige poisson-processor med udfald 0 og udfald 1 i den

n'te process:

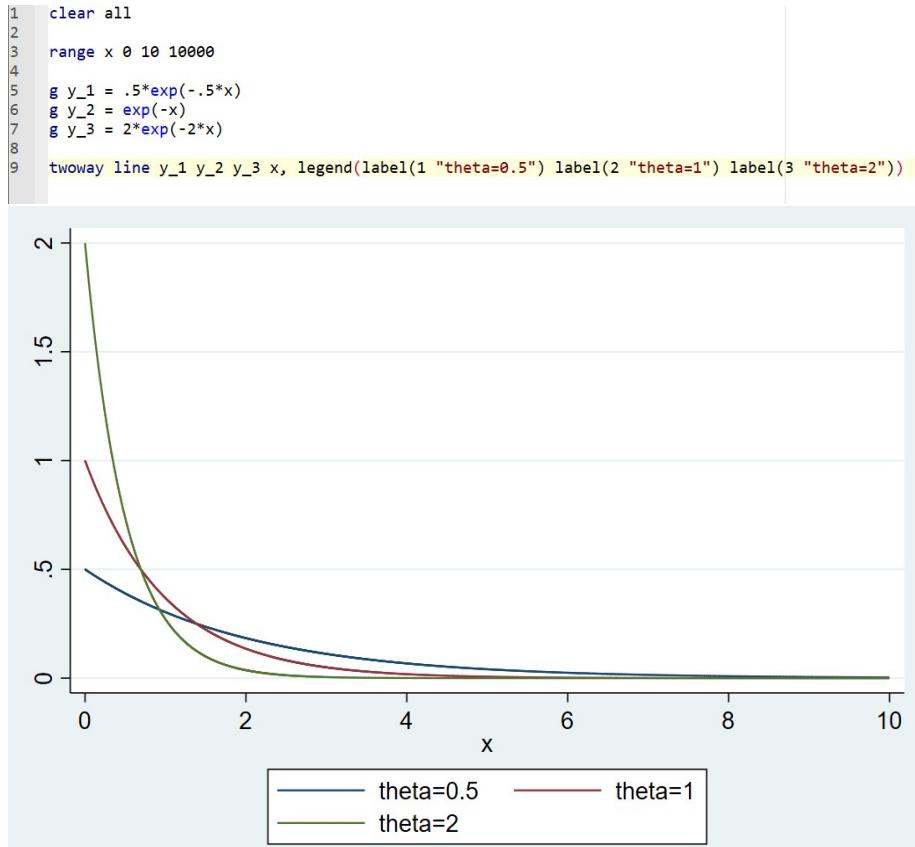
$$\begin{aligned}
 y_i = n : f_{Y_i}(y_i = n|\lambda) &= \left(\frac{\lambda^0}{0!} \exp(-\lambda) \right)^{n-1} \cdot \frac{\lambda^1}{1!} \exp(-\lambda) \\
 &= \exp(-\lambda)^{n-1} \cdot \lambda \exp(-\lambda) \\
 &= \lambda \exp(-\lambda n)
 \end{aligned}$$

Udskifter vi λ med θ får vi:

$$\begin{aligned}
 f_{Y_i}(y_i = n|\theta) &= \theta \exp\{-\theta n\} \Leftrightarrow \\
 f_{Y_i}(y_i|\theta) &= \theta \exp\{-\theta y_i\}
 \end{aligned}$$

Dermed kan vi fortolke y som nummeret på den tidsenhed hvor elsparepåren går ud. Eller dens levetid.

Illustration af tæthedsfunktionen for forskellige parametre (genereret i STATA med nedenstående kode):



Husk $E(Y_i) = \frac{1}{\theta}$. Jo højere parameteren λ er, jo mere sadsynlighedsmasse tæt ved 0, og jo lavere middelværdi.

2. Hvor i opstillingen ovenfor er det (implicit) angivet at de stokastiske variable er identisk fordelte?

Hvorfor er det vigtigt i opskrivningen af modellen?

Fordelingen for Y_i er angivet således:

$$Y_i \stackrel{d}{=} \text{exponential}(\theta), \quad 0 < \theta < \infty$$

Der er intet fodtegn på θ , så den samme parameter gælder for alle Y_i .

Dette er vigtigt i opskrivningen af modellen, da det er på grund af den *identiske fordeling* vi kan udlede generelle resultater for udfaldene af den stokastiske proces.

3. Antag, at de stokastiske variable, $\{Y_i\}_{i=1}^n$, er uafhængigt fordelte og opskriv en statistisk model for de observerede levetider, $\{y_i\}_{i=1}^n$, dvs. sample-likelihood funktionen,

$$L(\theta|y_1, y_2, \dots, y_n), \quad \theta \in \Theta$$

og parameter-området Θ .

Find log-likelihood funktionen for data-sættet $\{y_i\}_{i=1}^n$.

Sample-likelihood funktionen:

$$L(\theta|y_1, y_2, \dots, y_n), \quad \theta \in \Theta$$

ved identisk eksponentiel-fordelte og uafhængige data er givet ved:

$$\begin{aligned} L(\theta|y_1, y_2, \dots, y_n) &= \prod_{i=1}^n \ell(\theta|y_i) \\ &= \prod_{i=1}^n \theta \exp(-\theta y_i), \quad \theta \in \Theta = \{\theta \in \mathbb{R} : \theta > 0\} \end{aligned}$$

Log-likelihood:

$$\begin{aligned} \log L(\theta|y_1, y_2, \dots, y_n) &= \sum_{i=1}^n \log \ell(\theta|y_i) \\ &= \sum_{i=1}^n \log (\theta \exp(-\theta y_i)) \\ &= \sum_{i=1}^n (\log \theta - \theta y_i) \end{aligned}$$

4. Hvorfor er det vigtigt at antage uafhængighed?

Diskuter om i.i.d. antagelserne kan være problematiske i vores tilfælde.

Hvis vi ikke antager uafhængighed kan vi ikke udtrykke den simultane sandsynlighed af alle hændelser ved produktet af de marginale sandsynligheder. Dvs. ingen sample-likelihood funktion.

Hvis kort levetid af en elsparepære skyldes en maskinfel på en given produktionsdag, kan det medføre at observationer af kort levetid kommer i klynger og dermed ikke er uafhængige.

5. Find maximum likelihood estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$, ved at løse førsteordensbetingelsen for maksimum af log-likelihood funktionen.

Kontrollér at andenordens-betingelsen er opfyldt, sådan at $\hat{\theta}(Y_1, Y_2, \dots, Y_n)$ er et maksimum.

Samlet score:

$$\begin{aligned} S(\theta) &= \sum_{i=1}^n s_i(\theta) = \sum_{i=1}^n \frac{\partial \log \ell(\theta | Y_i)}{\partial \theta} \\ &= \sum_{i=1}^n \left(\frac{1}{\theta} - Y_i \right) \\ &= \frac{n}{\theta} - \sum_{i=1}^n Y_i \end{aligned}$$

FOC:

$$\begin{aligned} S(\hat{\theta}) &= \frac{n}{\hat{\theta}} - \sum_{i=1}^n Y_i = 0 \Leftrightarrow \\ \frac{n}{\hat{\theta}} &= \sum_{i=1}^n Y_i \Leftrightarrow \\ \hat{\theta} &= \frac{n}{\sum_{i=1}^n Y_i} \end{aligned}$$

$\hat{\theta}$ er altså aftagende i Y_i . Tænk på sammenhængen med $E(Y_i) = \frac{1}{\theta}$.

For at kontrollere andenordensbettingelsen finder vi først *Hessian-bidraget*:

$$H_i(\theta) = \frac{\partial^2 \log \ell(\theta | Y_i)}{\partial \theta \partial \theta'} = \frac{\partial}{\partial \theta} s_i(\theta) = -\frac{1}{\theta^2}$$

og summerer over alle hændelser for at få den anden afledte:

$$H(\theta) = \sum_{i=1}^n H_i(\theta) = -\frac{n}{\theta^2} < 0$$

Da dette er entydigt negativt har vi fundet et maksimum for log-likelihood funktionen, og dermed den parameter θ som maksimerer sandsynligheden af vores observationer.

6. Vi får nu de observerede levetider fra forbrugerne,

$$\{y_1, y_2, \dots, y_i, \dots, y_n\} = \{1.1, 2.3, \dots, 2.1\}.$$

Der er $n = 120$ levetider og den samlede levetid udregnes til $\sum_{i=1}^{120} y_i = 252,0$ år.

Find maximum likelihood estimatet, $\hat{\theta}(y_1, \dots, y_n)$.

Med formlen udledt ovenfor får vi følgende maximum likelihood estimat:

$$\hat{\theta}(y_1, \dots, y_{120}) = \frac{n}{\sum_{i=1}^n y_i} = \frac{120}{252} \approx \underline{\underline{0,48}}$$

7. Diskuter forskellen på $\hat{\theta}(y_1, \dots, y_{120})$ og $\hat{\theta}(Y_1, \dots, Y_{120})$

Hvis man læser i en artikel, at $\hat{\theta}_n = 9,25$, menes der så estimatet eller estimatoren?

Hvis man læser i en artikel, at $\hat{\theta}$ er god, fordi den har en lille varians menes der så estimatet eller estimatoren?

$\hat{\theta}(y_1, \dots, y_{120}) \approx 0,48$ er et tal. Det bedste bud på levetiden af en elsparepære i år.
 $\hat{\theta}(Y_1, \dots, Y_{120})$ er en stokastisk variabel med forventet værdi $E(Y_i) = 0,48$

Hvis artiklen påstår, at el sparepære holder i 9,25 år, så menes der et estimat af levetiden. Et fast tal.

Da varians er en egenskab ved en stokastisk variabel menes der estimatoren. Hvis levetiden kan beskrives ved en Poisson-process, så er variansen også 9,25, hvilket ikke er en lille varians.

Opgave 3

Det viser sig nu, at pærerne i opgave 2 ikke var identisk fordelte, men at der var to forskellige typer, nemlig nye og gamle, så den modificerede model er givet ved,

$$Y_i \stackrel{d}{=} \begin{cases} \text{exponential}(\theta_1) & \text{for } i = 1, 2, \dots, 75 \\ \text{exponential}(\theta_2) & \text{for } i = 76, 77, \dots, 120 \end{cases}$$

Y_i og Y_j er stadig uafhængige, for alle $i \neq j$.

- Opskriv sample likelihood funktionen for $\{y_1, \dots, y_{75}, y_{76}, \dots, y_{120}\}$, samt parameterrummet Θ , sådan at $\theta = (\theta_1, \theta_2)' \in \Theta$.

Da vi er så heldige at vide, hvilke pærer der er nye og gamle, kan vi lave to adskilte sample likelihood funktioner, og multiplicere dem for at få den samlede score:

$$L(\theta | y_1, \dots, y_{75}, y_{76}, \dots, y_{120}) = \prod_{i=1}^{75} \theta_1 \exp\{-\theta_1 y_i\} \times \prod_{i=76}^{120} \theta_2 \exp\{-\theta_2 y_i\},$$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \Theta = \{\theta \in \mathbb{R}^2 | \theta > 0\}$$

2. Opskriv log-likelihood funktionen for data-sættet, $\{y_i\}_{i=1}^n$.

$$\begin{aligned}\log L(\theta|y_1, \dots, y_{120}) &= \sum_{i=1}^{75} \log \ell(\theta_1|y_i) + \sum_{i=76}^{120} \log \ell(\theta_2|y_i) \\ &= \sum_{i=1}^{75} (\log \theta_1 - \theta_1 y_i) + \sum_{i=76}^{120} (\log \theta_2 - \theta_2 y_i)\end{aligned}$$

3. Find maximum likelihood estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$, ved at løse førsteordensbetingelsen.

Samlet score:

$$\begin{aligned}S(\theta) &= \frac{\partial \log L(\theta|Y_1, \dots, Y_{120})}{\partial \theta} \\ &= \left(\sum_{i=1}^{75} \frac{\partial \ell(\theta_1|Y_i)}{\partial \theta_1} \right) \\ &= \left(\sum_{i=76}^{120} \frac{\partial \ell(\theta_2|Y_i)}{\partial \theta_2} \right) \\ &= \left(\sum_{i=1}^{75} \left(\frac{1}{\theta_1} - Y_i \right) \right) \\ &= \left(\sum_{i=76}^{120} \left(\frac{1}{\theta_2} - Y_i \right) \right) \\ &= \left(\frac{75}{\theta_1} - \sum_{i=1}^{75} Y_i \right) \\ &= \left(\frac{45}{\theta_2} - \sum_{i=76}^{120} Y_i \right)\end{aligned}$$

Førsteordensbetinger:

$$\begin{aligned}S(\hat{\theta}_1) &= \frac{75}{\hat{\theta}_1} - \sum_{i=1}^{75} Y_i = 0 \Leftrightarrow \\ \hat{\theta}_1 &= \frac{75}{\sum_{i=1}^{75} Y_i}\end{aligned}$$

og

$$\begin{aligned}S(\hat{\theta}_2) &= \frac{45}{\hat{\theta}_2} - \sum_{i=76}^{120} Y_i = 0 \Leftrightarrow \\ \hat{\theta}_2 &= \frac{45}{\sum_{i=76}^{120} Y_i}\end{aligned}$$

4. Vi får nu at vide, at $\sum_{i=1}^{75} y_i = 160$ og $\sum_{i=76}^{120} y_i = 92$.

Find maximum likelihood estimaterne og kommenter på resultatet. Er de nye pærer bedre end de gamle?

Baseret på data får vi estimaterne:

$$\hat{\theta}_1 = \frac{75}{160} \approx 0,47$$

og

$$\hat{\theta}_2 = \frac{45}{92} \approx 0,49$$

Hvis de 75 pærer er nye, kan vi konkludere, at den forventede levetid for henh. nye og gamle pærer er:

$$E(Y_{ny}) = \frac{1}{\theta_1} = \frac{1}{0,47} \approx 2,13$$

$$E(Y_{gammel}) = \frac{1}{\theta_2} = \frac{1}{0,49} \approx 2,04$$

så de nye pærer holder lidt længere.

5. Forklar, hvordan vi kan for tolke den udvidede model som et eksempel på en betinget model for $Y_i|X_i = x_i$.

Hvis Y_i er en stokastisk process der angiver levetiden for en elsparepære, og X_i er en (binær) stokastisk variabel som angiver, hvorvidt en pære er *ny* eller *gammel*, så har vi i det foregående vist, at processerne ikke er uafhængige. Den forventede levetid for pæren er en funktion af x .

Opgave 4

Betrægt nu en diskret stokastisk variabel $Y_i \in \mathbb{Y} = \{0, 1, 2, 3\}$ med sandsynlighedsfunktion givet ved

$$P(Y_i = y_i) = \begin{cases} \frac{2\theta}{3} & \text{hvis } y = 0 \\ \frac{\theta}{3} & \text{hvis } y = 1 \\ \frac{2(1-\theta)}{3} & \text{hvis } y = 2 \\ \frac{(1-\theta)}{3} & \text{hvis } y = 3 \end{cases}$$

hvor parameteren $\theta \in [0; 1]$ er ukendt.

1. Vis, at for $y \in \mathbb{Y}$, kan sandsynligheden skrives som

$$f_{Y_i}(y|\theta) = \left(\frac{2\theta}{3}\right)^{\mathbb{I}(y=0)} \left(\frac{\theta}{3}\right)^{\mathbb{I}(y=1)} \left(\frac{2(1-\theta)}{3}\right)^{\mathbb{I}(y=2)} \left(\frac{(1-\theta)}{3}\right)^{\mathbb{I}(y=3)}$$

For et givet y vil funktionen netop antage den rette sandsynlighed, og de andre faktorer vil være lig 1.

Ex.:

$$f_{Y_i}(0|\theta) = \left(\frac{2\theta}{3}\right)^1 \left(\frac{\theta}{3}\right)^0 \left(\frac{2(1-\theta)}{3}\right)^0 \left(\frac{(1-\theta)}{3}\right)^0 = \left(\frac{2\theta}{3}\right)$$

2. Antag nu, at vi har $n = 10$ observationer givet ved

$$\{y_1, \dots, y_{10}\} = \{3, 0, 2, 1, 3, 2, 1, 0, 2, 1\}.$$

Opskriv sample likelihood funktionen for $\{y_i\}_{i=1}^{10}$ og log-likelihood funktionen.

Der er 2 observationer af 0 og 3 samt 3 observationer af 1 og 2.

Sample likelihood:

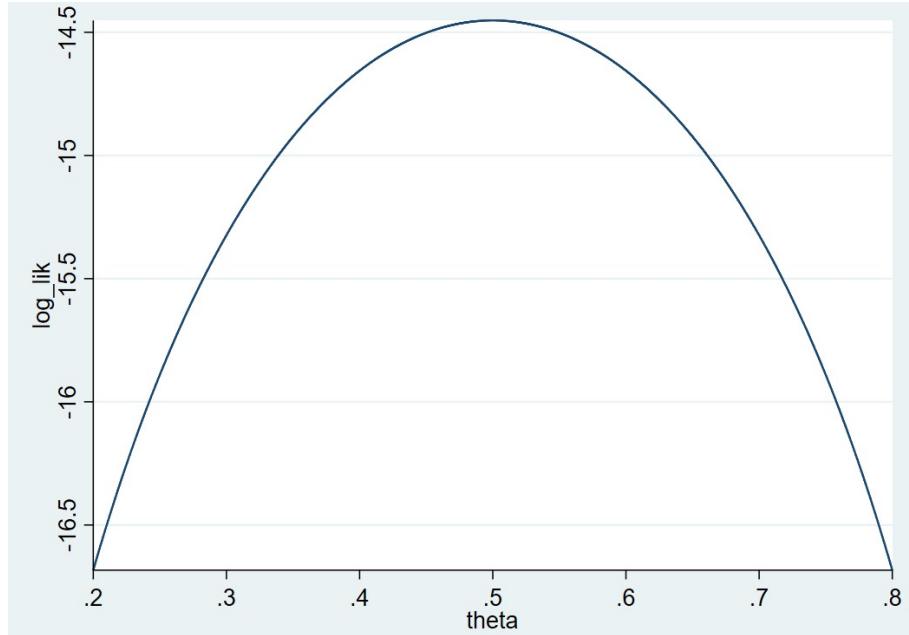
$$L(\theta|y_1, \dots, y_{10}) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{(1-\theta)}{3}\right)^2$$

Log-likelihood:

$$\begin{aligned} \log L(\theta|y_1, \dots, y_{10}) &= 2\log\left(\frac{2\theta}{3}\right) + 3\log\left(\frac{\theta}{3}\right) + 3\log\left(\frac{2(1-\theta)}{3}\right) + 2\log\left(\frac{(1-\theta)}{3}\right) \\ &= 2\log\left(\frac{2}{3}\right) + 2\log\theta + 3\log\left(\frac{1}{3}\right) + 3\log\theta + 3\log\left(\frac{2}{3}\right) + 3\log(1-\theta) + 2\log\left(\frac{1}{3}\right) + 2\log(1-\theta) \\ &= 5 \left[\log\left(\frac{2}{3}\right) + \log\left(\frac{1}{3}\right) + \log\theta + \log(1-\theta) \right] \\ &= 5 \left[\log\left(\frac{2}{9}\right) + \log(\theta - \theta^2) \right] \end{aligned}$$

3. Tegn log-likelihood funktionen $\log L_n(\theta)$ som funktion af θ .

STATA-plot:



4. Find maximum likelihood estimatet $\hat{\theta}(y_1, \dots, y_{10})$.

Vi finder her maximum likelihood estimatet på baggrund af de givne data, og ikke generelt. Førsteordensbetingelsen:

$$S(\hat{\theta}) = \frac{\partial \log L(y_1, \dots, y_{10})}{\partial \hat{\theta}} = 5 \cdot \frac{1 - 2\hat{\theta}}{\hat{\theta} - \hat{\theta}^2} = 0 \Leftrightarrow 1 - 2\hat{\theta} = 0 \Leftrightarrow \hat{\theta} = 0,5$$

5. Udregn sandsynlighederne, $P(Y_i = y_i)$, for $y_i \in \mathbb{Y}$

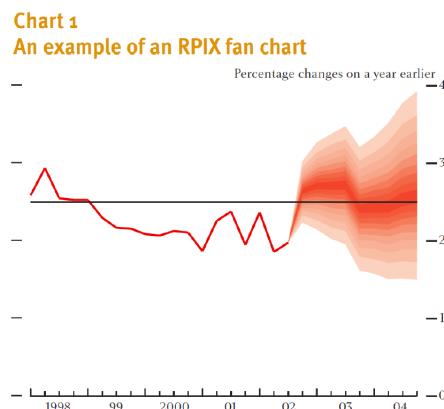
$$P(Y_i = y_i) = \begin{cases} \frac{1}{3} & \text{hvis } y = 0 \quad (\text{observeret andel : } \frac{2}{10}) \\ \frac{1}{6} & \text{hvis } y = 1 \quad (\text{observeret andel : } \frac{3}{10}) \\ \frac{1}{3} & \text{hvis } y = 2 \quad (\text{observeret andel : } \frac{3}{10}) \\ \frac{1}{6} & \text{hvis } y = 3 \quad (\text{observeret andel : } \frac{2}{10}) \end{cases}$$

6. I likelihood analysen er det vigtigt at undersøge, om den statistiske model ser ud til at passe nogenlunde på data. Undersøg om den påståede model giver sandsynligheder som ser ud til at passe nogenlunde til de observerede data.

Vi ser at sandsynlighederne beregnet på baggrund af maximum likelihood estimatet *ikke* passer godt med de observerede andele. Så modellen er højst sandsynligt misspecifieret.

Ekstraopgave

Bank of England offentliggør inflationsforecast i form af såkaldte fancharts og figuren nedenfor viser et eksempel på et fan-chart. Figuren er konstrueret så den faktiske inflation forventes at ligge indenfor det mørkeste område med 10% sandsynlighed, indenfor det næst-mørkeste med 20% sandsynlighed, osv., sådan at inflationen med 90% sandsynlighed vil ligge indenfor det lysest skraverede område.



1. Er forecast-fordelingen for den viste periode symmetrisk eller vurderes der at være markant anderledes risiko for afvigelser opad og nedad?

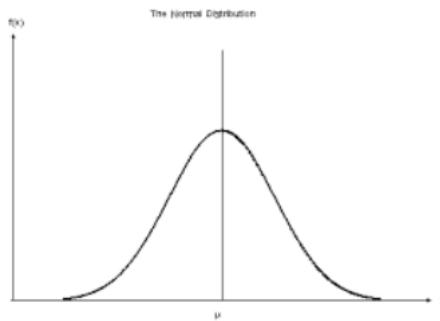
Overvej hvad en meget skæv fordeling for eksempel kunne skyldes.

Der lader til at være bredere konfidensbånd opad. Stigninger i inflation lader altså til at vurderes mere sandsynligt end fald.

Fordelingen er ikke meget skæv. En hvis tendens til at forvente stigning frem for fald i inflationen kunne skyldes, at makrotallene i økonomien peger i den retning. F.eks. arbejdsløshed.

2. Antag nu, at forecast-fordelingen i næste periode, dvs. 4. kvartal 2004, er givet ved $N(\mu, \sigma^2)$, med $\mu = 2,6$ og $\sigma^2 = 0,56$. Udregn den nødvendige information til at forlænge centralbankens fan-chart med én periode

Vi vil nu udregne grænserne for sandsynlighedsintervallerne i en normalfordeling. Vi husker først og fremmest, at normalfordelingen er symmetrisk



så hvis vi vil udregne hvad grænserne er for de 10% tættest på middelværdien, så skal vi have de kritiske værdier for 5% på begge sider af middelværdien.

I STATA kan vi beregne disse kritiske værdier. Vi bruger så at hvis $A \sim N(0, 1)$ og $B \sim N(\mu, \sigma^2)$, så er $B = \mu + \sigma A$.

Værdierne til centralbankens fan-chart én periode frem kan beregnes ved følgende STATA-kode:

```
Uge.47* ×
1  clear all
2
3  input y
4  9
5  8
6  7
7  6
8  5
9  4
10 3
11 2
12 1
13 end
14
15 gen x = y/20
16
17 gen z = -invnormal(x)|
```

```
18
19 gen percent = (1-x^2)*100
20 gen z_low = 2.6-sqrt(0.56)*z
21 gen z_high = 2.6+sqrt(0.56)*z
22
23 br
24
25
```

og de beregnede værdier er:

percent	z_low	z_high
10	2.505964	2.694036
20	2.410412	2.789587
30	2.311653	2.888347
40	2.207575	2.992425
50	2.095258	3.104742
60	1.970188	3.229812
70	1.824404	3.375596
80	1.640975	3.559025
90	1.369104	3.830896

Det ses, at transformationen ”flytter” værdierne med 2,6. Desuden formindsker den mindre varians ($0,56 < 1$) konfidensbåndene.

Maximum Likelihood Estimation og Grænseresultater for stokastiske variable

Assumption 3.1

- i) Den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y|\theta_i) \quad , \quad i = 1, 2, \dots, n,$$

hvor $f_{Y_i}(\cdot|\cdot)$ angiver sandsynlighedsfunktionen for Y_i . y_i er en realisering af Y_i og θ_i er sandsynlighedsfunktionens parameter.

- ii) De stokastiske variable er identisk fordelt således, at

$$\theta_i = \theta_j = \theta \quad \text{og} \quad \forall i, j \quad | \quad f_{Y_i}(y|\theta) = f_{Y_j}(y|\theta)$$

parameteren $\theta = (\theta_1, \dots, \theta_k)' \in \Theta \subseteq \mathbb{R}^k$

- iii) De stokastiske variable Y_i og Y_j er uafhængige for alle $i \neq j$ således, at den simultane sandsynlighed er givet ved

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n|\theta) &= f_{Y_1}(y_1|\theta) \cdot f_{Y_2}(y_2|\theta) \cdots \cdot f_{Y_n}(y_n|\theta) \\ &= \prod_{i=1}^n f_{Y_i}(y_i|\theta) \end{aligned}$$

Theorem C.1 (Store Tals Lov)

Lad $\{Y_1, Y_2, \dots\}$ være en følge af ukorrelerede stokastiske variable med identisk middelværdier og varianser, $E[Y_i] = \mu$ og $V[Y_i] < \infty$.

Da gælder følgende for ethvert $\delta > 0$:

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n Y_i - \mu \right| \geq \delta \right) \rightarrow 0 \quad \text{for } n \rightarrow \infty$$

Theorem C.4 (Den Centrale Grænseværdi sætning)

Lad $\{Y_1, Y_2, \dots\}$ være iid stokastiske variable, $E[Y_i] = \mu$ og $V[Y_i] < \infty$.

\bar{Y} er gennemsnittet for Y .

Da gælder følgende for $U_n = \sqrt{n} \frac{\bar{Y}_n - \mu}{\sigma}$:

$$P(U_n \leq u) \rightarrow \Phi(u) \quad \text{for } u \in \mathbb{R}$$

Opgave 1

Lad y_i , $i = 1, 2, 3, \dots$ være en følge af (i princippet uendeligt mange) tal trukket fra en normalfordeling,

$$N(5, 1)$$

så $\mu = 5$ og $\sigma^2 = 1$. Betragt gennemsnittet af de første n observationer, dvs.

$$m_y = \frac{1}{n} \sum_{i=1}^n y_i.$$

1. Brug STATA til at trække 50.000 tal fra normalfordelingen $N(5, 1)$. Man kan gøre det med følgende kode

```
set obs 50000
set seed 3500
gen y = 5 + rnormal();
```

2. Brug beskrivende statistik til at vise den empiriske fordeling af $\{y_i\}_{i=1}^{50000}$.

Den teoretiske fordeling for $y \sim N(5, 1)$ er

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - 5)^2}{2}\right)$$

Teoretiske momenter og fraktiler for denne fordeling er:

$$\begin{aligned} Mean &= 5 \\ Variance &= 1 \\ Skewness &= 0 \\ Kurtosis &= 3 \\ 5\% - fraktil &= \Phi^{-1}(0, 05) + 5 = 3, 355146 \\ Median &= \Phi^{-1}(0, 50) + 5 = 5 \\ 95\% - fraktil &= \Phi^{-1}(0, 95) + 5 = 6, 644854 \end{aligned}$$

$\Phi(\cdot)$ betegner fordelingsfunktionen for standard normalfordelingen.

Den empiriske fordeleing beregnet i STATA er:

Y			
	Percentiles	Smallest	
1%	2.655802	.8189617	
5%	3.353903	.875423	
10%	3.713241	.9589454	Obs 50,000
25%	4.325881	.9656613	Sum of Wgt. 50,000
50%	4.99811		Mean 4.998664
		Largest	Std. Dev. 1.002523
75%	5.671979	9.125916	
90%	6.276949	9.185016	Variance 1.005053
95%	6.646844	9.332473	Skewness .0008155
99%	7.346997	9.512827	Kurtosis 3.029749

3. Udregn gennemsnit af de første $n = 10$ observationer, dernæst de første $n = 100$, så $n = 500$, $n = 1000$, $n = 5000$, $n = 10000$, og endelig $n = 50000$.

Det kan gøres ved at bruge at `_n` i STATA refererer til observationsnummeret: `mean y if _n<=100` vil således give gennemsnit for de første 100 observationer.

Beskriv hvad der sker med det empiriske gennemsnit som funktion af n . Hvad er det udtryk for?

n	Gennemsnit
10	4,827469
100	4,915142
500	4,967967
1000	5,011143
5000	5,005519
10000	5,007435
50000	4,998664

Vi ser, at gennemsnittet konvergerer (ikke helt jævnt) til den teoretiske værdi og er tættest på denne ved $n = 50000$.

Opgave 2

Vi betragter nu en diskret variabel $\{y\}_{i=1}^n$ der karakteriserer befolkningens trafikvaner. Vi ser på tre mulige udfald og nedenstående tabel viser de absolutte frekvenser for de tre udfald i en tilfældig stikprøve af $n = 573$ personer.

	Gruppe	Antal
$y = 1$	Familien har ikke bil.	121
$y = 2$	Familien ejer en bil.	251
$y = 3$	Familien ejer to eller flere biler.	201
	I alt	573

Vi lader $s_j = \sum_{i=1}^n \mathbb{I}(y_i = j)$ betegne de absolutte frekvenser, $j = 1, 2, 3$ sådan, at $n = s_1 + s_2 + s_3$.

Til opstilling af den statistiske model lader vi y_i repræsentere af den stokastiske variabel Y_i og antager, at

$$P(Y_i = y) = \begin{cases} p_1 & \text{hvis } y = 1 \\ p_2 & \text{hvis } y = 2 \\ p_3 & \text{hvis } y = 3 \end{cases}$$

hvor det gælder, at $p_1 + p_2 + p_3 = 1$.

- Vis, at sandsynlighedsfunktionen, $P(Y_i = y)$, kan skrives som

$$f_{Y_i}(y|p_1, p_2, p_3) = p_1^{\mathbb{I}(y=1)} p_2^{\mathbb{I}(y=2)} p_3^{\mathbb{I}(y=3)}$$

hvor $\mathbb{I}(\cdot)$ er indikatorfunktionen.

Dette ses ved at betragte tilfældet $y = 3$:

$$f_{Y_i}(3|p_1, p_2, p_3) = p_1^0 \cdot p_2^0 \cdot p_3^1 = p_3$$

- Opskriv sample likelihood bidraget, $\ell(p_1, p_2, p_3|y_i)$, og vis at sample likelihood funktionen kan skrives som

$$L(p_1, p_2, p_3|y_1, \dots, y_n) = \prod_{i=1}^n \ell(p_1, p_2, p_3|y_i) = p_1^{s_1} p_2^{s_2} p_3^{s_3}$$

Sample likelihood bidraget (bidraget til den samlede likelihood fra hver marginal hændelse):

$$\ell(p_1, p_2, p_3|y_i) = f_{Y_i}(y|p_1, p_2, p_3) = p_1^{\mathbb{I}(y=1)} p_2^{\mathbb{I}(y=2)} p_3^{\mathbb{I}(y=3)}$$

Da vi har antaget, at alle udfald er uafhængige kan vi opstille den simultane sandsynlighed for hændelserne (den samlede likelihood) som produktet af de marginale hændelser:

$$\begin{aligned} L(p_1, p_2, p_3|y_1, \dots, y_n) &= \prod_{i=1}^n \ell(p_1, p_2, p_3|y_i) \\ &= \prod_{i=1}^n p_1^{\mathbb{I}(y_i=1)} p_2^{\mathbb{I}(y_i=2)} p_3^{\mathbb{I}(y_i=3)} \\ &= p_1^{121} p_2^{251} p_3^{201} \\ &= p_1^{s_1} p_2^{s_2} p_3^{s_3} \end{aligned}$$

3. Når nu sandsynlighederne p_1 , p_2 og p_3 summerer til 1, hvor mange frie parametre er der så i modellen?

$p_1, p_2 \in]0, 1[\times]0, 1[$ kan vælges frit, men det skal gælde, at $p_3 = 1 - p_1 - p_2$. Der er derfor to frie parametre i modellen.

Saml de frie parametre i en vektor θ og angiv parameter-området Θ , så $\theta \in \Theta$.

$$\theta = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \in \Theta = \{\theta \in \mathbb{R}_+^2 : p_1 + p_2 < 1\}$$

Opskriv likelihood funktionen som funktion af de frie parametre, θ

Vi udskifter p_3 med $1 - p_1 - p_2$ og s_3 med $n - s_1 - s_2$ i likelihood funtionen:

$$L(p_1, p_2 | y_1, \dots, y_n) = \prod_{i=1}^n \ell(p_1, p_2 | y_i) = p_1^{s_1} \cdot p_2^{s_2} \cdot (1 - p_1 - p_2)^{n - s_1 - s_2}$$

Opskriv også log-likelihood funktionen.

$$\begin{aligned} \log L(p_1, p_2 | y_1, \dots, y_n) &= \log \prod_{i=1}^n \ell(p_1, p_2 | y_i) \\ &= s_1 \log(p_1) + s_2 \log(p_2) + (n - s_1 - s_2) \log(1 - p_1 - p_2) \end{aligned}$$

4. Angiv præcis de antagelser du har brugt til at opskrive den statistiske model og forklar hvorfor de var vigtige

For at kunne opskrive den statistiske model har vi antaget at sandsynlighederne for at passe i én af kategorierne er den samme for alle 573 tilfældigt udvalgte personer (*identisk fordeling*). Dette er vigtigt fordi vi dermed kan antage, at de samme sandsynlighedsparametre gælder for hver marginal hændelse.

Det er desuden antaget at disse sandsynligheder er *uafhængige*. Sandsynligheden for, at en given person passer i en kategori er ikke influeret af, hvilke kategorier de andre adspurgte personer passer i. Dette er vigtigt da vi med uafhængighed kan udtrykke sandsynligheden for den samlede hændelse som produktet af de marginale sandsynligheder.

Kan de antages at være rimelige i dette tilfælde?

Forestiller man sig 573 personer valgt fra et afgrænset geografisk område er antagelsen om identisk fordeling ikke urimelig. Derimod er antagelsen om uafhængighed i dette tilfælde problematisk.

Er personerne derimod udvalgt fra et stort geografisk område, er antagelsen om uafhængighed plausibel. Men antagelsen om identisk fordeling er problematisk.

5. Find maksimum likelihood estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$.

Score-vektoren findes ved at differentiere likelihood-funktionen:

$$\begin{aligned}\frac{\partial \ell(p_1, p_2 | Y_i)}{\partial \theta} &= \left(\frac{\frac{\partial \ell(p_1, p_2 | Y_i)}{\partial p_1}}{\frac{\partial \ell(p_1, p_2 | Y_i)}{\partial p_2}} \right) \\ &= \left(\frac{\frac{s_1}{p_1} - \frac{n-s_1-s_2}{1-p_1-p_2}}{\frac{s_2}{p_2} - \frac{n-s_1-s_2}{1-p_1-p_2}} \right)\end{aligned}$$

Førsteordensbetingelserne:

$$\begin{aligned}\frac{s_1}{\hat{p}_1} - \frac{n-s_1-s_2}{1-\hat{p}_1-\hat{p}_2} &= 0 \quad \wedge \quad \frac{s_2}{\hat{p}_2} - \frac{n-s_1-s_2}{1-\hat{p}_1-\hat{p}_2} = 0 \Leftrightarrow \\ \frac{s_1}{\hat{p}_1} &= \frac{s_2}{\hat{p}_2} \Leftrightarrow \\ \hat{p}_1 &= \hat{p}_2 \frac{s_1}{s_2} \Leftrightarrow \\ \frac{s_2}{\hat{p}_2} &= \frac{n-s_1-s_2}{1-\hat{p}_2 \frac{s_1}{s_2} - \hat{p}_2} \Leftrightarrow \\ s_2 - \hat{p}_2 s_1 - \hat{p}_2 s_2 &= \hat{p}_2(n - s_1 - s_2) \Leftrightarrow \\ \hat{p}_2(n - s_1 - s_2 + s_1 + s_2) &= s_2 \Leftrightarrow \\ \hat{p}_2 &= \frac{s_2}{\underline{\underline{n}}} \Leftrightarrow \\ \hat{p}_1 &= \frac{s_2}{n} \frac{s_1}{s_2} \\ &= \frac{s_1}{\underline{\underline{n}}}\end{aligned}$$

Brug informationen i tabellen ovenfor til også at finde estimatet, $\hat{\theta}(y_1, \dots, y_n)$.
Kommentér på resultaterne.

Med resultaterne fra tabellen bliver sandsynlighederne:

$$\begin{aligned}\hat{p}_1 &= \frac{121}{573} \approx \underline{\underline{0,21}} \\ \hat{p}_2 &= \frac{251}{573} \approx \underline{\underline{0,44}} \\ \hat{p}_3 &= 1 - p_1 - p_2 \approx \underline{\underline{0,35}}\end{aligned}$$

Hvis antagelserne om uafhængighed og identiske fordelinger i modellen holder er sandsynlighed for at en tilfældigt udvalgt familie har bil altså ca. 80%. Blandt familier med bil har de fleste kun én bil.

Opgave 3

Betragt en maskine der producerer reb. Hvert minut produceres der i gennemsnit en reblængde svarende til 2 meter med en standardafvigelse på 10cm. Antag at rebproduktionen i forskellige minutter er identisk og uafhængigt fordelt. Lad rebproduktionen i minut i være repræsenteret ved den stokastiske variabel Y_i , $i = 1, 2, \dots, n$.

Den gennemsnitlige reblængde som maskinen har produceret på n minutter er givet ved

$$X_n = \frac{1}{n} \sum_{i=1}^n Y_i,$$

så den samlede reblængde er

$$nX_n = \sum_{i=1}^n Y_i.$$

- Brug den centrale grænseværdidisætning til at karakterisere den approximative fordeling af den gennemsnitlige reblængde, X_n når n bliver stor.

Vi gentager *Theorem C.4* om den centrale grænse værdidisætning:

Theorem C.4 (Den Centrale Grænseværdi sætning)

Lad $\{Y_1, Y_2, \dots\}$ være iid stokastiske variable, $E[Y_i] = \mu$ og $V[Y_i] < \infty$.

Da gælder følgende for $U_n = \sqrt{n} \frac{\bar{Y}_n - \mu}{\sigma}$:

$$P(U_n \leq u) \rightarrow \Phi(u) \quad \text{for } u \in \mathbb{R}$$

Den empiriske middelværdi for vores stokastiske variabel, rebproduktion pr. minut er lig med det beregnede gennemsnit

$$X_n = \frac{1}{n} \sum_{i=1}^n Y_i$$

og vi ved fra *Store Tals Lov*, at

$$\frac{1}{n} \sum_{i=1}^n Y_i \rightarrow E[Y_i] \quad \text{for } n \rightarrow \infty$$

Da alle Y_i er identisk fordelt (samme varians) og uafhængige (kovarians lig 0) er den empriske varians

$$\begin{aligned} V(X_n) &= V\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n V(Y_i) \\ &= \frac{1}{n} V(Y_i) \\ &= \frac{0,10^2}{n} \end{aligned}$$

Derfor får vi ifølge den centrale grænseværdidisætning at

$$X_n \sim N\left(E(Y_i), \frac{V(Y_i)}{n}\right)$$

og med de givne oplysninger

$$X_n \sim N\left(2, \frac{0,10^2}{n}\right)$$

2. Brug den asymptotiske approksimation til at udregne sandsynligheden for, at der på en time produceres mindst 125 meter

Vi ved at den gennemsnitlige produktion pr. minut i 60 minutter kan approksimeres med

$$X_{60} \sim N\left(2; \frac{0,10^2}{60}\right)$$

hvilket giver følgende fordeling efter 60 minutter:

$$60 \times X_{60} \sim N\left(2 \times 60; 60^2 \frac{0,10^2}{60}\right) = N(120; 0,6)$$

Hvis vi gerne vil udtrykke fordelingen af time-produktionen ved *standard-normalfordelingen* kan vi transformere $60X_{60}$:

$$\frac{60X_{60} - 120}{\sqrt{0,6}} = U_n \sim N(0, 1)$$

På baggrund af fordelingen kan sandsynligheden for, at der produceres mindst 125 meter udregnes:

$$\begin{aligned} P(60X_{60} > 125) &= 1 - P(60X_{60} < 125) \\ &= 1 - \Phi\left(\frac{125 - 120}{\sqrt{0,6}}\right) \\ &\approx \underline{\underline{0}} \end{aligned}$$

3. Ville den centrale grænseværdidisætning også gælde, hvis fordelingen af maskinens rebproduktion pr. minut Y_i var beskrevet ved en Cauchy fordeling?

Vi kigger lige på definitionen igen:

Theorem C.4 (Den Centrale Grænseværdi sætning)

Lad $\{Y_1, Y_2, \dots\}$ være iid stokastiske variable, $E[Y_i] = \mu$ og $V[Y_i] < \infty$.

Da gælder følgende for $U_n = \sqrt{n} \frac{\bar{Y}_n - \mu}{\sigma}$:

$$P(U_n \leq u) \rightarrow \Phi(u) \quad \text{for } u \in \mathbb{R}$$

Læg mærke til, at en af forudsætningerne er, at de første to momenter, middelværdi og varians skal være defineret.

[Wikipedia-artikel om Cauchy-fordeling](#)

Da et gennemsnit af n Cauchy-fordelinger har samme fordeling som den enkelte fordelinger er der ingen konvergens i fordeling, og Den Centrale Grænseværdi sætning gælder derfor ikke.

Opgave 4

Lad $\{Y_i\}_{i=1}^n$ være n identiske og uafhængige normalfordelte stokastiske variable,

$$Y_i \stackrel{d}{=} N(\mu, \sigma^2),$$

med $\mu \in \mathbb{R}$ og $\sigma^2 > 0$ så tæthedsfunktionen er givet ved

$$f_{Y_i}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

- Opskriv likelihood funktionen $L(\mu, \sigma^2 | Y_1, \dots, Y_n)$.

Da der er givet en tæthedsfunktion og vi har *iid*-antagelser kan vi opstille likelihood funktionen som produkt af de identisk fordelte likelihood bidrag:

$$\begin{aligned} L(\mu, \sigma^2 | Y_1, \dots, Y_n) &= \prod_{i=1}^n \ell(\mu, \sigma^2 | Y_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i-\mu)^2}{2\sigma^2}\right) \\ &\quad \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \theta \in \Theta = \{\theta \in \mathbb{R}^2 : \sigma^2 > 0\} \end{aligned}$$

Opskriv også log-likelihood funktionen, $\log L(\mu, \sigma^2 | Y_1, \dots, Y_n)$ og find maximum likelihood estimatorne, $\hat{\mu}(Y_1, \dots, Y_n)$ og $\hat{\sigma}^2(Y_1, \dots, Y_n)$.

Log-likelihood funktionen:

$$\begin{aligned} \log L(\mu, \sigma^2 | Y_1, \dots, Y_n) &= \log \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_i-\mu)^2}{2\sigma^2}\right) \\ &= \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (Y_i - \mu)^2 \right) \end{aligned}$$

Den samlede score fås som summen af score-bidragene:

$$\begin{aligned} S(\mu, \sigma^2) &= \left(\sum_{i=1}^n s_i(\mu) \right) \\ &= \left(\sum_{i=1}^n \frac{\partial \log \ell(\mu, \sigma^2 | Y_i)}{\partial \mu} \right) \\ &= \left(\sum_{i=1}^n \left(\frac{\sum_{i=1}^n \frac{Y_i - \mu}{\sigma^2}}{\sum_{i=1}^n \left(\frac{(Y_i - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)} \right) \right) \\ &= \left(\sum_{i=1}^n \frac{\sum_{i=1}^n \frac{Y_i - \mu}{\sigma^2}}{\sum_{i=1}^n \frac{(Y_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2}} \right) \end{aligned}$$

Vi løser først førsteordensbetingelsen for μ :

$$\begin{aligned} \sum_{i=1}^n \frac{Y_i - \hat{\mu}}{\sigma^2} = 0 &\Leftrightarrow \\ \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{\mu} &\Leftrightarrow \\ \sum_{i=1}^n Y_i = n\hat{\mu} &\Leftrightarrow \\ \hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i & \end{aligned}$$

Den optimale estimator for middelværdien er lig med gennemsnittet af observationerne, eller *den empiriske middelværdi*.

Vi indsætter det udledte udtryk for $\hat{\mu}$ i førsteordensbetingelsen for σ^2 :

$$\begin{aligned} \sum_{i=1}^n \frac{(Y_i - \hat{\mu})^2}{2\hat{\sigma}^4} - \frac{n}{2\hat{\sigma}^2} = 0 &\Leftrightarrow \\ \sum_{i=1}^n \frac{(Y_i - \frac{1}{n} \sum_{i=1}^n Y_i)^2}{2\hat{\sigma}^4} = \frac{n}{2\hat{\sigma}^2} &\Leftrightarrow \\ \sum_{i=1}^n (Y_i - \frac{1}{n} \sum_{i=1}^n Y_i)^2 = n \frac{\hat{\sigma}^4}{\hat{\sigma}^2} &\Leftrightarrow \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \frac{1}{n} \sum_{i=1}^n Y_i)^2 & \end{aligned}$$

Den optimale estimator for variansen er gennemsnittet af de kvadrerede afvigelser fra gennemsnittet, eller *den empiriske varians*.

2. I et konkret tilfælde med observationer $\{y_i\}_{i=1}^{125}$, finder vi, at $\sum_{i=1}^{125} y_i = 627,60$ og $\sum_{i=1}^{125} y_i^2 = 3807,2$. Brug dette til også at finde maksimum likelihood estimaterne, $\hat{\mu}(y_1, \dots, y_n)$ og $\hat{\sigma}^2(y_1, \dots, y_n)$.

Vi indsætter de observerede værdier i de udledte udtryk for $\hat{\mu}$ og $\hat{\sigma}^2$:

$$\begin{aligned} \hat{\sigma}^2(y_1, \dots, y_{125}) &= \frac{1}{125} \sum_{i=1}^{125} (y_i - \frac{1}{125} \sum_{i=1}^{125} y_i)^2 \\ \hat{\mu}(y_1, \dots, y_{125}) &= \frac{1}{125} \sum_{i=1}^{125} y_i \\ &= \frac{627,60}{125} \\ &\approx \underline{\underline{5,02}} \\ &= \frac{3807,2}{125} + \frac{1}{125} \sum_{i=1}^{125} 5,02^2 - \frac{1}{125} \sum_{i=1}^{125} 2 \times 5,02^2 \\ &= 30,46 + 25,21 - 50,42 \\ &= \underline{\underline{5,25}} \end{aligned}$$

3. Tilfældet, hvor middelværdi og varians er den samme, $\mu = \sigma^2 = \phi$, kaldes equi-dispersion, og er blandt andet kendt fra Poisson-fordelingen.

Vi vil analysere normalfordelings-modellen under antagelse af equi-dispersion, dvs. hvor tæthedsfunktionen er givet ved

$$f_{Y_i}(y|\phi) = \frac{1}{\sqrt{2\pi\phi}} \exp\left(\frac{-(y-\phi)^2}{2\phi}\right),$$

hvor $\phi > 0$ nu er en ukendt parameter der skal estimeres

Skriv igen likelihood funktionen, $L(\phi|Y_1, \dots, Y_n)$ og log-likelihood funktionen, $\log L(\phi|Y_1, \dots, Y_n)$.

Likelihood funktionen:

$$L(\phi|Y_1, \dots, Y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\phi}} \exp\left(\frac{-(Y_i-\phi)^2}{2\phi}\right) \quad \phi \in \mathbb{R}_+$$

Log-likelihood funktionen:

$$\begin{aligned} \log L(\phi|Y_1, \dots, Y_n) &= \log \prod_{i=1}^n \frac{1}{\sqrt{2\pi\phi}} \exp\left(\frac{-(Y_i-\phi)^2}{2\phi}\right) \\ &= \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\phi) - \frac{1}{2\phi} (Y_i - \phi)^2 \right) \end{aligned}$$

4. Vis at scoren er givet ved

$$\frac{\sum_{i=1}^n Y_i^2 - n\phi^2 - n\phi}{2\phi^2}$$

Score-bidrag:

$$\begin{aligned} s_i(\phi) &= \frac{\partial \ell(\phi|Y_i)}{\partial \phi} \\ &= -\frac{1}{2\phi} - \frac{-2(Y_i - \phi)2\phi - 2(Y_i - \phi)^2}{4\phi^2} \\ &= \frac{2(Y_i - \phi)2\phi + 2(Y_i - \phi)^2 - 2\phi}{4\phi^2} \\ &= \frac{4\phi Y_i - 4\phi^2 + 2Y_i^2 + 2\phi^2 - 4\phi Y_i - 2\phi}{4\phi^2} \\ &= \frac{Y_i^2 - \phi^2 - \phi}{2\phi^2} \end{aligned}$$

Samlet score:

$$\begin{aligned} S(\phi) &= \sum_{i=1}^n s_i(\phi) \\ &= \sum_{i=1}^n \frac{Y_i^2 - \phi^2 - \phi}{2\phi^2} \\ &= \frac{\sum_{i=1}^n Y_i^2 - \sum_{i=1}^n \phi^2 - \sum_{i=1}^n \phi}{2\phi^2} \\ &= \frac{\sum_{i=1}^n Y_i^2 - n\phi^2 - n\phi}{2\phi^2} \end{aligned}$$

5. Find estimatoren, $\hat{\phi}(Y_1, \dots, Y_n)$, og sammenlign med estimatorerne ovenfor, $\hat{\mu}(y_1, \dots, y_n)$ og $\hat{\sigma}^2(y_1, \dots, y_n)$

Vi finder $\hat{\phi}(Y_1, \dots, Y_n)$ ved at løse førsteordensbetingelsen

$$\begin{aligned}\frac{\sum_{i=1}^n Y_i^2 - n\hat{\phi}^2 - n\hat{\phi}}{2\hat{\phi}^2} &= 0 \Leftrightarrow \\ \sum_{i=1}^n Y_i^2 - n\hat{\phi}^2 - n\hat{\phi} &= 0 \Leftrightarrow \\ \hat{\phi}^2 + \hat{\phi} - \frac{1}{n} \sum_{i=1}^n Y_i^2 &= 0 \Leftrightarrow \\ \hat{\phi} &= \frac{-1 \pm \sqrt{1 + \frac{4}{n} \sum_{i=1}^n Y_i^2}}{2}\end{aligned}$$

Da $\phi > 0$ er $\hat{\phi}$ én tydigt bestemt:

$$\hat{\phi} = \frac{-1 + \sqrt{1 + 4 \frac{1}{n} \sum_{i=1}^n Y_i^2}}{2}$$

6. Find også estimatet, $\hat{\phi}(y_1, \dots, y_n)$, og sammenlign med ovenfor.

Vi indsætter de observerede værdier:

$$\begin{aligned}\hat{\phi} &= \frac{-1 + \sqrt{1 + 4 \frac{1}{125} \sum_{i=1}^{125} y_i^2}}{2} \\ &= \frac{-1 + \sqrt{1 + 4 \frac{3807,2}{125}}}{2} \\ &= \frac{-1 + 11,08}{2} \\ &\approx \underline{\underline{5,04}}\end{aligned}$$

Vi får et estimat af ϕ som ligger mellem estimererne for μ og σ^2 .

Opgave 5

Lad de stokastiske variable, $\{Y\}_{i=1}^n$, være uafhængigt og identisk fordelt med en gamma-fordeling

$$Y_i \stackrel{d}{=} \text{Gamma}(\alpha, \beta), \quad Y_i \in \mathbb{Y} = \{y \in \mathbb{R} | y > 0\}.$$

Tæthedsfunktionen er givet ved

$$f_{Y_i}(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad \alpha > 0, \quad \beta > 0$$

hvor $\Gamma(\cdot)$ er en kompliceret funktion kendt som gamma-funktionen. Det gælder for gamma-fordelingen at $E(Y_i) = \frac{\alpha}{\beta}$.

1. Skriv likelihood funktionen og log-likelihood funktionen.

Likelihood funktionen:

$$L(\alpha, \beta | Y_1, \dots, Y_n) = \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} Y_i^{\alpha-1} \exp(-\beta Y_i)$$

$$\binom{\alpha}{\beta} = \theta \in \Theta = \{\theta \in \mathbb{R}^2 : \alpha > 0, \beta > 0\}$$

Log-likelihood funktionen:

$$\begin{aligned} \log L(\alpha, \beta | Y_1, \dots, Y_n) &= \log \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} Y_i^{\alpha-1} \exp(-\beta Y_i) \\ &= \sum_{i=1}^n \left(\alpha \log(\beta) - \log(\Gamma(\alpha)) + (\alpha - 1) \log(Y_i) - \beta Y_i \right) \\ &= n\alpha \log(\beta) - n \log(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^n \log(Y_i) - \beta \sum_{i=1}^n Y_i \end{aligned}$$

2. Vis, at førsteordensbetingelserne for maksimum af log-likelihood funktionen kan skrives som

$$0 = n \frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n y_i \tag{1}$$

$$0 = n \log(\hat{\beta}) - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(Y_i) \tag{2}$$

hvor den logaritmisk afledte, $\psi(\alpha) = \frac{\partial}{\partial \alpha} \log \Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$, er kendt som di-gamma funktionen.

Samlet score:

$$\begin{aligned} S(\alpha, \beta) &= \left(\frac{\frac{\partial \log L(\alpha, \beta | Y_1, \dots, Y_n)}{\partial \alpha}}{\frac{\partial \log L(\alpha, \beta | Y_1, \dots, Y_n)}{\partial \beta}} \right) \\ &= \left(n \log(\beta) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log(Y_i) \atop n \frac{\alpha}{\beta} - \sum_{i=1}^n (Y_i) \right) \end{aligned}$$

Førsteordensbetingelserne for maximum likelihood estimerterne:

$$\begin{aligned} 0 &= n \log(\hat{\beta}) - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(Y_i) \\ 0 &= n \frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n (Y_i) \end{aligned}$$

3. Løs (1) for $\frac{\hat{\alpha}}{\hat{\beta}}$. Virker estimatoren rimelig?

$$0 = n \frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n Y_i \Leftrightarrow$$

$$\frac{\hat{\alpha}}{\hat{\beta}} = \frac{1}{n} \sum_{i=1}^n Y_i$$

Estimatoren for $\frac{\alpha}{\beta}$ er lig med den empiriske middelværdi hvilket virker logisk, da $E(Y_i) = \frac{\alpha}{\beta}$ i en gamma-fordeling.

4. Vis ved indsættelse at ligning 2 kan skrives som

$$0 = n \left(\log(n\hat{\alpha}) - \log \left(\sum_{i=1}^n Y_i \right) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \right) + \sum_{i=1}^n \log(Y_i).$$

Bemærk at (1) kan skrives om på følgende måde:

$$0 = n \frac{\hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n Y_i \Leftrightarrow$$

$$\hat{\beta} = n\hat{\alpha} \left(\sum_{i=1}^n Y_i \right)^{-1}$$

Dette indsættes i (2):

$$0 = n \log(\hat{\beta}) - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(Y_i)$$

$$= n \log \left(n\hat{\alpha} \left(\sum_{i=1}^n Y_i \right)^{-1} \right) - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(Y_i)$$

$$= n \left(\log(n\hat{\alpha}) - \log \left(\sum_{i=1}^n Y_i \right) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} \right) + \sum_{i=1}^n \log(Y_i)$$

Denne ligning har ingen analytisk løsning, men kan løses numerisk.

5. Filen gamma.xls indeholder $n = 250$ observationer fra en gamma-fordeling
Brug beskrivende statistik til at beskrive data.

	Percentiles	Smallest		
1%	1.0608	.66647		
5%	1.9581	.99459		
10%	2.46845	1.0608	Obs	250
25%	3.3655	1.0777	Sum of Wgt.	250
50%	4.47325		Mean	4.81098
		Largest	Std. Dev.	2.120183
75%	5.9675	11.044	Variance	4.495175
90%	7.6491	11.352	Skewness	.9009869
95%	8.4153	12.028	Kurtosis	4.0494
99%	11.352	12.394		

6. Vi får at vide, at estimatet for α er $\hat{\alpha} = 5,09981$. Find estimatet for β .

Vi ved, at middelværdien i gamma-fordelingen er

$$E(Y_i) = \frac{\alpha}{\beta}$$

I den beskrivende statistik ovenfor har vi et konsistent estimat for $E(Y_i)$, nemlig $\frac{1}{n} \sum_{i=1}^n y_i$. Givet et estimat for α , $\hat{\alpha} = 5,09981$ og et estimat for middelværdien kan vi udlede et estimat for β :

$$\begin{aligned}\hat{\beta} &= \frac{\hat{\alpha}}{\frac{1}{n} \sum_{i=1}^n y_i} \\ &= \frac{5,09981}{4,81098} \\ &= \underline{\underline{1,06004}}\end{aligned}$$

Opgave 6

Betrægt igen datasættet med jeres spørgeskemasvar, givet i filen `sand_spskema2019.dta`. Vi vil nu se på rygning og brugen af cykelhjelm. Konstruer variablen:

$$y_i = \text{smoker}_i = \begin{cases} 1 & \text{hvis person } i \text{ mindst ryger til fester} \\ 0 & \text{ellers} \end{cases}$$

Lad Y_i være en stokastisk variabel svarende til observationerne $y_i = \text{hjelm}_i$. Antag at alle har samme sandsynlighed for at bruge hjelm, θ , $0 < \theta < 1$, og at observationerne er uafhængige.

1. Foreslå en statistisk model for datasættet $\{y_i\}_{i=1}^n$.

Angiv de antagelser du anvender og diskutér om de forekommer rimelige.

Da data er binære ereEn passende statistisk model for datasættet $\{y_i\}_{i=1}^n$ er en *Bernoulli*-fordeling

$$Y_i \stackrel{d}{=} \text{Bernoulli}(\theta_i)$$

. Under givne antagelser kan vi estimere parameteren θ_i i denne Bernoulli-fordeling ved Maximum Likelihood Estimation.

Antagelserne for at opstille en likelihood-model er:

- Den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y|\theta_i) = \theta_i^y (1 - \theta_i)^{1-y}$$

- Alle stokastiske variable er identisk fordelte

$$f_{Y_i}(y|\theta_i) = f_{Y_j}(y|\theta_j) \Leftrightarrow \theta_i = \theta_j = \theta$$

- Alle stokastiske variable er uafhængige

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n | \theta) = \prod_{i=1}^n f_{Y_i}(y_i | \theta)$$

Disse antagelser kan nemt problematiseres da der må forventes forskelle mellem eks. køn (brud på antagelsen om identisk fordeling), samt at brugen af cykelhjelm påvirkes af om andre bruger cykelhjelm (brud på antagelsen om uafhængighed).

2. Opskriv likelihood funktionen og log-likelihood funktionen

Likelihood funktion:

$$L(\theta | Y_1, \dots, Y_n) = \prod_{i=1}^n \ell(\theta | Y_i) = \prod_{i=1}^n \theta^{Y_i} (1 - \theta)^{1-Y_i}$$

$$\theta \in \Theta = \{\theta \in \mathbb{R} : 0 < \theta < 1\}$$

Log-likelihood funktion:

$$\log L(\theta | Y_1, \dots, Y_n) = \log \prod_{i=1}^n \ell(\theta | Y_i) = \sum_{i=1}^n \log \ell(\theta | Y_i) = \sum_{i=1}^n \left(Y_i \log(\theta) + (1 - Y_i) \log(1 - \theta) \right)$$

3. Find maximum likelihood estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$.

Estimatoren $\hat{\theta}(Y_1, \dots, Y_n)$ løser førsteordensbetingelsen for $\log L(\theta|Y_1, \dots, Y_n)$:

$$\begin{aligned} S(\hat{\theta}) &= \frac{\partial \log L(\theta|Y_1, \dots, Y_n)}{\partial \hat{\theta}} = \sum_{i=1}^n \frac{\partial \ell(\hat{\theta}|Y_i)}{\partial \hat{\theta}} = 0 \Leftrightarrow \\ &\sum_{i=1}^n \left(\frac{Y_i}{\hat{\theta}} - \frac{1 - Y_i}{1 - \hat{\theta}} \right) = 0 \Leftrightarrow \\ &\frac{1}{\hat{\theta}} \sum_{i=1}^n Y_i = \frac{1}{1 - \hat{\theta}} \sum_{i=1}^n (1 - Y_i) \Leftrightarrow \\ &\frac{1 - \hat{\theta}}{\hat{\theta}} = \frac{n - \sum_{i=1}^n Y_i}{\sum_{i=1}^n (1 - Y_i)} \Leftrightarrow \\ &\frac{1}{\hat{\theta}} - 1 = \frac{n}{\sum_{i=1}^n (1 - Y_i)} - 1 \Leftrightarrow \\ &\hat{\theta} = \underline{\underline{\frac{\sum_{i=1}^n Y_i}{n}}} \end{aligned}$$

4. Angiv den sufficiente statistik

Udregn den sufficiente statistik for det observerede datasæt og brug det til at finde maximum likelihood estimatet, $\hat{\theta}(y_1, \dots, y_n)$.

Den sufficiente statistik er $\sum_{i=1}^n y_i$ da udledningen af $\hat{\theta}$ kun afhænger heraf.

I datasættet er $\sum_{i=1}^{126} y_i = 31$:

$$\hat{\theta}(y_1, \dots, y_n) = \frac{31}{126} \approx \underline{\underline{0,2460}}$$

5. Vi indrager nu også variablen for brug af cykelhjelm: Generér variablen:

$$x_i = hjelm_i = \begin{cases} 1 & \text{hvis cykelhjem} \in \{1, 2\} \text{ for person } i \\ 0 & \text{ellers,} \end{cases}$$

og vi er interesserede i en betinget model der kan bruges til at undersøge sandsynligheden for at ryge for folk der buger hjelm mod folk, der ikke bruger cykelhjelm, dvs. en model for $Y_i|X_i = x_i$, hvor

$$P(Y_i = 1) = \begin{cases} \theta_1 & \text{hvis person } i \text{ bruger hjelm} \\ \theta_2 & \text{hvis person } i \text{ ej bruger hjelm.} \end{cases}$$

Foreslå den betingede model, dvs. likelihood funktionen og parameterområdet Θ , så $\theta = (\theta_1, \theta_2)' \in \Theta$.

Find estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$, og estimatet $\hat{\theta}(y_1, \dots, y_n)$. Fortolk dine resultater.

Den betingede model indeholder to parametre, θ_1 hvis personen bruger hjelm ($x_i = 0$), og θ_2 hvis personen ikke bruger hjelm ($x_i = 1$).

Dermed kan vi skrive likelihood funktionen som:

$$L(\theta_1, \theta_2|Y_1, \dots, Y_n, X_1, \dots, X_n) = \prod_{i=1}^n \left(\theta_1^{X_i} \theta_2^{1-X_i} \right)^{Y_i} \left((1 - \theta_1)^{X_i} (1 - \theta_2)^{1-X_i} \right)^{1-Y_i}$$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \Theta = \{\theta \in \mathbb{R}^2 : 0 < \theta_i < 1, i = 1, 2\}$$

Log-likelihood funktion:

$$\begin{aligned} \log L(\theta_1, \theta_2 | Y_1, \dots, Y_n, X_1, \dots, X_n) &= \sum_{i=1}^n \log \left(\theta_1^{X_i} \theta_2^{1-X_i} \right)^{Y_i} \left((1-\theta_1)^{X_i} (1-\theta_2)^{1-X_i} \right)^{1-Y_i} \\ &= \sum_{i=1}^n Y_i \log \left(\theta_1^{X_i} \theta_2^{1-X_i} \right) + \sum_{i=1}^n (1-Y_i) \log \left((1-\theta_1)^{X_i} (1-\theta_2)^{1-X_i} \right) \\ &= \sum_{i=1}^n \left(Y_i X_i \log(\theta_1) + Y_i (1-X_i) \log(\theta_2) \right) + \sum_{i=1}^n \left((1-Y_1) X_i \log(1-\theta_1) + (1-Y_i) (1-X_i) \log(1-\theta_2) \right) \\ &= \log(\theta_1) \sum_{i=1}^n Y_i X_i + \log(1-\theta_1) \sum_{i=1}^n (1-Y_i) X_i + \log(\theta_2) \sum_{i=1}^n Y_i (1-X_i) + (1-\theta_2) \sum_{i=1}^n (1-Y_i) (1-X_i) \end{aligned}$$

Samlet score:

$$\begin{aligned} S(\theta) &= \frac{\partial \log L(\theta_1, \theta_2 | Y_1, \dots, Y_n, X_1, \dots, X_n)}{\partial \theta} \\ &= \begin{cases} \frac{\partial \log L(\theta_1, \theta_2 | Y_1, \dots, Y_n, X_1, \dots, X_n)}{\partial \theta_1} \\ \frac{\partial \log L(\theta_1, \theta_2 | Y_1, \dots, Y_n, X_1, \dots, X_n)}{\partial \theta_2} \end{cases} \\ &= \begin{cases} \frac{\sum_{i=1}^n Y_i X_i}{\hat{\theta}_1} - \frac{\sum_{i=1}^n (1-Y_i) X_i}{1-\hat{\theta}_1} \\ \frac{\sum_{i=1}^n Y_i (1-X_i)}{\hat{\theta}_2} - \frac{\sum_{i=1}^n (1-Y_i) (1-X_i)}{1-\hat{\theta}_2} \end{cases} \end{aligned}$$

FOC for θ_1 :

$$\begin{aligned} \frac{\sum_{i=1}^n Y_i X_i}{\hat{\theta}_1} - \frac{\sum_{i=1}^n (1-Y_i) X_i}{1-\hat{\theta}_1} &= 0 \Leftrightarrow \\ \frac{1-\hat{\theta}_1}{\hat{\theta}_1} &= \frac{\sum_{i=1}^n (1-Y_i) X_i}{\sum_{i=1}^n Y_i X_i} \Leftrightarrow \\ \frac{1}{\hat{\theta}_1} - 1 &= \frac{\sum_{i=1}^n (X_i - Y_i X_i)}{\sum_{i=1}^n Y_i X_i} \Leftrightarrow \\ \frac{1}{\hat{\theta}_1} &= \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n Y_i X_i + \sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n Y_i X_i} \Leftrightarrow \\ \hat{\theta}_1 &= \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} \end{aligned}$$

FOC for θ_2 :

$$\begin{aligned}
 \frac{\sum_{i=1}^n Y_i(1-X_i)}{\hat{\theta}_2} - \frac{\sum_{i=1}^n (1-Y_i)(1-X_i)}{1-\hat{\theta}_2} = 0 &\Leftrightarrow \\
 \frac{1-\hat{\theta}_2}{\hat{\theta}_2} &= \frac{\sum_{i=1}^n (1-Y_i)(1-X_i)}{\sum_{i=1}^n Y_i(1-X_i)} \Leftrightarrow \\
 \frac{1}{\hat{\theta}_2} - 1 &= \frac{\sum_{i=1}^n (1-Y_i)(1-X_i)}{\sum_{i=1}^n Y_i(1-X_i)} \Leftrightarrow \\
 \frac{1}{\hat{\theta}_2} &= \frac{\sum_{i=1}^n ((1-Y_i)(1-X_i) + Y_i(1-X_i))}{\sum_{i=1}^n Y_i(1-X_i)} \Leftrightarrow \\
 \frac{1}{\hat{\theta}_2} &= \frac{\sum_{i=1}^n ((1-X_i) - (1-X_i)Y_i + Y_i(1-X_i))}{\sum_{i=1}^n Y_i(1-X_i)} \Leftrightarrow \\
 \frac{1}{\hat{\theta}_2} &= \frac{\sum_{i=1}^n (1-X_i)}{\sum_{i=1}^n Y_i(1-X_i)} \Leftrightarrow \\
 \hat{\theta}_2 &= \frac{\sum_{i=1}^n Y_i(1-X_i)}{\sum_{i=1}^n (1-X_i)}
 \end{aligned}$$

Dermed har vi udledt estimatoren:

$$\hat{\theta}(Y_1, \dots, Y_n) = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} \\ \frac{\sum_{i=1}^n Y_i(1-X_i)}{\sum_{i=1}^n (1-X_i)} \end{pmatrix}$$

og estimatet på baggrund af data bliver:

$$\begin{aligned}
 \hat{\theta}(y_1, \dots, y_{165}) &= \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \\
 &= \begin{pmatrix} \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i} \\ \frac{\sum_{i=1}^n y_i(1-x_i)}{\sum_{i=1}^n (1-x_i)} \end{pmatrix} \\
 &= \begin{pmatrix} \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i} \\ \frac{\sum_{i=1}^n y_i - \sum_{i=1}^n y_i x_i}{n - \sum_{i=1}^n x_i} \end{pmatrix} \\
 &= \begin{pmatrix} \frac{9}{44} \\ \frac{31-9}{126-44} \end{pmatrix} \\
 &= \underline{\underline{\begin{pmatrix} 0,2045 \\ 0,2683 \end{pmatrix}}}.
 \end{aligned}$$

Der lader altså til at være en højere sandsynlighed for at være ryger, blandt personer som ikke benytter cykelhjelm.

6. Man kunne også have parametrizeret modellen som

$$P(Y_i = 1) = \begin{cases} \gamma & \text{hvis person } i \text{ bruger hjelm} \\ \gamma + \delta & \text{hvis person } i \text{ ej bruger hjelm,} \end{cases}$$

så parametrene er givet ved $\theta = (\gamma, \delta)'$. Skriv parameterrummet, Θ , i dette tilfælde, så $\theta \in \Theta$.

Hvad vil fortolkningen af δ være?

Likelihood funktionen bliver da:

$$L(\gamma, \delta | Y_1, \dots, Y_n, X_1, \dots, X_n) = \prod_{i=1}^n (\gamma + \delta(1 - X_i))^{Y_i} (1 - \gamma - \delta(1 - X_i))^{1-Y_i}$$

med parameterrummet: $\begin{pmatrix} \gamma \\ \delta \end{pmatrix} = \theta \in \Theta = \{\theta \in \mathbb{R}^2 : 0 < \gamma < 1, -1 < \delta < 1, 0 < \theta + \delta < 1\}$

δ kan i denne model fortolkes som mersandsynligheden for at være ryger hvis man ikke bruger cykelhjelm.

- Hvis $\delta < 0$ er sandsynligheden for at være ryger lavere for personer der ikke bruger cykelhjelm.
- Hvis $\delta = 0$ er sandsynligheden for at være ryger den samme for personer der bruger og ikke bruger cykelhjelm.
- Hvis $\delta > 0$ er sandsynligheden for at være ryger højere for personer der ikke bruger cykelhjelm.

I overensstemmelse med den analytiske løsning i delopgave 5 får vi:

$$\gamma = 0,2045 \text{ og } \delta = 0,0637.$$

7. Koder til estimationerne er vedlagt i do-fil.

Maximum Likelihood Estimation og Grænsesresultater for stokastiske variable

Følgende antagelser skal være opfyldt for at parametrene i en model for den stokastiske variabel Y_i kan estimeres ved Maximum Likelihood Estimation.

Assumption 3.1

- i) Den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y|\theta_i), \quad i = 1, 2, \dots, n,$$

hvor $f_{Y_i}(\cdot|\cdot)$ angiver sandsynlighedsfunktionen for Y_i . y_i er en realisering af Y_i og θ_i er sandsynlighedsfunktionens parameter.

- ii) De stokastiske variable er identisk fordelt således, at

$$\theta_i = \theta_j = \theta \quad \text{og} \quad \forall i, j | f_{Y_i}(y|\theta) = f_{Y_j}(y|\theta)$$

parameteren $\theta = (\theta_1, \dots, \theta_k)' \in \Theta \subseteq \mathbb{R}^k$

- iii) De stokastiske variable Y_i og Y_j er uafhængige for alle $i \neq j$ således, at den simultane sandsynlighed er givet ved

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n|\theta) &= f_{Y_1}(y_1|\theta) \cdot f_{Y_2}(y_2|\theta) \cdot \dots \cdot f_{Y_n}(y_n|\theta) \\ &= \prod_{i=1}^n f_{Y_i}(y_i|\theta) \end{aligned}$$

Theorem C.1 (Store Tals Lov)

Lad $\{Y_1, Y_2, \dots\}$ være en følge af ukorrelerede stokastiske variable med identisk middelværdier og varianser, $E[Y_i] = \mu$ og $V[Y_i] < \infty$.

Da gælder følgende for ethvert $\delta > 0$:

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n Y_i - \mu \right| \geq \delta \right) \rightarrow 0 \quad \text{for } n \rightarrow \infty$$

Theorem C.4 (Den Centrale Grænseværdi sætning)

Lad $\{Y_1, Y_2, \dots\}$ være iid stokastiske variable, $E[Y_i] = \mu$ og $V[Y_i] < \infty$.

\bar{Y} er gennemsnittet for Y . Da gælder følgende for $U_n = \sqrt{n} \frac{\bar{Y}_n - \mu}{\sigma}$:

$$P(U_n \leq u) \rightarrow \Phi(u) \quad \text{for } u \in \mathbb{R}$$

Opgave 1

Du får angivet en sample likelihood funktion

$$L(\theta|y_1, \dots, y_n), \quad \theta \in \mathbb{R}$$

Det opnåede maximum likelihood estimat er $\hat{\theta}_n = 12,41$ med en estimeret standardafvigelse på $se(\hat{\theta}_n) = 1,07$. Antag at de tilstrækkelige betingelser for Nielsen (2017, Teorem 1) er opfyldt.

- Angiv et 95% konfidensinterval for den sande værdi af parameteren, θ_0 .

Forklar fortolkningen af konfidensintervallet

Vi ved, at for en standard-normalfordelt stokastisk variabel $Z \sim N(0, 1)$ gælder følgende:

$$P(-1,96 \leq Z \leq 1,96) = \int_{-1,96}^{1,96} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = 0,95$$

For et maximum likelihood estimat $\hat{\theta}_n$ gælder *den centrale grænseværdisætning* såfremt nulhypotesen er korrekt:

$$\left(\frac{\hat{\theta}_n - \theta_0}{\sqrt{V(\hat{\theta}_n)}} \right) \sim N(0, 1)$$

θ_0 angiver her den sande værdi af θ .

Givet, at vi kender estimatet, $\hat{\theta}_n$ og standardafvigelsen, $se(\hat{\theta}_n) = \sqrt{V(\hat{\theta}_n)}$ kan vi altså beregne et konfidensinterval for den sande værdi af parameteren θ_0 .

$$\begin{aligned} P\left(-1,96 \leq \frac{\hat{\theta}_n - \theta_0}{se(\hat{\theta}_n)} \leq 1,96\right) &= 0,95 \Leftrightarrow \\ P\left(-1,96 \cdot se(\hat{\theta}_n) \leq \hat{\theta}_n - \theta_0 \leq 1,96 \cdot se(\hat{\theta}_n)\right) &= 0,95 \Leftrightarrow \\ P\left(1,96 \cdot se(\hat{\theta}_n) \geq \theta_0 - \hat{\theta}_n \geq -1,96 \cdot se(\hat{\theta}_n)\right) &= 0,95 \Leftrightarrow \\ P\left(\hat{\theta}_n - 1,96 \cdot se(\hat{\theta}_n) \leq \theta_0 \leq \hat{\theta}_n + 1,96 \cdot se(\hat{\theta}_n)\right) &= 0,95 \end{aligned}$$

Indsætter vi $\hat{\theta}_n = 12,41$ og $se(\hat{\theta}_n) = 1,07$ får vi:

$$P(12,41 - 1,96 \cdot 1,07 \leq \theta_0 \leq 12,41 + 1,96 \cdot 1,07) = P(10,31 \leq \theta_0 \leq 14,51) = 0,95$$

95% konfidensintervallet er derfor: 10,31;14,51

Konfidensintervallet angiver det interval hvor vi, under antagelse af asymptotisk konvergens af θ , med 95% sandsynlighed kan forvente at finde θ_0 .

- Fra økonomisk teori har vi en idé om, at parameteren θ_0 er lig 10.

Formulér denne idé som en statistisk hypotese. Vær præcis om den urekstrikterede model, nulhypotesen og alternativ-hypotesen. Specificér parameter-rummet for hver model.

Den urekstrikterede model:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log L(\theta | y_1, \dots, y_n), \quad \theta \in \Theta = \{\theta \in \mathbb{R}\}$$

Den statistiske hypotese opdeler parameterrummet i den urekstrikterede model i en *nulhypotese* og en *alternativhypotese*. Afhængig af testens resultat kan vi så afvise eller ikke afvise nulhypotesen. Hvis vi afviser nulhypotesen konkluderer vi, at alternativhypotesen gælder. De to hypoteser skal derfor indeholde alle mulige udfald.

I vores tilfælde:

$$\text{Nulhypotese, } \mathcal{H}_0 : \theta_0 = 10$$

$$\text{Alternativhypotese, } \mathcal{H}_A : \theta_0 \neq 10$$

Under nulhypotesen har vi altså følgende restrikterede model:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log L(\theta | y_1, \dots, y_n), \quad \theta \in \Theta = \{\theta = 10\}$$

og under alternativhypotesen har vi følgende restrikterede model:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log L(\theta | y_1, \dots, y_n), \quad \theta \in \Theta = \{\theta \in \mathbb{R} \setminus \{10\}\}$$

3. *Test hypotesen med et z-test. Find test-størrelsen, den kritiske værdi og p-værdien.*

Forklar hvordan test-størrelsen opfører sig hvis nul-hypotesen er korrekt og hvis den er forkert.

Test-størrelsen:

$$z_n(\theta_0 = 10) = \frac{\hat{\theta}_n - 10}{se(\hat{\theta}_n)} = \frac{12,41 - 10}{1,07} = \underline{\underline{2,25}}$$

Den kritiske værdi i en dobbeltsidet test med 95 % konfidensniveau er $\pm 1,96$

P-værdi:

$$P_n(\theta_0 = 10) = 2 \cdot \left(1 - \Phi(|z_n(\theta_0 = 10)|)\right) = 2 \cdot \left(1 - \Phi(2,25)\right) \approx \underline{\underline{0,0245}}$$

Hvis nulhypotesen $\mathcal{H}_0 : \theta_0 = a$ er korrekt gælder følgende for fordelingen af $\hat{\theta}_n$:

$$\hat{\theta}_n \sim N\left(a, \frac{1}{n}V(\theta)\right) \Leftrightarrow \frac{\hat{\theta}_n - a}{\sqrt{\frac{1}{n}V(\theta)}} \sim N(0, 1)$$

Med en korrekt nulhypotese går teststørrelsen z_n altså mod 0:

$$z_n(\theta_0 = a) = \frac{\hat{\theta}_n - a}{se(\hat{\theta}_n)} = \frac{\hat{\theta}_n - a}{\sqrt{\frac{1}{n}V(\theta)}} = \sqrt{n} \frac{\hat{\theta}_n - a}{\sqrt{V(\theta)}} \rightarrow N(0, 1)$$

fordi *Store Tals Lov* tilsiger at $\hat{\theta}_n \rightarrow \theta_0 = a \Leftrightarrow (\hat{\theta}_n - \theta_0) \rightarrow 0$.

Hvis $\theta_0 > a$:

$$z_n(\theta_0 = a) = \sqrt{n} \frac{\hat{\theta}_n - a}{\sqrt{V(\theta)}} \rightarrow \infty \text{ for } n \rightarrow \infty$$

da $\hat{\theta}_n - a > 0$.

Hvis $\theta_0 < a$:

$$z_n(\theta_0 = a) = \sqrt{n} \frac{\hat{\theta}_n - a}{\sqrt{V(\theta)}} \rightarrow -\infty \text{ for } n \rightarrow \infty$$

da $\hat{\theta}_n - a < 0$.

4. *Test hypotesen med et kvadreret Wald-test, dvs. z-test størrelsen i anden. Find test-størrelsen, den kritiske værdi og p-værdien.*

Forklar hvordan test-størrelsen opfører sig hvis nulhypotesen er korrekt og hvis den er forkert.

Test-størrelse w_n for kvadreret Wald-test:

$$w_n(\theta_0 = 10) = 2,25^2 \approx 5,07$$

Kritisk værdi: $|1,96|^2 = 3,84 = (\chi_1^2)^{-1}(0,95)$

P-værdi: $p_n(\theta_0 = 10) = 1 - F(5,07; 1) \approx \underline{\underline{0,0243}}$ hvor $F(\cdot; 1)$ betegner fordelingsfunktionen for $\chi^2(1)$ -fordelingen.

Test-størrelsens opførsel under korrekt og forkert nulhypoteze er ligesom ved et z-test.

Da test-størrelsen går mod en kendt fordeling hvis nulhypotesen er korrekt, og mod $\pm\infty$ hvis nulhypotesen er forkert kan vi kan slutte følgende om *Type 1* og *Type 2-fejl*:

Type 1 fejl er at forkaste en korrekt hypoteze.

Sandsynligheden for at begå en type 1 fejl går asymptotisk mod α
for en test på konfidensniveau $1 - \alpha$.

Type 2 fejl er ikke at forkaste en forkert hypoteze.

Sandsynligheden for at begå en type 2 fejl går asymptotisk mod 0
for test på alle konfidensniveauer.

Opgave 2

Betrægt igen levetiden af elsparepærerne fra ugeseddel 47, hvor vi anvendte eksponentialfordelingens tæthed,

$$f_{Y_i}(y|\theta) = \theta \exp\{-\theta y\}$$

for $y \in \mathbb{Y} = \mathbb{R}_+$ og $0 < \theta < \infty$, og antag at likelihood funktionen er korrekt specifiseret, sådan at

$$Y_i \stackrel{d}{=} \text{exponential}(\theta_0),$$

hvor θ_0 er den sande værdi af parameteren.

1. Nogle gange skrives eksponentialfordelingens tæthed med $\mu = E(Y) = \theta^{-1}$ som parameter, nemlig som

$$f_{Y_i}(y|\mu) = \frac{1}{\mu} \exp\left\{-\frac{y}{\mu}\right\}.$$

Find maximum likelihood estimatoren for μ .

Sammenlign med estimatoren for θ og kommentér på resultaterne.

Likelihood funktionen:

$$L(\mu|Y_1, \dots, Y_n) = \prod_{i=1}^n \frac{1}{\mu} \exp\left\{-\frac{Y_i}{\mu}\right\}$$

$$\mu \in \Theta = \{\mu \in R_+\}$$

Log-likelihood funktion:

$$\log L = \sum_{i=1}^n \left(-\log \mu - \frac{Y_i}{\mu} \right)$$

FOC:

$$\begin{aligned} \frac{\partial \log L}{\partial \hat{\mu}} &= \sum_{i=1}^n s_i(\hat{\mu}) = 0 \Leftrightarrow \\ &\sum_{i=1}^n \left(-\frac{1}{\hat{\mu}} + \frac{Y_i}{\hat{\mu}^2} \right) = 0 \Leftrightarrow \\ &\frac{n}{\hat{\mu}} = \frac{\sum_{i=1}^n Y_i}{\hat{\mu}^2} \\ &\hat{\mu} = \frac{\sum_{i=1}^n Y_i}{\underline{\underline{n}}} \end{aligned}$$

MLE, når $f_{Y_i}(y|\theta) = \theta \exp\{-\theta y\}$: $\hat{\theta} = \frac{n}{\sum_{i=1}^n Y_i}$.

Vi får altså at $\hat{\mu} = \hat{\theta}^{-1}$

Specificeringen af tæthedsfunktionen for Y med $\mu = E(Y)$ har den fordel at den forventede værdi af den stokastiske variabel kan aflæses direkte.

2. Vis ved direkte udregninger på formlen for estimatoren, $\hat{\mu}_n$, at den er unbiased (middelret).

Hvis estimatoren er unbiased gælder: $E(\hat{\mu}_n) = \mu_0 = E(Y)$. μ_0 betegner den sande værdi af μ .

$$E(\hat{\mu}_n) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \underbrace{\frac{1}{n} \sum_{i=1}^n E(Y_i)}_{Y \text{ er iid}} = E(Y) = \mu_0$$

Hermed er det vist, at $\hat{\mu}_n$ er unbiased.

3. Find ved direkte udregninger variansen af estimatoren, $V(\hat{\mu}_n)$

Husk at for iid stokastiske variable, Y_i , gælder: $V(\sum_{i=1}^n Y_i) = nV(Y_i)$.

$$V(\hat{\mu}_n) = V\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n Y_i\right) = \underbrace{\frac{1}{n} V(Y_i)}$$

Husk følgende resultat for eksponentialfordelingen:

$$f_{Y_i}(y_i) = \frac{1}{\mu} \exp\left\{-\frac{y_i}{\mu}\right\} \Leftrightarrow V(Y_i) = \mu^2$$

Vi kan altså skrive: $\frac{1}{n}V(Y_i) = \frac{1}{n}\mu_0^2$

Variansen kan også findes ved følgende lidt mere komplikerede udregning:

$$\begin{aligned} V(\hat{\mu}_n) &= E(\hat{\mu}_n^2) - (E(\hat{\mu}_n))^2 \\ &= E\left(\left(\frac{1}{n} \sum_{i=1}^n Y_i\right)^2\right) - \mu_0^2 \\ &= \frac{1}{n^2} E\left(\left(\sum_{i=1}^n Y_i\right)^2\right) - \mu_0^2 \quad (\text{benyt regneregel for kvadrat af } n\text{-leddet sum}) \\ &= \frac{1}{n^2} E\left(\sum_{i=1}^n Y_i^2 + 2 \sum_{\substack{i=1 \\ i+j=n \\ i < j}} Y_i Y_j\right) - \mu_0^2 \quad (\text{Benyt at } Y \text{ er iid}) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n E(Y_i^2) + 2 \sum_{i=1}^{n-1} E(Y_i)E(Y_{i+1}) \right) - \mu_0^2 \quad (\text{Benyt at } (1+2+\dots+n) = \frac{1}{2}(n+1)n) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n E(Y_i^2) + n(n-1)\mu_0^2 \right) - \mu_0^2 \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \left(V(Y_i) + (E(Y_i))^2 \right) \right) + \frac{n-1}{n}\mu_0^2 - \mu_0^2 \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n V(Y_i) \right) + \frac{1}{n^2} \sum_{i=1}^n \mu_0^2 - \frac{1}{n}\mu_0^2 \\ &= \frac{1}{n}V(Y_i) + \frac{1}{n}\mu_0^2 - \frac{1}{n}\mu_0^2 \\ &= \underline{\underline{\frac{1}{n}V(Y_i)}} \quad \left(= \frac{1}{n}\mu_0^2 \right) \end{aligned}$$

4. Angiv den asymptotiske fordeling af estimatoren

Vær præcis med hvilke antagelser der er tilstrækkelige.

Ifølge den centrale grænseværdidisætning gælder følgende:

Hvis $\{Y_1, Y_2, \dots\}$ er iid stokastiske variable med middelværdi, $E(Y_i) = \mu$ og endelig varians $V(Y_i) < \infty$, så er den asymptotiske fordeling af $\hat{\mu}_n = \frac{\sum_{i=1}^n Y_i}{n}$ givet ved:

$$\hat{\mu}_n \xrightarrow{a} N\left(\mu_0, \frac{V(Y_i)}{n}\right)$$

hvoraf følger, at:

$$\sqrt{n} \frac{\hat{\mu}_n - \mu_0}{se(Y_i)} \xrightarrow{a} N(0, 1)$$

Opgave 3

Vi fortsætter analysen ovenfor, og bruger den generelle udledning af variansen på estimatoren, baseret på den anden afledte af log-likelihood funktionen, se Teorem 1 i Nielsen (2017).

1. Opskriv log-likelihood bidraget for den statistiske model,

$$\log \ell(\mu|y_i) = \log f_{Y_i}(y_i|\mu).$$

Log-likelihood bidrag:

$$\begin{aligned} \log \ell(\mu|y_i) &= \log \left(\frac{1}{\mu} \exp \left\{ -\frac{y_i}{\mu} \right\} \right) \\ &= -\log(\mu) - \frac{y_i}{\mu} \end{aligned}$$

2. Find Hesse-bidraget, H_i , som den anden-ordens afledte,

$$H_i(\mu_0) = \frac{\partial^2 \log \ell(\mu|y_i)}{\partial \mu' \partial \mu} \Big|_{\mu=\mu_0}$$

Hesse-bidraget $H_i(\mu)$ er den første afledte af score-bidraget $s_i(\mu)$:

$$\begin{aligned} H_i(\mu_0) &= \frac{\partial^2 \log \ell(\mu|y_i)}{\partial \mu' \partial \mu} \Big|_{\mu=\mu_0} \\ &= \frac{\partial s_i(\mu)}{\partial \mu} \Big|_{\mu=\mu_0} \\ &= \frac{d}{d\mu} \left(-\frac{1}{\mu_0} + \frac{y_i}{\mu_0^2} \right) \\ &= \frac{1}{\mu_0^2} - 2 \frac{y_i}{\mu_0^3} \end{aligned}$$

3. Find informationen, defineret som

$$I(\mu_0) = -E(H_i(\mu_0)),$$

og find informationen evalueret i estimatoren, $I(\hat{\mu}_n)$.

Informationen:

$$\begin{aligned}
 I(\mu_0) &= -E(H_i(\mu_0)) \\
 &= E\left(2\frac{y_i}{\mu_0^3} - \frac{1}{\mu_0^2}\right) \\
 &= \frac{2E(y_i)}{\mu_0^3} - \frac{1}{\mu_0^2} \\
 &= \frac{2\mu_0}{\mu_0^3} - \frac{1}{\mu_0^2} \\
 &= \frac{1}{\underline{\mu_0^2}}
 \end{aligned}$$

Informationen evalueret i estimatoren $I(\hat{\mu}_n)$:

$$I(\hat{\mu}_n) = \frac{1}{\hat{\mu}_n^2}$$

4. Find variansen på estimatoren,

$$V(\hat{\mu}_n) = \frac{I(\hat{\mu}_n)^{-1}}{n},$$

og sammenlign med resultatet ovenfor.

Vi ved fra udregning i opgave 2, at $V(\hat{\mu}_n) = \frac{1}{n}\mu_0^2$. Vi bruger udtrykket for informationen udledt ovenfor:

$$\begin{aligned}
 I(\mu_0) &= \frac{1}{\mu_0^2} \\
 &= \frac{1}{nV(\hat{\mu}_n)} \Leftrightarrow \\
 V(\hat{\mu}_n) &= \frac{I(\mu_0)^{-1}}{n}
 \end{aligned}$$

Når vi har en endelig sample at estimere μ ud fra approksimerer vi den sande værdi μ_0 ud fra sample estimatet $\hat{\mu}_n$:

$$\mu_0 \approx \hat{\mu}_n$$

Som beregnet ovenfor er variansen af esimatoren en funktion af den sande værdi af μ :

$$V(\hat{\mu}_n) = \frac{1}{n}\mu_0^2 = \frac{I(\mu_0)^{-1}}{n}$$

Da vi ikke kender μ_0 er vores bedste bud på variansen at indsætte estimatoren $\hat{\mu}_n$ i informationen:

$$V(\hat{\mu}_n) = \frac{1}{n}\mu_0^2 \approx \frac{I(\hat{\mu}_n)^{-1}}{n}$$

Opgave 4

Betrægt igen datasættet med overlevende fra skibet *Titanic*, givet i filen *titanic.xls*. Husk at der er information om de 2201 passagerers overlevelsesstatus, køn, alder og rejsekasse:

$$\begin{aligned} \text{survived}_i &= \begin{cases} 1 & \text{hvis person } i \text{ overlevede} \\ 0 & \text{ellers,} \end{cases} \\ \text{female}_i &= \begin{cases} 1 & \text{hvis person } i \text{ var en kvinde} \\ 0 & \text{ellers,} \end{cases} \\ \text{child}_i &= \begin{cases} 1 & \text{hvis person } i \text{ var et barn} \\ 0 & \text{ellers,} \end{cases} \\ \text{class}_i &= \begin{cases} 1 & \text{hvis person } i \text{ rejste på første klasse} \\ 2 & \text{hvis person } i \text{ rejste på anden klasse} \\ 3 & \text{hvis person } i \text{ rejste på tredje klasse} \\ 0 & \text{hvis person } i \text{ var besætningsmedlem.} \end{cases} \end{aligned}$$

Lad Y_i være en stokastisk variabel svarende til observationerne $y_i = \text{survived}_i$. Antag at alle har samme sandsynlighed for at overleve, θ , $0 < \theta < 1$, og at observationerne er uafhængige.

1. Foreslå en statistisk model for datasættet $\{y_i\}_{i=1}^n$.

Angiv de antagelser du anvender og diskutér om de forekommer rimelige.

Find maximum likelihood estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$, og maximum likelihood estimatet, $\hat{\theta}(y_1, \dots, y_n)$.

Da den stokstatiske variabel Y_i har et binært udfaldsrum, $Y_i \in \{0, 1\}$ er en statistisk model med iid. *Bernoulli-fordelte* stokastiske variable den rette.

I denne model antager vi følgende:

- i) Sandsynligheden for et givet udfald, y_i , af den stokastiske variabel Y_i , betinget på parameteren, θ , er givet ved

$$f_{Y_i}(y_i) = \theta^{y_i} (1 - \theta)^{1-y_i}.$$

- ii) Alle udfald er realisationer fra samme *identiske* fordeling.
- iii) Alle udfald er uafhængige.

Antagelserne om identisk og uafhængig fordeling forekommer ikke hele rimelige. Vi ved fra tidligere, at overlevelseschancen om bord varierede mellem besætning og passagerer og mellem passagerernes rejsekasse.

Likelihood funktion ($\ell(\theta|Y_i)$ angiver likelihood bidrag):

$$\begin{aligned} L(\theta|Y_1, \dots, Y_n) &= \prod_{i=1}^n \ell(\theta|Y_i) \\ &= \prod_{i=1}^n \theta^{Y_i} (1 - \theta)^{1-Y_i} \\ &\quad \theta \in \Theta = \{\theta \in \mathbb{R} : 0 < \theta < 1\} \end{aligned}$$

Log-Likelihood funktion:

$$\begin{aligned}\log L(\theta|Y_1, \dots, Y_n) &= \sum_{i=1}^n \log \ell(\theta|Y_i) \\ &= \sum_{i=1}^n \left(Y_i \log \theta + (1 - Y_i) \log(1 - \theta) \right)\end{aligned}$$

Maximum likelihood estimatoren er det argument til $\log L(\theta|Y_1, \dots, Y_n)$ som maksimerer funktionsværdien:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \log L(\theta|Y_1, \dots, Y_n)$$

FOC:

$$\begin{aligned}\frac{\partial \log L(\theta|Y_1, \dots, Y_n)}{\partial \theta} &= \sum_{i=1}^n \frac{\partial \ell(\theta|Y_i)}{\partial \theta} = 0 \Leftrightarrow \\ \sum_{i=1}^n \left(\frac{Y_i}{\hat{\theta}} - \frac{1 - Y_i}{1 - \hat{\theta}} \right) &= 0 \Leftrightarrow \\ \frac{1}{\hat{\theta}} \sum_{i=1}^n Y_i &= \frac{1}{1 - \hat{\theta}} \sum_{i=1}^n (1 - Y_i) \Leftrightarrow \\ \frac{1 - \hat{\theta}}{\hat{\theta}} &= \frac{n - \sum_{i=1}^n Y_i}{\sum_{i=1}^n Y_i} \Leftrightarrow \\ \frac{1}{\hat{\theta}} - 1 &= \frac{n}{\sum_{i=1}^n Y_i} - 1 \Leftrightarrow \\ \hat{\theta}(Y_1, \dots, Y_n) &= \frac{\sum_{i=1}^n Y_i}{\underline{\underline{n}}}\end{aligned}$$

Maximum likelihood estimatet er estimatoren $\hat{\theta}$ som funktion af realisationer af den stokastiske variabel Y_i :

$$\hat{\theta}(y_1, \dots, y_n) = \frac{\sum_{i=1}^n y_i}{n}$$

I *Titanic*-datasættet er $\sum_{i=1}^n y_i = 711$ og $n = 2201$, hvilket giver estimatet

$$\hat{\theta}(y_1, \dots, y_{2201}) = \frac{711}{2201} = \underline{\underline{0,323}}$$

2. Find variansen på estimatoren, $V(\hat{\theta}(Y_1, \dots, Y_n))$, og angiv den asymptotiske fordeling for estimatoren.

$$\begin{aligned}V(\hat{\theta}(Y_1, \dots, Y_n)) &= V\left(\frac{\sum_{i=1}^n Y_i}{n}\right) \\ &= \frac{1}{n^2} V\left(\sum_{i=1}^n Y_i\right) \\ &= \frac{n}{n^2} V(Y_i) \\ &= \frac{1}{n} V(Y_i)\end{aligned}$$

Bemærk, at denne udregning benytter at de stokastiske variable, Y_i , $i = 1, \dots, n$, er identisk fordelt og uafhængige og dermed er $V(\sum_{i=1}^n Y_i) = nV(Y_i)$.

Den asymptotiske fordeling af estimatoren $\hat{\theta}(Y_1, \dots, Y_n)$ er givet ved *den centrale grænseværdidisætning* som siger, at fordelingen af estimatoren konvergerer mod en normalfordeling når n går mod uendelig:

$$\begin{aligned}\hat{\theta} &\sim N\left(\theta_0, \frac{1}{n}V(Y_i)\right) \Leftrightarrow \\ \hat{\theta} - \theta_0 &\sim N\left(0, \frac{1}{n}V(Y_i)\right) \Leftrightarrow \\ \sqrt{n}(\hat{\theta} - \theta_0) &\sim N(0, V(Y_i)) \Leftrightarrow \\ \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{se(Y_i)} &\sim N(0, 1), \text{ for } n \rightarrow \infty\end{aligned}$$

3. Vi vil nu også finde estimatet ved at maksimere log-likelihood funktionen numerisk. Maksimér log-likelihood funktionen og sammenlign med resultaterne ovenfor.

Følgende kode kan benyttes i STATA:

```
mat define init = J(1,1,0.2) /*starting values*/
mlexp ( survived*log({theta})+(1-survived)*log(1-{theta}) ) , from(init)
```

Hvilket giver følgende output:

```
Iteration 0:  log likelihood = -1476.7942
Iteration 1:  log likelihood = -1385.0065
Iteration 2:  log likelihood = -1384.7284
Iteration 3:  log likelihood = -1384.7284

Maximum likelihood estimation

Log likelihood = -1384.7284                               Number of obs      =     2,201


```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	.323035	.0099678	32.41	0.000	.3034985 .3425714

Den numeriske optimering giver altså estimatet $\hat{\theta}(y_1, \dots, y_{2201}) = 0,323$ hvilket er identisk med det ovenfor udregnede.

4. Vi vil nu undersøge scoren i vores initiale startværdi, dvs $S(\theta_{init})$, hvor θ_{init} er vores start-værdier. Vi kan gøre det ved kun at tage én numerisk iteration og bede STATA printe gradienten.

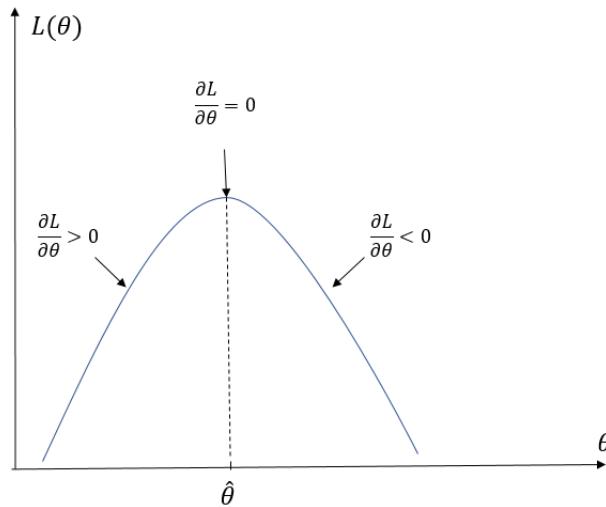
Følgende kode kan benyttes:

```
19 mlexp (survived*log({theta})+(1-survived)*log(1-{theta})) ,from(init) iter(0)
20 mat list e(gradient)
21
```

Hvis $S(\theta) > 0$, hvad er det så udtryk for? Scoren ved initial-værdien $\theta = 0,2$ er $S(\theta) \approx 1.693 > 0$. Likelihood funktionen er strengt konkav, da den samlede Hessian er entydigt negativ:

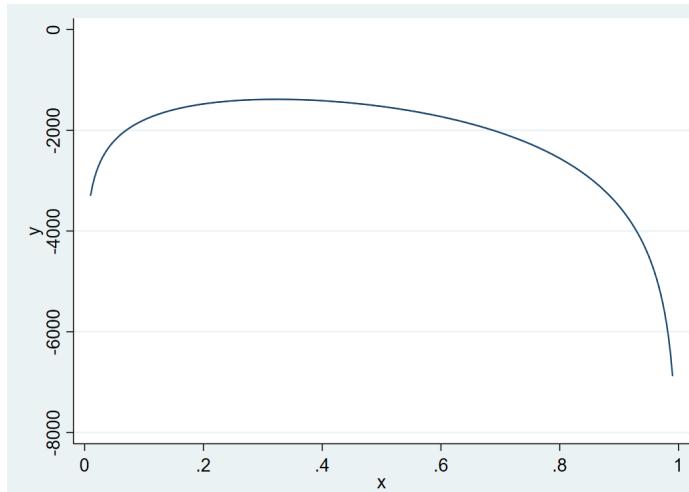
$$H(\theta) = \sum_{i=1}^n \frac{\partial^2 \ell(\theta|Y_i)}{\partial \theta' \partial \theta} = \sum_{i=1}^n \left(-\frac{Y_i}{\theta^2} - \frac{1-Y_i}{(1-\theta)^2} \right) < 0$$

En positiv score betyder derfor at vi ligger ”til venstre” for den værdi af θ som maksimerer funktionen. Man kan altså øge likelihood-værdien ved at øge θ .



Log-likelihood for survived tegnet i STATA:

```
twoway (function y=711*log(x)+(2201-711)*log(1-x), range(0.01 0.99))
```



5. Vi vil nu undersøge om sandsynligheden varierer mellem mænd og kvinder.
Vi opstiller derfor en ny model baseret på antagelsen

$$P(Y_i = 1) = \begin{cases} \theta & \text{hvis } \text{female}_i = 0 \\ \theta + \delta & \text{hvis } \text{female}_i = 1. \end{cases}$$

Vi kan fx definere en individ-specifik sandsynlighed

$$P(Y_i = 1) = \theta_i = \theta + \delta \cdot \text{female}_i,$$

sådan at sample likelihood funktionen bliver

$$L(\theta, \delta) = \prod_{i=1}^n \theta_i^{y_i} (1 - \theta_i)^{1-y_i}.$$

Find igen log-likelihood funktionen og omskriv koden til numerisk likelihood estimation til det nye tilfælde.

Estimér parametrene og kommentér på resultaterne.

Vi indfører nu den stokastiske variabel, X_i , svarende til observationerne $x_i = \text{female}_i$. Vi har, at $X_i = 1$ hvis person i er en kvinde, og $X_i = 0$ hvis person i ikke er en kvinde.

Likelihood funktionen bliver da:

$$L(\theta, \delta | y_1, \dots, y_n, x_1, \dots, x_n) = \prod_{i=1}^n (\theta + \delta x_i)^{y_i} (1 - \theta - \delta x_i)^{1-y_i}$$

$$\binom{\theta}{\delta} = \theta \in \Theta = \{\theta \in \mathbb{R}^2 : 0 < \theta < 1, 0 < \theta + \delta < 1\}$$

Log-likelihood funktion:

$$\log L(\theta, \delta | y_1, \dots, y_n, x_1, \dots, x_n) = \sum_{i=1}^n \left(y_i \log (\theta + \delta x_i) + (1 - y_i) \log (1 - \theta - \delta x_i) \right)$$

θ svarer til sandsynligheden for at overleve for mænd, og $\theta + \delta$ svarer til sandsynligheden for at overleve for kvinder.

Til den numeriske optimering bruges STATA-koden

```
mat define init = J(1,2,0.2)
mlexp ( survived*log({theta}+female*{delta})+(1-survived)*log(1-{theta}-female*{delta}) ) , from(init)
```

Output:

```

Iteration 0:  log likelihood = -1274.5996
Iteration 1:  log likelihood = -1168.3596
Iteration 2:  log likelihood = -1167.4972
Iteration 3:  log likelihood = -1167.4939
Iteration 4:  log likelihood = -1167.4939

Maximum likelihood estimation

Log likelihood = -1167.4939                               Number of obs      =     2,201


```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	.2120162	.0098241	21.58	0.000	.1927612 .2312711
/delta	.5198987	.0226714	22.93	0.000	.4754635 .5643339

Vi får altså estimaterne $\theta = 0,212$ og $\delta = 0,520$, hvilket fortæller at overlevelseschancen var 21,2% for ikke-kvinder, og $21,2\% + 52\% = 73,2\%$ for kvinder ombord på Titanic.

6. Forklar at denne likelihood estimation er et eksempel på en likelihood funktion baseret på en betinget fordeling.

Når vi inkorporerer varaiblen `female` i likelihood funktionen, så giver vi mulighed for, at den forventede overlevelseschance varierer med udfaldet af `female`. Dermed bliver udfaldet af `survived` betinget på udfaldet af `female`.

Find estimatorer for de betingede middelværdier

Vi definerer $n_1 = \sum_{i=1}^n Y_i(1 - X_i)$ som antallet af overlevede mænd ombord på Titanic og $n_2 = \sum_{i=1}^n Y_i X_i$ som antallet af overlevede kvinder ombord.

Estimatorer for de betingede middelværdier er da:

$$E(Y_i|female_i = 0) = \frac{\sum_{i=1}^n Y_i(1 - X_i)}{\sum_{i=1}^n (1 - X_i)}$$

$$E(Y_i|female_i = 1) = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i}$$

Opgave 5

Fortsæt med analysen ovenfor.

1. Opstil idéen om at mænd og kvinder blev behandlet ens under Titanics forlis som en statistisk hypotese.

Vær præcis om den urestrikterede model, nul-hypotesen og alternativ-hypotesen. Specificér parameterrummet for hver model.

Den urestrikterede model er:

$$Y_i \sim Bernoulli(\theta + \delta)$$

$$\begin{pmatrix} \theta \\ \delta \end{pmatrix} \in \Theta = \left\{ \begin{pmatrix} \theta \\ \delta \end{pmatrix} \in \mathbb{R}^2 : 0 < \theta < 1, 0 < \theta + \delta < 1 \right\}.$$

Hvis mænd og kvinder blev behandlet ens under Titanics forlis vil det gælde, at δ ikke er signifikant forskellig fra 0 statistisk set.

For at teste dette opstiller vi nulhypotesen:

$$\mathcal{H}_0 : \delta = 0$$

med alternativhypotesen

$$\mathcal{H}_A : \delta \neq 0.$$

Under nulhypotesen er parameterrummet:

$$\begin{pmatrix} \theta \\ \delta \end{pmatrix} \in \Theta = \left\{ \begin{pmatrix} \theta \\ \delta \end{pmatrix} \in \mathbb{R}^2 : 0 < \theta < 1, \delta = 0 \right\}.$$

Under alternativhypotesen er parameterrummet:

$$\begin{pmatrix} \theta \\ \delta \end{pmatrix} \in \Theta = \left\{ \begin{pmatrix} \theta \\ \delta \end{pmatrix} \in \mathbb{R}^2 : 0 < \theta < 1, 0 < \theta + \delta < 1, \delta \neq 0 \right\}.$$

Tilsammen udspænder de restrikterede parameterrum under \mathcal{H}_0 og \mathcal{H}_A det urestrikterede parameterrum.

2. *Test hypotesen med et z-test. Find test-størrelsen, den kritiske værdi og p-værdien.*

Forklar hvordan test-størrelsen opfører sig hvis nulhypotesen er korrekt og hvis den er forkert.

Fra outputtet ovenfor har vi et estimat og en standard afvigelse for δ .

$$\hat{\delta} = 0,520$$

$$se(\hat{\delta}) = 0,023$$

og udfra disse kan vi beregne vores test-størrelsen z_n :

$$z_n(\delta = 0) = \frac{0,520}{0,023} = 22,61.$$

Vi ved, at hvis nulhypotesen er korrekt, så er z_n standard normalfordelt:

$$z_n(\delta = 0) = \frac{\hat{\delta}}{se(\hat{\delta})} \sim N(0, 1)$$

og de kritiske værdier for en dobbeltsidet test på 95% konfidensniveau er 2,5%- og 97,5%-fraktilene, $\pm 1,96$.

Vi ser at test-størrelsen er langt over den kritiske værdi.

P-værdien er:

$$p_n(\delta = 0) = 2(1 - \Phi(22,61)) = 0,000.$$

P-værdien udtrykker sandsynligheden for den givne teststatistik eller en teststatistik endnu længere fra nul under nulhypotesen. Sandsynligheden for den her udregnede teststatistik er altså 0,000 under forudsætning af korrekt nulhypotese.

Hvis nulhypotesen er falsk ved vi at test-størrelsen vil gå mod uendelig, da $\sqrt{n}(\hat{\delta} - \delta_0) \rightarrow \pm\infty$.

Vi konkluderer, at dette er tilfældet i vores statistiske model og forkaster nulhypotesen.

3. *Test hypotesen med et kvadreret Wald test. Find test-størrelsen, den kritiske værdi og p-værdien.*

Forklar hvordan test-størrelsen opfører sig hvis nulhypotesen er korrekt og hvis den er forkert.

Test-størrelsen (w_n) for et kvadreret Wald test er kvadratet på test-størrelsen for et z-test:

$$w_n(\delta = 0) = (z_n(\delta = 0))^2 = 22,61^2 = 511,21$$

Den kritiske værdi er 95%-fraktilen i en $\chi^2(1)$ -fordeling, 3,84 ($= \pm 1,96^2$).

P-værdien er:

$$p_n = 1 - F(511, 21; 1) = 0,000$$

hvor $F(\cdot; 1)$ betegner fordelingsfunktionen for en $\chi^2(1)$ -fordeling.

Opførslen for test-størrelsen under korrekt of forkert nulhypotese er det samme som ved et z-test.

Baseret på test-størrelse og p-værdi forkaster vi også nulhypotesen efter et kvadreret Wald test.

4. *Test hypotesen med et likelihood ratio test. Find test-størrelsen, den kritiske værdi og p-værdien. Forklar hvordan test-størrelsen opfører sig hvis nulhypotesen er korrekt og hvis den er forkert.*

I et likelihood ratio (LR) test udregner vi test-størrelsen vha. log-likelihood værdierne fra den restrikterede og den urestrikterede model. Testen går ud på at måle om faldet i log-likelihood ved at restriktere modellen er signifikant forskelligt fra nul. Hvis nulhypotesen er sand er test-størrelsen i en LR-test $\chi^2(1)$ -fordelt, og den kritiske værdi er dermed 3,84.

Fra output ovenfor ser vi at log-likelihood for den restrikterede model $\delta = 0$ er $-1384,73$.

For den urestrikterede model er log-likelihood $-1167,49$.

Likelihood ratio:

$$LR_n = 2(\log L(\theta, \delta) - \log L(\theta, 0)) = 2(-1167,49 - (-1384,73)) = 434,48$$

P-værdien er:

$$p_n = 1 - F(434,48; 1) = 0,000$$

hvor $F(\cdot; 1)$ betegner fordelingsfunktionen for en $\chi^2(1)$ -fordeling.

Opførslen for test-størrelsen under korrekt og forkert nulhypotese er det samme som ved et z-test og et Wald test.

Baseret på test-størrelse og p-værdi forkaster vi også nulhypotesen efter et likelihood ratio test.

Opgave 6

Fortsæt med analysen ovenfor.

1. *Opstil hypotesen om at mænd og kvinder blev behandlet ens og at overlevelsessandsynligheden var præcis $\theta = 0,35$ for alle på skibet.*

Vær igen præcis om den urestrikterede model, nul-hypotesen og alternativ-hypotesen. Specificér parameter-rummet for hver model.

Vores hypotese indeholder nu to restriktioner på data.

I den urestrikterede model er parameterrummet:

$$\begin{pmatrix} \theta \\ \delta \end{pmatrix} = \theta \in \Theta = \left\{ \theta \in \mathbb{R}^2 : 0 < \theta < 1, 0 < \theta + \delta < 1 \right\}$$

Nulhypotesen med tilørende parameterrum:

$$\mathcal{H}_0 : \theta = \begin{pmatrix} 0,35 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \theta \\ \delta \end{pmatrix} = \theta \in \Theta = \left\{ \theta = \begin{pmatrix} \theta \\ \delta \end{pmatrix} = \begin{pmatrix} 0,35 \\ 0 \end{pmatrix} \right\}$$

Alternativhypotesen med tilhørende parameterrum:

$$\mathcal{H}_A : \theta \neq \begin{pmatrix} 0,35 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \theta \\ \delta \end{pmatrix} = \theta \in \Theta = \left\{ \theta \in \mathbb{R}^2 : 0 < \theta < 1, 0 < \theta + \delta < 1, \begin{pmatrix} \theta \\ \delta \end{pmatrix} \neq \begin{pmatrix} 0,35 \\ 0 \end{pmatrix} \right\}$$

2. Test hypotesen med et likelihood ratio test

I dette tilfælde kan likelihood funktionen ikke analyseres i STATA, fordi der ikke er nogen parametre at estimere under nul-hypotesen. Opskriv og udregn derfor log-likelihood funktionen i hånden.

Log-likelihood funktion for den urestrikterede model:

$$\begin{aligned} \log L(\theta, \delta | y_1, \dots, y_n, x_1, \dots, x_n) &= \sum_{i=1}^n \left(y_i \log(\theta + \delta x_i) + (1 - y_i) \log(1 - \theta - \delta x_i) \right) \\ &= \sum_{i=1}^n y_i \log(\theta + \delta x_i) + \sum_{i=1}^n (1 - y_i) \log(1 - \theta - \delta x_i) \end{aligned}$$

Antallet af overlevende: $y_i = 711$

Antallet af ikke-overlevende: $n - y_i = 1490$

Antallet af overlevende kvinder: $y_i x_1 = 344$

Antallet af ikke-overlevende kvinder: $x_i - y_i x_1 = 126$

Estimater i urestrikteret model: $\hat{\theta} = \begin{pmatrix} \hat{\theta} \\ \hat{\delta} \end{pmatrix} = \begin{pmatrix} 0,21 \\ 0,52 \end{pmatrix}$

Dermed får vi følgende log-likelihood værdi fra den urestrikterede model:

$$\begin{aligned} \log L(0, 21; 0, 52 | y_1, \dots, y_{2201}, x_1, \dots, x_{2201}) &= (711 - 344) \log(0, 21) \\ &\quad + 344 \log(0, 21 + 0, 52) + (1490 - 126) \log(1 - 0, 21) \\ &\quad + 126 \log(1 - 0, 21 - 0, 52) \\ &= -1167,52 \end{aligned}$$

Estimater i restrikteret model: $\tilde{\theta} = \begin{pmatrix} \tilde{\theta} \\ \tilde{\delta} \end{pmatrix} = \begin{pmatrix} 0,35 \\ 0 \end{pmatrix}$

Dermed får vi følgende log-likelihood værdi fra den restrikterede model:

$$\begin{aligned} \log L(0, 35; 0 | y_1, \dots, y_{2201}, x_1, \dots, x_{2201}) &= (711 - 344) \log(0, 35) \\ &\quad + 344 \log(0, 35 + 0) + (1490 - 126) \log(1 - 0, 35) \\ &\quad + 126 \log(1 - 0, 35 - 0) \\ &= -1388,29 \end{aligned}$$

Med log-likelihood værdierne fra den restrikterede og den urestrikterede model kan vi beregne teststørrelsen i et likelihood ratio test:

$$\begin{aligned} LR_n \left(\theta = \begin{pmatrix} 0,35 \\ 0 \end{pmatrix} \right) &= 2(\log L(\theta | y_1, \dots, y_n) - \log L(\theta | \tilde{y}_1, \dots, \tilde{y}_n)) \\ &= 2(-1167,52 - (-1388,29)) \\ &= 441,7 \end{aligned}$$

Hvis nulhypotesen er sand er teststørrelsen fordelt med en $\chi^2(2)$ -fordeling.

$$LR_n \sim \chi^2(2) \text{ (hvis } H_0 \text{ er sand)}$$

Den kritiske værdi for en test på 95% konfidensniveau er 95%-fraktilen i denne fordeling: 5,99.

Bemærk at denne kritiske værdi er højere end i en $\chi^2(1)$ -fordeling da vi nu har to frihedsgrader som følge af, at vores hypotese lægger to restriktioner på data.

Da test-størrelsen er langt højere end den kritiske værdi forkaster vi nulhypotesen om, at overlevels chance for alle ombord på Titanic var 35% og at der ikke var forskel på mænd og kvinder.

3. *Kan denne hypotese testes med et z-test.*

Z-test kan kun benyttes til at teste en restriktion på data.

Når man vil teste restriktioner på flere variable samtidigt kan man ikke benytte et z-test. Her benytter man LR-test eller F-test.

Maximum Likelihood Estimation og Grænseresultater for stokastiske variable

Centrale begreber:

Assumption 3.1

- i) Den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y|\theta_i) \quad , \quad i = 1, 2, \dots, n,$$

hvor $f_{Y_i}(\cdot|\cdot)$ angiver sandsynlighedsfunktionen for Y_i . y_i er en realisering af Y_i og θ_i er sandsynlighedsfunktionens parameter.

- ii) De stokastiske variable er identisk fordelt således, at

$$\theta_i = \theta_j = \theta \quad \text{og} \quad \forall i, j \quad | \quad f_{Y_i}(y|\theta) = f_{Y_j}(y|\theta)$$

parameteren $\theta = (\theta_1, \dots, \theta_k)' \in \Theta \subseteq \mathbb{R}^k$

- iii) De stokastiske variable Y_i og Y_j er uafhængige for alle $i \neq j$ således, at den simultane sandsynlighed er givet ved

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n|\theta) &= f_{Y_1}(y_1|\theta) \cdot f_{Y_2}(y_2|\theta) \cdot \dots \cdot f_{Y_n}(y_n|\theta) \\ &= \prod_{i=1}^n f_{Y_i}(y_i|\theta) \end{aligned}$$

Theorem 4.1

Under Assumption 3.1, such that the emodel is correctly specified with the p.f. or p.d.f. for the DGP given by $f_{Y_i}(y_1, \dots, y_n|\theta_0)$, and under some regularity conditions it holds that

- (a) The MLE is consistent $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \rightarrow \infty$

- (b) The MLE is asymptotically normal, i.e. as $n \rightarrow \infty$,

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Omega_\theta)$$

- (c) The asymptotic variance is given by $\Omega_\theta = I(\theta_0)^{-1}$, where

$$I(\theta_0) = -E(H_i(\theta_0))$$

is the information matrix, and

$$H_i(\theta_0) = \left. \frac{\partial^2 \log \ell(\theta|Y_i)}{\partial \theta \partial \theta'} \right|_{\theta=\theta_0}$$

is the Hessian matrix corresponding to a single observation.

- (d) The MLE is asymptotically efficient: all other consistent and asymptotically normal estimators have asymptotic variance larger than or equal to $I(\theta)^{-1}$.

Opgave 1

Hele likelihood analysen er nu dækket af forelæsninger. Det skulle gerne være fremgået at likelihood-analysen er meget skematisk og altid indeholder de samme trin.

1. Start med tætheden

$$f_{Y_i}(y | \theta).$$

Lav en kogebog med de samme trin du ville gå igennem for at finde maksimum likelihood estimatoren og estimatelet.

Vi husker de nødvendige antagelser for Maximum Likelihood Estimering:

Assumption 3.1

- i) Den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y | \theta_i) \quad , \quad i = 1, 2, \dots, n,$$

hvor $f_{Y_i}(\cdot | \cdot)$ angiver sandsynlighedsfunktionen for Y_i . y_i er en realisering af Y_i og θ_i er sandsynlighedsfunktionens parameter.

- ii) De stokastiske variable er identisk fordelt således, at

$$\theta_i = \theta_j = \theta \quad \text{og} \quad \forall i, j \mid f_{Y_i}(y | \theta) = f_{Y_j}(y | \theta)$$

parameteren $\theta = (\theta_1, \dots, \theta_k)' \in \Theta \subseteq \mathbb{R}^k$

- iii) De stokastiske variable Y_i og Y_j er uafhængige for alle $i \neq j$ således, at den simultane sandsynlighed er givet ved

$$\begin{aligned} f_{Y_1, \dots, Y_n}(y_1, \dots, y_n | \theta) &= f_{Y_1}(y_1 | \theta) \cdot f_{Y_2}(y_2 | \theta) \cdot \dots \cdot f_{Y_n}(y_n | \theta) \\ &= \prod_{i=1}^n f_{Y_i}(y_i | \theta) \end{aligned}$$

Vi antager, at vores observerede data kan beskrives ved en stokastisk proces, med en kendt sandsynlighedstæthed $f_{Y_i}(y | \theta)$, hvor de stokastiske variable er identisk og uafhængigt fordelt. Ud fra disse antagelser kan vi formulere *likelihood funktionen* L som produktet af alle *likelihood bidragene* ℓ :

$$L(\theta | Y_1, \dots, Y_n) = \prod_{i=1}^n \ell(\theta | Y_i)$$

hvor n er antallet af observationer, og $\ell(\theta | Y_i) = f_{Y_i}(Y_i | \theta)$.

For at kunne udtrykke den samlede likelihood som en sum i stedet for et produkt (nemmere at differentiere) udleder vi *log-likelihood funktionen*

$$\log L(\theta | Y_1, \dots, Y_n) = \sum_{i=1}^n \log \ell(\theta | Y_i)$$

Da transformationen ved den naturlige logaritme er en monoton transformation vil det samme argument til funktionen maksimere L og $\log L$.

Vi vil gerne finde det globale maksimum for likelihood funktionen som fortæller os hvilken værdi af θ der giver den største sandsynlighed for vores observationer, under de tre antagelser. Vi løser derfor førsteordensbetingelsen for funktionen ved at sætte den første afledte (*scoren*) lig 0.

Da de stokastiske variable Y_i , $i = 1, \dots, n$ er identisk fordelte kan vi udlede den samlede score $S(\theta)$ som summen fra 1 til n af score-bidragene:

$$S(\theta|Y_1, \dots, Y_n) = \sum_{i=1}^n s_i(\theta|Y_i) = \sum_{i=1}^n \frac{\partial \ell(\theta|Y_i)}{\partial \theta} = \frac{\partial L(\theta|Y_1, \dots, Y_n)}{\partial \theta}$$

Maximum likelihood estimatoren ($\hat{\theta}_n$) findes da som den værdi af θ der løser førsteordensbetingelsen

$$S(\hat{\theta}|Y_1, \dots, Y_n) = 0$$

og dermed har vi udledt *maximum likelihood estimatoren* (*MLE*) $\hat{\theta}_n(Y_1, \dots, Y_n)$.

Maximum likelihood estimatoren $\hat{\theta}_n(y_1, \dots, y_n)$ findes ved at indsætte de observerede værdier y_i .

Strengt taget bør det vises, at den anden afledte af L og $\log L$ er negativ definit i parameterrummet. Dette sikrer at den fundne estimator maksimerer funktionen.

2. Lav en tilsvarende kogebog for at finde variansen på maksimum likelihood estimatoren ud fra den anden afledte og informationen.

I *Theorem 4.1* på side 81 i bogen står der under punkt b og c:

- (b) *The MLE is asymptotically normal, i.e. as $n \rightarrow \infty$,*

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Omega_\theta).$$

- (c) *The asymptotic variance is given by $\Omega_\theta = I(\theta_0)^{-1}$, where*

$$I(\theta_0) = -E(H_i(\theta_0))$$

is the information matrix, and

$$H_i(\theta_0) = \left. \frac{\partial^2 \log \ell(\theta|Y_i)}{\partial \theta \partial \theta'} \right|_{\theta=\theta_0}$$

is the Hessian matrix corresponding to a single observation.

Vi kan altså finde variansen på maksimum likelihood estimatoren via den anden afledte af likelihood bidraget. Informationen er givet ved den negative værdi af forventningen (vi ved at den anden afledte må være negativ, da likelihood bidraget når sit maksimum ved $\theta = \theta_0$. For at få en positiv varians ganger vi derfor med -1).

Hvis man allerede har udledt maksimum likelihood estimatoren, så har man et udtryk for score bidraget. Differentieres dette fås *Hesse bidraget*:

$$H_i(\theta) = \frac{\partial s(\theta|Y_i)}{\partial \theta}.$$

Da vi ikke kender θ_0 benytter vi vores maksimum likelihood estimator $\hat{\theta}_n$:

$$H_i(\theta_0) \approx H_i(\hat{\theta}_n) = \frac{\partial s(\hat{\theta}_n|Y_i)}{\partial \hat{\theta}}.$$

Informationen findes da ved at benytte regneregler for forventet værdi:

$$I(\theta_0) \approx -E(H_i(\hat{\theta}_n))$$

Ifølge *Theorem 4.1b* gælder:

$$\begin{aligned} \sqrt{n} (\hat{\theta}_n - \theta_0) &\xrightarrow{d} N(0, \Omega_\theta) \Leftrightarrow \\ \hat{\theta}_n &\xrightarrow{d} N(\theta_0, \frac{\Omega_\theta}{n}) \end{aligned}$$

og da vi fra (c) har, at $\Omega_\theta = I(\theta_0)^{-1} \approx -E(H_i(\hat{\theta}))^{-1}$ kan vi nu opstille et udtryk for variansen af vores estimator baseret på sampelen:

$$Var(\hat{\theta}_n) = \frac{I(\theta_0)^{-1}}{n} \approx \frac{-E(H_i(\hat{\theta}_n))^{-1}}{n}$$

Hvis den forventede værdi af den anden afledte er høj får vi altså en lille varians.

Prøv at tegne en likelihood funktion med en "spids" top ved $\hat{\theta}_n$ og en likelihood funktion med "flad" top ved $\hat{\theta}_n$.

- Likelihood funktionen med spids top ved $\hat{\theta}_n$ vil have relativt højere værdi af den anden afledte, da den første afledte ændrer sig hurtigere. Dette giver relativt mange værdier af θ tæt på $\hat{\theta}_n$ og dermed mindre varians.
 - Likelihood funktionen med flad top ved $\hat{\theta}_n$ vil have relativt lavere værdi af den anden afledte, da den første afledte ændrer sig langsommere. Dette giver relativt færre værdier af θ tæt på $\hat{\theta}_n$ og dermed højere varians.
-

Opgave 2

Vi vil undersøge elevers skolefravær i 7.-8. klasse. Filen `skole.xls` indeholder information om 165 elevers fravær. Variablen $\{y_i\}_{i=1}^{165} = \{\text{uger}_i\}_{i=1}^{165}$ mäter hvor mange uger i træk en given elev møder i skole uden fravær, mens $\{\text{dreng}_i\}_{i=1}^{165}$ er en dummy-variabel der tager værdien 1 hvis eleven er en dreng.

Eleverne er tilfældigt udvalgt fra forskellige skoler, og vi antager som udgangspunkt, at der hver uge er en konstant sandsynlighed, θ , for at elev i er syg i en given uge, og bliver hjemme fra skolen. I det tilfælde vil den stokastiske variabel Y_i , der repræsenterer det observerede antal uger, være geometrisk fordelt,

$$Y_i \stackrel{d}{=} \text{Geometric}(\theta)$$

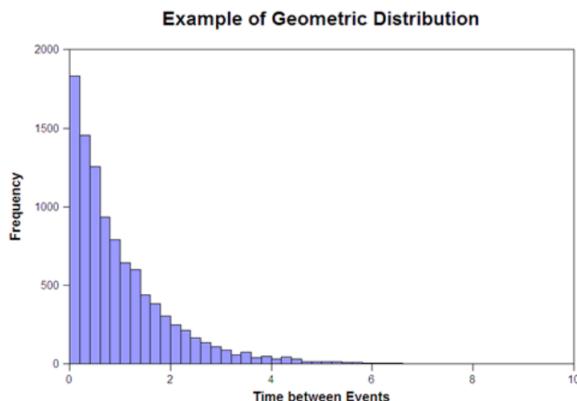
med sandsynlighedsfunktionen givet ved

$$P(Y_i = y) = (1 - \theta)^y \theta, \quad y \in \{0, 1, 2, 3, \dots\}$$

1. Brug beskrivende statistik til at karakterisere data.

I indeværende datasæt, hvor vi kan opdele eleverne på køn, er det interessant og se på både *ubetingede* og *betingede* data. Ubetingede data er for alle skoleeleverne, og betingede kunne være opdelt på køn. Data bør fortælle om minimum, maksimum, median, gennemsnit og spredning. Til spredning er standardafvigelsen ofte et bedre mål end varians, da standardafvigelsen er i samme måleenhed som gennemsnittet (varianser er i kvadrerede værdier).

Til grafisk sammenligning er her en typisk geometrisk fordeling:



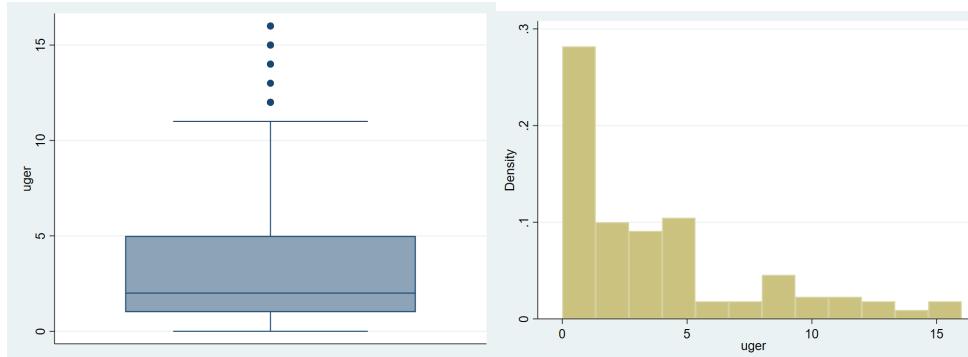
Fra STATA:

- Ubetingede data:

variable	mean	p50	sd	skewness	kurtosis	max	min	N
uger	3.690909	2	3.976483	1.301522	3.912086	16	0	165

Hvad fortæller data os?

- Boxplot og Histogram:



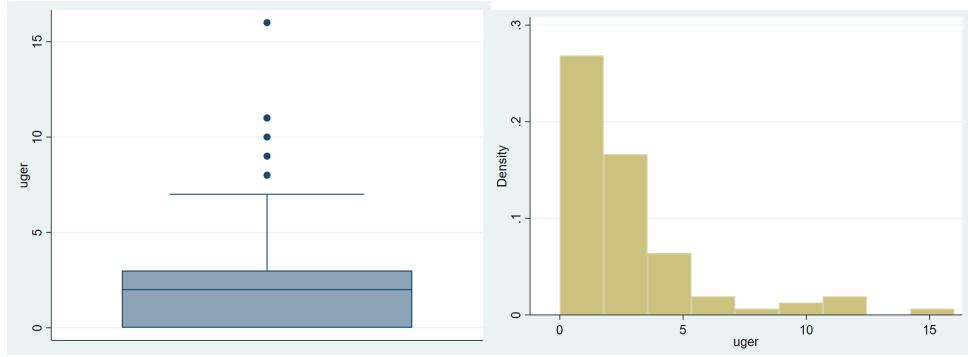
- Hvad fortæller boxplottet os?
- Ligner histogrammet en geometrisk fordeling?
- Betingede data for drenge:

-> dreng = 1

variable	mean	p50	sd	skewness	kurtosis	max	min	N
uger	2.511364	2	3.039942	2.036844	7.679182	16	0	88

Hvad fortæller data os. Hvordan sammenligner observationer sig med ubetingede data?

- Boxplot og Histogram:



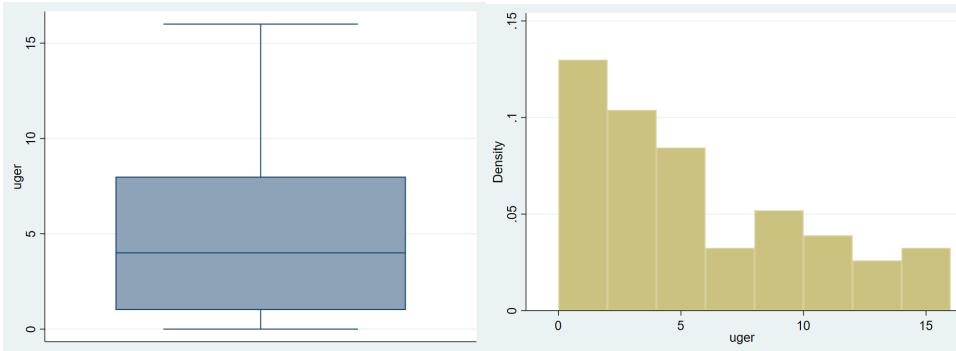
- Hvad fortæller boxplottet os?
- Ligner histogrammet en geometrisk fordeling?
- Betingede data for piger:

-> dreng = 0

variable	mean	p50	sd	skewness	kurtosis	max	min	N
uger	5.038961	4	4.48225	.7436852	2.488347	16	0	77

Hvad fortæller data os. Hvordan sammenligner observationer sig med ubetingede data?

- Boxplot og Histogram:



- Hvad fortæller boxplottet os?
- Ligner histogrammet en geometrisk fordeling?

2. Opskriv en statistisk model, dvs. likelihood funktionen og parameterrummet, svarende til beskrivelsen ovenfor.

Angiv de antagelser du anvender og diskutér om de forekommer rimelige i dette eksempel.

Da vi har en sandsynlighedsfunktion for data, $P(Y_i = y) = (1 - \theta)^y \theta$, og antagelser om identisk og uafhængig fordeling kan vi opstille en likelihood funktion for data. For at udlede parameterrummet for vores parameter, θ er vi dog nødt til at vide hvad den udtrykker.

Tænk på udtrykket

$$f_{Y_i}(y|\theta) = (1 - \theta)^y \theta$$

som en bernoulli-fordeling

$$f_{Y_i}(y|\theta) = \theta^y (1 - \theta)^{1-y}, \quad y \in \{0, 1\}$$

hvor man har haft udfaldet ($Y_i = 0$) de første y gange og udfaldet ($Y_i = 1$) den $(y + 1)$ 'te gang.

Med andre ord udtrykker *den geometriske fordeling* sandsynligheden for ventetiden inden et "gunstigt" udfald. θ er sandsynligheden for dette gunstige udfald ved hver (uafhængig) gentagelse. Den ækvivalente kontinuære fordeling er eksponentialfordelingen (se Opgave 2 på ugeseddel 47).

Da vi ved, at θ er en sandsynlighed kan vi opskrive likelihood funktionen med tilhørende parameterrum ($\ell(\cdot|\cdot)$ angiver likelihood-bidrag):

$$\begin{aligned} L(\theta|Y_1, \dots, Y_n) &= \prod_{i=1}^n \ell(\theta|Y_i) \\ &= \prod_{i=1}^n (1 - \theta)^{Y_i} \theta, \quad \theta \in \Theta = \{\theta \in \mathbb{R} : 0 < \theta < 1\} \end{aligned}$$

Som altid bygger maksimum likelihood estimering på tre antagelser:

- i) Den stokastiske variabel, hvis realiseringer vi observerer i data er beskrevet ved en sandsynlighedsfunktion. I vores tilfælde sandsynlighedsfunktionen for den geometriske fordeling. Dette virker rimeligt, og den geometriske fordeling virker som en god model til at beskrive "ventetiden" på, at en elev har fravær.
- ii) De stokastiske variable er identisk fordelt. θ er ens for alle elever.
Dette forekommer *ikke* rimeligt. Den deskriptive statistik peger bl.a. på en systematisk forskel på θ mellem køn.

- iii) De stokastiske variable Y_i og Y_j er uafhængige. θ for en elev er ikke påvirket af fraværet fra andre i samme klasse.

Dette forekommer heller ikke rimeligt. Hvis en elev er i en klasse med meget lavt fravær, vil det højest sandsynligt føre til lavere fravær for denne elev. Altså bliver θ lavere, hvis andre i nærheden har lave θ .

3. Find maksimum likelihood estimatoren, $\hat{\theta}(Y_1, \dots, Y_n)$

Vi finder log-likelihood funktionen for lettere at kunne differentiere.

$$\begin{aligned} \log L(\theta | Y_1, \dots, Y_n) &= \log \prod_{i=1}^n \ell(\theta | Y_i) \\ &= \sum_{i=1}^n \log \ell(\theta | Y_i) \\ &= \sum_{i=1}^n \log (1 - \theta)^{Y_i} \theta \\ &= \sum_{i=1}^n \left(Y_i \log(1 - \theta) + \log(\theta) \right) \end{aligned}$$

Samlet score

$$\begin{aligned} S(\theta | Y_1, \dots, Y_n) &= \frac{\partial L(\theta | Y_1, \dots, Y_n)}{\partial \theta} \\ &= \sum_{i=1}^n \frac{\partial \ell(\theta | Y_i)}{\partial \theta} \\ &= \sum_{i=1}^n \left(-\frac{Y_i}{1 - \theta} + \frac{1}{\theta} \right) \\ &= \frac{n}{\theta} - \frac{\sum_{i=1}^n Y_i}{(1 - \theta)}. \end{aligned}$$

Maksimum likelihood estimatoren $\hat{\theta}(Y_1, \dots, Y_n)$ er den værdi af θ der løser førsteordensbetingelsen

$$\begin{aligned} S(\hat{\theta} | Y_1, \dots, Y_n) = 0 &\Leftrightarrow \\ \frac{n}{\hat{\theta}} - \frac{\sum_{i=1}^n Y_i}{(1 - \hat{\theta})} &\Leftrightarrow \\ \frac{1}{\hat{\theta}} - 1 = \frac{\sum_{i=1}^n Y_i}{n} &\Leftrightarrow \\ \frac{1}{\hat{\theta}} = \frac{\sum_{i=1}^n Y_i + n}{n} &\Leftrightarrow \\ \hat{\theta} &= \underline{\frac{n}{\sum_{i=1}^n Y_i + n}} \end{aligned}$$

Prøv at tænke over θ når $\sum_{i=1}^n Y_i = 0$ og når $\sum_{i=1}^n Y_i \rightarrow \infty$.

4. Brug informationen fra de observerede data til også at finde estimatet $\hat{\theta}(y_1, \dots, y_n)$

Vi indsætter $n = 165$ og $\sum_{i=1}^n = 609$ (beregnet i STATA):

$$\begin{aligned}\hat{\theta}(y_1, \dots, y_n) &= \frac{165}{609 + 165} \\ &\approx 0,2132\end{aligned}$$

Baseret på data har en tilfældigt udvalgt elev altså 21,32% risiko for fravær i en givet uge.

5. Vis, at Hesse-matricen, $H_i(\theta)$ er givet ved

$$H_i(\theta) = \frac{\partial^2 \log \ell(\theta|Y_i)}{\partial \theta \partial \theta} = -\frac{1}{\theta^2} - \frac{Y_i}{(1-\theta)^2}.$$

Hesse-matricen, eller Hesse bidraget er den anden afledte af log-likelihood bidraget. Vi kender allerede score bidraget som er den første afledte af log-likelihood bidraget, så vi kan differentiere dette:

$$\begin{aligned}H_i(\theta) &= \frac{\partial^2 \log \ell(\theta|Y_i)}{\partial \theta \partial \theta} \\ &= \frac{\partial s(\theta|Y_i)}{\partial \theta} \\ &= \frac{\partial}{\partial \theta} \left(\frac{1}{\theta} - \frac{Y_i}{(1-\theta)} \right) \\ &= \underline{-\frac{1}{\theta^2}} - \underline{\frac{Y_i}{(1-\theta)^2}}\end{aligned}$$

6. Find informationen, $I(\theta) = -E(H_i(\theta))$, idet vi bemærker, at $E(Y_i) = \frac{1-\theta_0}{\theta_0}$, hvor θ_0 er den sande værdi.

Brug det til at finde variansen på estimatoren,

$$V(\hat{\theta}(Y_1, \dots, Y_n)) = \frac{I(\theta_0)^{-1}}{n} \approx \frac{I(\hat{\theta})^{-1}}{n}$$

hvor informationen er evalueret i estimatoren.

$$\begin{aligned}I(\theta) &= -E(H_i(\theta)) \\ &= -E \left(-\frac{1}{\theta^2} - \frac{Y_i}{(1-\theta)^2} \right) \\ &= \frac{1}{\theta^2} + \frac{E(Y_i)}{(1-\theta)^2} \\ &= \frac{1}{\theta^2} + \frac{1-\theta_0}{\theta_0(1-\theta)^2}\end{aligned}$$

Læg mærke til at informationen er en funktion af θ og θ_0 er en (ukendt) konstant.

Når vi udregner variansen bruger vi informationen evalueret i estimatoren $\hat{\theta}$. Og på θ_0 's plads indsætter vi også $\hat{\theta}$, da det er vores bedste bud på den sande værdi:

$$\begin{aligned}
 V(\hat{\theta}(Y_1, \dots, Y_n)) &= \frac{I(\theta_0)^{-1}}{n} \\
 &\approx \frac{I(\hat{\theta})^{-1}}{n} \\
 &= \frac{1}{n} \left(\frac{1}{\hat{\theta}^2} + \frac{1 - \hat{\theta}}{\hat{\theta}(1 - \hat{\theta})^2} \right)^{-1} \\
 &= \frac{1}{n} \left(\frac{1}{\hat{\theta}^2} + \frac{1}{\hat{\theta}(1 - \hat{\theta})} \right)^{-1} \\
 &= \frac{1}{n} \left(\frac{1 - \hat{\theta}}{\hat{\theta}^2(1 - \hat{\theta})} + \frac{\hat{\theta}}{\hat{\theta}^2(1 - \hat{\theta})} \right)^{-1} \\
 &= \frac{1}{n} \left(\frac{1}{\hat{\theta}^2(1 - \hat{\theta})} \right)^{-1} \\
 &= \frac{\hat{\theta}^2(1 - \hat{\theta})}{n}
 \end{aligned}$$

Variansen af estimatoren baseret på vores sample:

$$\begin{aligned}
 V(\hat{\theta}(y_1, \dots, y_{165})) &= \frac{0,2132^2(1 - 0,2132)}{165} \\
 &\approx 0,0002
 \end{aligned}$$

7. Angiv et 95% konfidensinterval for θ_0 .

Vi antager at *den centrale grænse værdi sætning* gælder for vores estimator $\hat{\theta}$.

Det gælder derfor, at

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega_\theta), \quad \text{for } n \rightarrow \infty$$

I dette tilfælde ved vi, at 95% konfidensintervallet for θ_0 er $\hat{\theta} \pm 1,96 \cdot se(\hat{\theta})$

Formuleret på en anden måde:

$$P(\hat{\theta} - 1,96 \cdot se(\hat{\theta}) \leq \theta_0 \leq \hat{\theta} + 1,96 \cdot se(\hat{\theta})) = \Phi(1,96) - \Phi(-1,96) = 0,95$$

Baseret på oberseveret data er

$$se(\hat{\theta}) = \sqrt{0,0002} = 0,0147.$$

så vores konfidensinterval for θ_0 er:

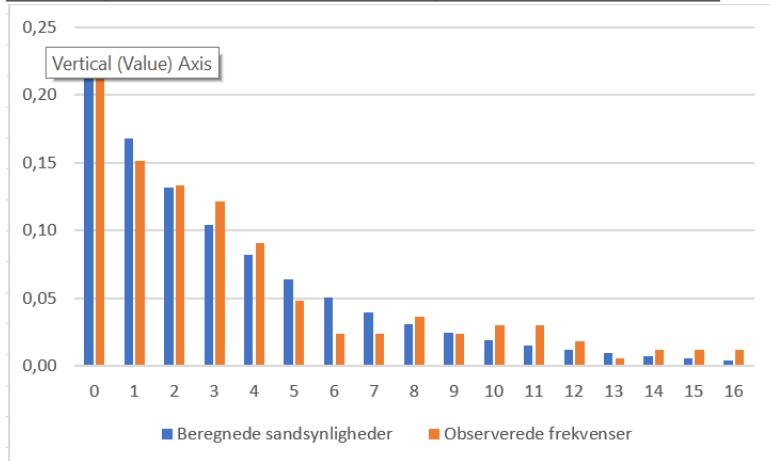
$$\{0,2132 - 1,96 \cdot 0,0147 \leq \theta_0 \leq 0,2132 + 1,96 \cdot 0,0147\} = \underline{\underline{\{0,1843 \leq \theta_0 \leq 0,2420\}}}$$

8. Som modelkontrol bedes du udregne de predikterede sandsynligheder

$$P(Y = y) = (1 - \hat{\theta})^y \hat{\theta}, \quad y \in \{0, 1, 2, 3, \dots\},$$

og sammenligne med de observerede frekvenser.

	Beregnet sandsynlighed	Observeret frekvens
y=0	0,2131	0,2242
y=1	0,1677	0,1515
y=2	0,1320	0,1333
y=3	0,1038	0,1212
y=4	0,0817	0,0909
y=5	0,0643	0,0485
y=6	0,0506	0,0242
y=7	0,0398	0,0242
y=8	0,0313	0,0364
y=9	0,0246	0,0242
y=10	0,0194	0,0303
y=11	0,0153	0,0303
y=12	0,0120	0,0182
y=13	0,0094	0,0061
y=14	0,0074	0,0121
y=15	0,0058	0,0121
y=16	0,0046	0,0121



Ser antagelsen om en geometrisk fordeling ud til at være rimelig?

- Den overordnede tendens lader til at passe godt med en geometrisk fordeling. Der er flest observationer ved $y = 0$ og derfra faldende sandsynlighed ved højere værdier af y . De observerede sandsynligheder falder dog ikke monoton som de teoretiske.

9. Gentag estimationen i STATA, hvor log-likelihood funktionen optimeres numerisk. Sammenlign med resultaterne ovenfor.

Output fra estimeringen:

```
. mlexp (uger*log(1-{theta})+log({theta})) , from(init)

Iteration 0:  log likelihood = -401.45168
Iteration 1:  log likelihood = -401.03919
Iteration 2:  log likelihood = -401.0383
Iteration 3:  log likelihood = -401.0383

Maximum likelihood estimation

Log likelihood = -401.0383                         Number of obs      = 165


```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	.2131783	.0147211	14.48	0.000	.1843256 .2420311

Opgave 3

Vi forsætter analysen ovenfor og vil undersøge om skolefraværet er forskelligt hos piger og drenge.

1. Formulér en statistisk model hvor sandsynligheden for sygdom for individ i afhænger af køn, dvs

$$Y_i | \text{dreng} \stackrel{d}{=} \text{Geometric}(\theta_i)$$

sådan at

$$\theta_i = \beta_1 + \beta_2 \cdot \text{dreng}_i.$$

Opskriv likelihood funktionen og parameterrummet

Hvad er fortolkningen af parametrene?

Vi skal nu tage højde for, at piger og drenge ikke nødvendigvis har samme sandsynlighed for fravær i en given uge.

Vi kan derfor derfor indføre en stokastisk variabel

$$X_i = \begin{cases} 0 & \text{hvis person } i \text{ er en pige} \\ 1 & \text{hvis person } i \text{ er en dreng} \end{cases}$$

og formulere en Likelihood-funktion som tillader sandsynlighedsparameteren at være *individ-specifik*:

$$L(\beta_1, \beta_2 | Y_1, \dots, Y_n, X_1, \dots, X_n) = (\beta_1 + X_i \beta_2) (1 - \beta_1 - X_i \beta_2)^{Y_i}$$

En simpelere model er, at antage ordnede data, så de første 77 observationer er piger og de følgende 88 observationer er drenge. I dette tilfælde kan vi dele vores sample op i to. 77 piger med sandsynlighed β_1 og 88 drenge med sandsynlighed $\beta_1 + \beta_2$.

Betegner vi antallet af piger med n_1 og antallet af drenge med n_2 kan vi opstille likelihood funktionen

$$\begin{aligned} L(\beta_1, \beta_2 | Y_1, \dots, Y_n) &= \prod_{i=1}^{n_1} \ell(\beta_1 | Y_i) \prod_{i=1}^{n_2} \ell(\beta_1, \beta_2 | Y_i) \\ &= \prod_{i=1}^{n_1} (1 - \beta_1)^{Y_i} \beta_1 \prod_{i=1}^{n_2} (1 - (\beta_1 + \beta_2))^{Y_i} (\beta_1 + \beta_2) \\ \Theta &= \left\{ \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \in \mathbb{R}^2 \mid 0 < \beta_1 < 1, 0 < \beta_1 + \beta_2 < 1 \right\} \end{aligned}$$

Som nævnt er β_1 sandsynligheden for fravær i en given uge for piger. β_2 er *tillægs-sandsynligheden* for drenge. Hvis der ikke er forskel på piger og drenge er β_2 lig nul. Hvis drenge har lavere sandsynlighed end piger er β_2 negativ og hvis drenge har større sandsynlighed (hvad vores deskriptive statistik i opg.2 indikerer) er β_2 positiv.

Ved ordnede data kan β_1 og $\beta_1 + \beta_2$ udledes analytisk som i opgave 2.

2. Modificér koden ovenfor og estimér parametrene, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

Output:

```
. mlexp (uger*log(1- ({beta1}+{beta2}*dreng))+log(({beta1}+{beta2}*dreng)), from (init)
Iteration 0: log likelihood = -409.12419
Iteration 1: log likelihood = -393.44564
Iteration 2: log likelihood = -393.30739
Iteration 3: log likelihood = -393.30716
Iteration 4: log likelihood = -393.30716

Maximum likelihood estimation
Log likelihood = -393.30716 Number of obs = 165

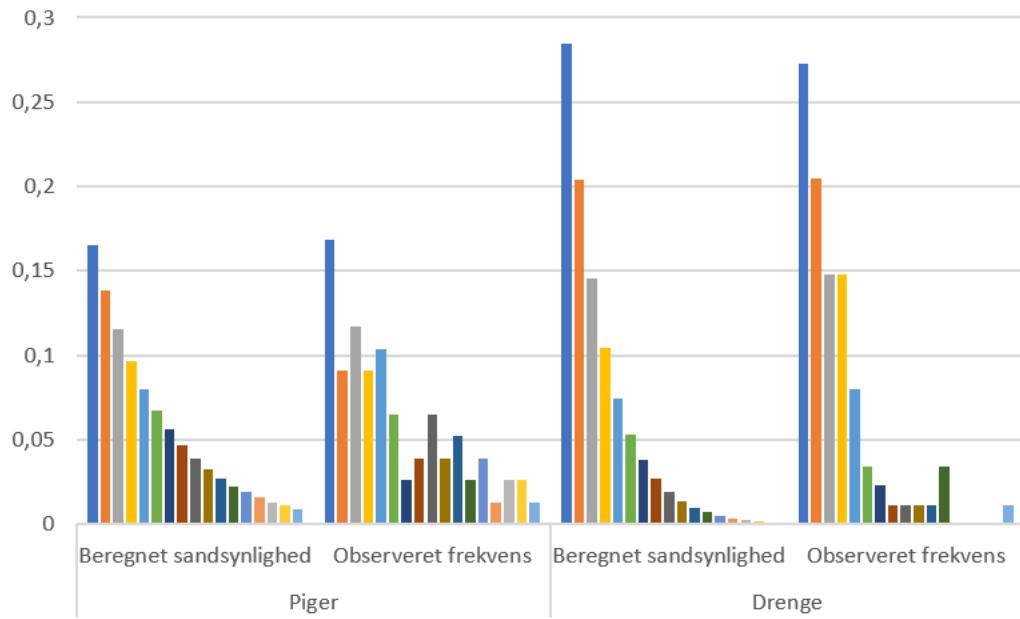

```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/beta1	.1655914	.0172378	9.61	0.000	.1318059 .1993769
/beta2	.1191982	.0309243	3.85	0.000	.0585877 .1798088

β_1 estimeres til 0,1656 og β_2 estimeres til 0,1192. Baseret på vores sample har piger altså en sandsynlighed for fravær på 16,56% i en given uge, mens sandsynligheden for drenge er 11,92% højere, 28,48%. z-værdien for β_2 er 3,85 og β_2 er dermed signifikant forskellig fra nul på et 99% konfidensniveau.

3. Udregn som modelkontrol de predikterede sandsynligheder for piger henholdsvis drenge, og sammenlign med de observerede frekvenser.

	Beregnet sandsynlighed		Observeret frekvens	
	piger	drenge	piger	drenge
y=0	0,1656	0,2848	0,1688	0,2727
y=1	0,1382	0,2037	0,0909	0,2045
y=2	0,1153	0,1457	0,1169	0,1477
y=3	0,0962	0,1042	0,0909	0,1477
y=4	0,0803	0,0745	0,1039	0,0796
y=5	0,0670	0,0533	0,0649	0,0341
y=6	0,0559	0,0381	0,0260	0,0227
y=7	0,0466	0,0273	0,0390	0,0113
y=8	0,0389	0,0195	0,0649	0,0113
y=9	0,0325	0,0139	0,0390	0,0113
y=10	0,0271	0,0100	0,0520	0,0113
y=11	0,0226	0,0071	0,0260	0,0341
y=12	0,0189	0,0051	0,0390	-
y=13	0,0157	0,0036	0,0130	-
y=14	0,0131	0,0026	0,0260	-
y=15	0,0110	0,0019	0,0260	-
y=16	0,0091	0,0013	0,0130	0,0113



Ser antagelsen om en geometrisk fordeling ud til at passe bedre i den udvidede model?

- Igen passer den overordnede tendens. Specielt for piger er den observerede frekvens dog ikke monoton faldende i y . Måske ville flere observationer ”glatte” observationsfrekvenserne ud, så de ligner den geometriske fordeling mere.

4. Formulér idéen om at drenge har en fraværs-sandsynlighed hver uge som er 10 procent højere end piger som en statistisk hypotese.

- Nul-hypotese, $\mathcal{H}_0 : \beta_2 = 0, 10$
- Alternativ-hypotese, $\mathcal{H}_A : \beta_2 \neq 0, 10$

Test hypotesen med et z -test og med et likelihood ratio test.

Fra output i delopgave 2 ved vi, at $\hat{\beta}_2 = 0,119197$ og $se(\hat{\beta}_2) = 0,03092$.

Test-statistik til z -test:

$$\begin{aligned} z_n(\beta_2 = 0, 10) &= \frac{\hat{\beta}_2 - 0, 1}{se(\hat{\beta}_2)} \\ &= \frac{0,119197 - 0, 1}{0,03092} \\ &\approx 0,62 \end{aligned}$$

Under antagelse af, at nul-hypotesen er sand, gælder

$$z_n(\beta_2 = 0, 10) \xrightarrow{d} N(0, 1)$$

og da er den kritiske værdi for 95% konfidensniveau $\pm 1,96$.

Test-statistikken ligger inden for den kritiske værdi og p-værdien er:

$$\begin{aligned} p_n(\beta_2 = 0, 10) &= 2 \cdot (1 - \Phi(0, 62)) \\ &= 0,5353 \end{aligned}$$

Vi kan derfor ikke på baggrund af et z -test afvise nulhypotesen.

For at foretage en likelihood ratio test skal vi bruge log-likelihood fra den urestrikterede model (hvor β_1 og β_2 kan vælges frit) og den restrikterede model ($\beta_2 = 0, 1$).

For den urestrikterede model definerer vi parameter-vektoren

$$\hat{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

og for den restrikterede model definerer vi parameter-vektoren

$$\tilde{\beta} = \begin{pmatrix} \beta_1 \\ 0, 1 \end{pmatrix}$$

Den urestrikterede log-likelihood har vi fra outputtet i delopgave 2:

$$L_n(\hat{\beta}) = -393,31$$

Den restrikterede loglikelihood må vi selv beregne ud fra likelihood funktionen:

$$\begin{aligned} L(\beta_1, \beta_2 | Y_1, \dots, Y_n) &= \prod_{i=1}^{n_1} (1 - \beta_1)^{Y_i} \beta_1 \prod_{i=1}^{n_2} (1 - (\beta_1 + \beta_2))^{Y_i} (\beta_1 + \beta_2) \Leftrightarrow \\ \log L(\beta_1, \beta_2 | Y_1, \dots, Y_n) &= \sum_{i=1}^{n_1} Y_i \log (1 - \beta_1) + n_1 \log \beta_1 \\ &+ \sum_{i=1}^{n_2} Y_i \log (1 - (\beta_1 + \beta_2)) + n_2 \log (\beta_1 + \beta_2) \end{aligned}$$

Vi indsætter data fra vores sample:

$$\begin{aligned} \log L(\tilde{\beta} | Y_1, \dots, Y_{165}) &= 388 \cdot \log(1 - 0, 165592) + 77 \cdot \log(0, 165592) \\ &\quad + 221 \cdot \log(1 - (0, 165592 + 0, 1)) + 88 \cdot \log(0, 165592 + 0, 1) \\ &\approx -393,5 \end{aligned}$$

Likelihood ratio test statistik:

$$\begin{aligned} LR_n(\beta_{20} = 0, 1) &= 2(\log L_n(\hat{\beta}_n) - \log \tilde{\beta}_n) \\ &= 2(-393,31 - (-393,5)) \\ &= 0,38 \end{aligned}$$

Under antagelse af, at nul-hypotesen er sand, gælder

$$LR_n(\beta_{20} = 0, 1) \stackrel{d}{\sim} \chi^2(1)$$

og da er den kritiske værdi for 95% konfidensniveau 3,84.

Da vores test-statistik ligger under den kritiske værdi kan vi heller ikke ved likelihood ratio test afvise nulhypotesen.

Opgave 4

Betrægt situationen hvor vi er ved at planlægge en spørgeskemaundersøgelse om folks TV-vaner, specielt hvor stor en del af seerne der bager mere derhjemme efter at have set programmet Den Store Bagedyst i fjernsynet. Vi vil gerne have et præcist resultat, hvilket taler for at spørge så mange som muligt, men spørgeskema-proceduren er dyr og vi er blevet bedt om at spare, hvilket taler for at spørge så få som muligt. Spørgeskemaet er enkelt, med svarene nej eller ja, og når undersøgelsen er gennemført har vi n observationer, $\{y_i\}_{i=1}^n$ med $y_i \in \{0, 1\}$.

Vores procedure til at udvælge seere garanterer, at de er identiske og uafhængigt fordelt.

- Angiv en statistisk model til at undersøge hvor stor sandsynligheden er for at seere bager mere derhjemme efter at have set Den Store Bagedyst. Kald sandsynligheden p

Eftersom udfalsrummet er binært, $y_i \in \{0, 1\}$, er den rigtige statistiske model en *Bernoulli-fordeling*. Observationerne y_i er da realisationer af en stokastisk variabel Y_i , og

$$Y_i \stackrel{d}{=} \text{Bernoulli}(\theta).$$

Dermed antager vi, udover identisk og uafhængig fordeling, at den stokastiske variabel Y_i er beskrevet ved sandsynlighedsfunktionen

$$f_{Y_i}(y|p) = p^y (1-p)^{1-y}$$

- Find maksimum likelihood estimatoren, \hat{p}_n , for sandsynligheden p .

Likelihoodfunktionen $L(p|Y_1, \dots, Y_n)$ er produktet af alle likelihood bidragene $\ell(p|Y_i)$:

$$\begin{aligned} L(p|Y_1, \dots, Y_n) &= \prod_{i=1}^n \ell(p|Y_i) \\ &= \prod_{i=1}^n p^{Y_i} (1-p)^{1-Y_i} \end{aligned}$$

Log-likelihood funktionen:

$$\begin{aligned}
 \log L(p|Y_1, \dots, Y_n) &= \log \prod_{i=1}^n \ell(p|Y_i) \\
 &= \log \prod_{i=1}^n p^{Y_i} (1-p)^{1-Y_i} \\
 &= \sum_{i=1}^n \left(Y_i \log(p) + (1-Y_i) \log(1-p) \right)
 \end{aligned}$$

Scorebidraget s_i der den første afledte af likelihood bidraget:

$$\begin{aligned}
 s_i(p) &= \frac{\partial \ell(p|Y_i)}{\partial p} \\
 &= \frac{Y_i}{p} - \frac{1-Y_i}{1-p}
 \end{aligned}$$

Scoren S er den første afledte af likelihood funktionen og lig med summen af alle score bidragene:

$$\begin{aligned}
 S(p) &= \frac{\partial L(p|Y_1, \dots, Y_n)}{\partial p} \\
 &= \sum_{i=1}^n \frac{\partial \ell(p|Y_i)}{\partial p} \\
 &= \sum_{i=1}^n \left(\frac{Y_i}{p} - \frac{1-Y_i}{1-p} \right) \\
 &= \frac{\sum_{i=1}^n Y_i}{p} - \frac{n - \sum_{i=1}^n Y_i}{1-p}
 \end{aligned}$$

Maksimum likelihood estimatoren, \hat{p}_n , er løsningen til førsteordensbetingelsen:

$$\begin{aligned}
 S(p) = 0 &\Leftrightarrow \\
 \frac{\sum_{i=1}^n Y_i}{\hat{p}_n} &= \frac{n - \sum_{i=1}^n Y_i}{1 - \hat{p}_n} \Leftrightarrow \\
 \frac{1 - \hat{p}_n}{\hat{p}_n} &= \frac{n - \sum_{i=1}^n Y_i}{\sum_{i=1}^n Y_i} \Leftrightarrow \\
 \frac{1}{\hat{p}_n} - 1 &= \frac{n}{\sum_{i=1}^n Y_i} - 1 \Leftrightarrow \\
 \hat{p}_n &= \frac{\sum_{i=1}^n Y_i}{n}
 \end{aligned}$$

3. Lad den sande andel der bager mere efter at have set Den Store Bagedyst være $p_0 = 0,4$. Vi vil indrette vores undersøgelse, sådan at estimatoren, \hat{p}_n , med 99% sikkerhed kun afviger $d = 0,04$ fra den sande værdi p_0 .

Hvor mange seere skal vi spørge for at opnå dette?

Den kritiske værdi for en dobbeltsidet test af en standard normalfordelt variabel på et 99% konfidenzniveau er 2,58 (brug ”NORM.S.INV” i EXCEL).

Kravet om, at \hat{p}_n kun må afvige med 0,04 fra p_0 svarer til at kræve:

$$\begin{aligned}
 se(\hat{p}_n) \cdot 2,58 &\leq 0,04 \Leftrightarrow \\
 se(\hat{p}_n) &\leq \frac{0,04}{2,58}.
 \end{aligned}$$

(Husk metoden til at konstruere konfidens-intervaller).

Antallet af adspurgte seere, n , indgår i beregningen af variansen af \hat{p}_n , og dermed også i udtrykket for $se(\hat{p}_n)$.

Vi ved fra opgave 2,6 at variansen af estimatoren er givet ved

$$Var(\hat{p}_n(Y_1, \dots, Y_n)) = \frac{I(p_0)^{-1}}{n}$$

hvor informationen er givet ved

$$I(p_0) = -E(H_i(p_0))$$

Hesse bidraget H_i er den anden afledte af log-likelihood bidraget $\log \ell$, eller den første afledte af score bidraget s_i :

$$\begin{aligned} H_i(p_0) &= \frac{\partial^2 \log \ell(p_0 | Y_i)}{\partial p_0 \partial p_0} \\ &= \frac{\partial s(p_0 | Y_i)}{\partial p_0} \\ &= \frac{\partial}{\partial p_0} \left(\frac{Y_i}{p_0} - \frac{1 - Y_i}{1 - p_0} \right) \\ &= -\frac{Y_i}{p_0^2} - \frac{1 - Y_i}{(1 - p_0)^2} \end{aligned}$$

Informationen er da givet ved:

$$\begin{aligned} I(p_0) &= -E(H_i(p_0)) \\ &= E \left(\frac{Y_i}{p_0^2} + \frac{1 - Y_i}{(1 - p_0)^2} \right) \\ &= \frac{E(Y_i)}{p_0^2} + \frac{1 - E(Y_i)}{(1 - p_0)^2} \\ &= \frac{p_0}{p_0^2} + \frac{1 - p_0}{(1 - p_0)^2} \end{aligned}$$

hvor vi i sidste linie bruger middelværdien i en Bernoulli-fordeling.

Vi sætter nu på fælles brøkstreg for at kunne invertere:

$$\begin{aligned} I(p_0) &= \frac{p_0}{p_0^2} + \frac{1 - p_0}{(1 - p_0)^2} \\ &= \frac{1}{p_0} + \frac{1}{1 - p_0} \\ &= \frac{1 - p_0}{p_0(1 - p_0)} + \frac{p_0}{p_0(1 - p_0)} \\ &= \frac{1}{p_0(1 - p_0)} \end{aligned}$$

Og ud fra dette finder vi variansen af \hat{p}_n :

$$\begin{aligned} Var(\hat{p}_n) &= \frac{I(p_0)^{-1}}{n} \\ &= \frac{p_0(1 - p_0)}{n} \end{aligned}$$

Læg mærke til at $p_0(1-p_0)$, variansen af hver enkelt observation, er lig med variansen i en binomialfordeling med antalsparameter 1, hvilket er identisk med en Bernoulli-fordeling.

I opgave 2,5 approksimerede vi $I(p_0)$ med $I(\hat{p}_n)$ da vi ikke kendte p_0 . Det behøver vi ikke her.

Standardfejlen for \hat{p}_n er:

$$\begin{aligned} se(\hat{p}_n) &= \sqrt{Var(\hat{p}_n)} \\ &= \sqrt{\frac{p_0(1-p_0)}{n}} \\ &= \sqrt{\frac{0,4(1-0,4)}{n}} \\ &= \frac{\sqrt{0,24}}{\sqrt{n}} \end{aligned}$$

Vi kan nu finde den værdi af n der sikrer at estimatoren, \hat{p}_n , med 99% sikkerhed kun afviger $d = 0,04$ fra den sande værdi p_0 :

$$\begin{aligned} se(\hat{p}_n) \leq \frac{0,04}{2,58} &\Leftrightarrow \\ \frac{\sqrt{0,24}}{\sqrt{n}} \leq \frac{0,04}{2,58} &\Leftrightarrow \\ n \geq 0,24 \cdot \left(\frac{2,58}{0,04}\right)^2 &\approx \underline{\underline{998,5}} \end{aligned}$$

Vi skal altså spørge mindst 999 seere for at få det ønskede kondidens-interval.

4. Hvor meget større skal n være, hvis vi kræver den dobbelte præcision, $d = 0,02$?

Vi erstatter 0,04 med 0,02 i udregningen ovenfor:

$$n \geq 0,24 \cdot \left(\frac{2,58}{0,02}\right)^2 \approx \underline{\underline{3994}}$$

For at indsætte konfindens-intervallet med 50% skal vi altså øge antal adspurgt til det 4-dobbelte.

Hvorfor er det ikke bare dobbelt så mange?

Det er \sqrt{n} der indgår beregningen af standardfejlen og dermed i beregningen af konfindensintervallet. Dermed kommer n til at afhænge kvadratisk af den ønskede bredde af konfindensintervallet. Dette ses i udtrykket ovenfor, hvor n er en funktion af den maksimale afvigelse kvadreret.

5. Skriv en generel løsning for den stikprøve-størrelse, n , der sikrer at afstanden til p_0 med sandsynligheden β er højst d , dvs.

$$P(|\hat{p}_n - p_0| \leq d) = \beta.$$

Hvis vi benytter at:

$$\begin{aligned} P(|\hat{p}_n - p_0| \leq d) &= P(0 \leq \hat{p}_n - p_0 \leq d) \cup P(-d \leq \hat{p}_n - p_0 < 0) \\ &= P(0 \leq \hat{p}_n - p_0 \leq d) + P(-d \leq \hat{p}_n - p_0 < 0), \end{aligned}$$

og

$$\begin{aligned} \frac{\hat{p}_n - p_0}{se(\hat{p}_n)} &\sim N(0, 1) \Leftrightarrow P\left(\frac{\hat{p}_n - p_0}{se(\hat{p}_n)} \leq d\right) = \Phi(d) \\ &\wedge P\left(\frac{\hat{p}_n - p_0}{se(\hat{p}_n)} \leq -d\right) = \Phi(-d) = 1 - \Phi(d), \end{aligned}$$

kan vi lave følgende omregning:

$$\begin{aligned} P(|\hat{p}_n - p_0| \leq d) &= \beta \Leftrightarrow \\ P(0 \leq \hat{p}_n - p_0 \leq d) &\cup P(-d \leq \hat{p}_n - p_0 < 0) = \beta \Leftrightarrow \\ P\left(0 \leq \frac{\hat{p}_n - p_0}{se(\hat{p}_n)} \leq \frac{d}{se(\hat{p}_n)}\right) &\cup P\left(\frac{-d}{se(\hat{p}_n)} \leq \frac{\hat{p}_n - p_0}{se(\hat{p}_n)} < 0\right) = \beta \Leftrightarrow \\ \Phi\left(\frac{d}{se(\hat{p}_n)}\right) - \Phi(0) &+ \Phi(0) - \Phi\left(-\frac{d}{se(\hat{p}_n)}\right) = \beta \Leftrightarrow \\ \Phi\left(\frac{d}{se(\hat{p}_n)}\right) - \Phi(0) &+ \Phi(0) - \left(1 - \Phi\left(\frac{d}{se(\hat{p}_n)}\right)\right) = \beta \Leftrightarrow \\ 2\Phi\left(\frac{d}{se(\hat{p}_n)}\right) - 1 &= \beta \Leftrightarrow \\ \frac{d}{se(\hat{p}_n)} &= \Phi^{-1}\left(\frac{1+\beta}{2}\right) \Leftrightarrow \\ d &= se(\hat{p}_n) \cdot \Phi^{-1}\left(\frac{1+\beta}{2}\right) \end{aligned}$$

Hvis $\beta = 0,95$ ved vi, at $d = 1,96 \cdot se(\hat{p}_n) = \Phi^{-1}\left(\frac{1+\beta}{2}\right) \cdot se(\hat{p}_n)$.

Vi ved at n indgår i udtrykket for $se(\hat{p}_n)$:

$$se(\hat{p}_n) = \sqrt{\frac{p_0(1-p_0)}{n}}$$

så vi kan nu skrive en generel formel op:

$$\begin{aligned}
 P(|\hat{p}_n - p_0| \leq d) = \beta &\Leftrightarrow \\
 \Phi^{-1} \left(\frac{1+\beta}{2} \right) \cdot se(\hat{p}_n) = d &\Leftrightarrow \\
 \sqrt{\frac{p_0 (1-p_0)}{n}} = \frac{d}{\Phi^{-1} \left(\frac{1+\beta}{2} \right)} &\Leftrightarrow \\
 \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{p_0 (1-p_0)}} \cdot \frac{d}{\Phi^{-1} \left(\frac{1+\beta}{2} \right)} &\Leftrightarrow \\
 n = \left(\frac{p_0 (1-p_0)}{d^2} \right) \left(\Phi^{-1} \left(\frac{1+\beta}{2} \right) \right)^2 &
 \end{aligned}$$

6. Meningsmålingsinstitutter spørger ofte ca. $n = 1000$ personer i sådanne undersøgelser. Hvilken præcision, d , svarer det med sandsynlighed $\beta = 0,99$ til?

Vi omskriver resultatet fra sidste opgave for at isolere d :

$$\begin{aligned}
 n = \left(\frac{p_0 (1-p_0)}{d^2} \right) \left(\Phi^{-1} \left(\frac{1+\beta}{2} \right) \right)^2 &\Leftrightarrow \\
 d^2 = \left(\frac{p_0 (1-p_0)}{n} \right) \left(\Phi^{-1} \left(\frac{1+\beta}{2} \right) \right)^2 &\Leftrightarrow \\
 d = \left(\frac{p_0 (1-p_0)}{n} \right)^{1/2} \Phi^{-1} \left(\frac{1+\beta}{2} \right) &
 \end{aligned}$$

og indsætter de givne værdier:

$$\begin{aligned}
 d &= \left(\frac{0,4 (1-0,4)}{1000} \right)^{1/2} \Phi^{-1} \left(\frac{1+0,99}{2} \right) \\
 &= \left(\frac{0,24}{1000} \right)^{1/2} \cdot 2,58 \approx \underline{\underline{0,0399}}
 \end{aligned}$$

hvis man spørger 1000 personer får man altså en præcision på $d = 0,0399$. Dette er ikke overraskende hvis vi kigger på resultater fra opgave 4,3.

Opgave 5

Vi betragter følgende tabel for $n = 111$ individers indtag af kaffe om dagen og information om rygevaner:

Kaffe	Rygning		Sum	
	Nej ($x = 1$)	Ja ($x = 2$)		
0	($y = 1$)	17	10	27
1-5 kopper	($y = 2$)	43	27	70
Mere	($y = 3$)	3	11	14
Sum		63	48	111

For individ i , $i = 1, 2, \dots, n$, repræsenterer vi det daglige kaffe-indtag med den stokastiske variabel $Y_i \in \{1, 2, 3\}$ og rygevanerne med $X_i \in \{1, 2\}$, så vi samlet betragter en to-dimensional stokastisk variabel,

$$\{Z_i\}_{i=1}^n = \{(Y_i, X_i)\}_{i=1}^n, \quad Z_i \in \{1, 2, 3\} \times \{1, 2\}$$

Der er således 6 mulige udfald i tabellen,

$$Z_i = (Y_i, X_i) \in \mathbb{Z} = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\}$$

Vi antager i udgangspunktet at de stokastiske variable, $\{Z_i\}_{i=1}^n$, er uafhængigt og identisk fordele, og som statistisk model foreslår vi en generel diskret fordeling, med sandsynlighedsfunktion givet ved sandsynlighederne,

$$P(Z_i = (y, x)) = p_{ab} \quad \text{hvis } (y, x) = (a, b)$$

sådan, at $0 < p_{ab} < 1$ og

$$\sum_{y=1}^3 \sum_{x=1}^2 p_{yx} = 1$$

1. Vis, at sandsynlighederne under ét kan skrives som,

$$f_{Z_i}(z|p_{11}, p_{12}, p_{21}, p_{22}, p_{31}, p_{32}) = p_{11}^{\mathbb{I}(z=(1,1))} \cdot p_{12}^{\mathbb{I}(z=(1,2))} \cdot p_{21}^{\mathbb{I}(z=(2,1))} \cdot p_{22}^{\mathbb{I}(z=(2,2))} \cdot p_{31}^{\mathbb{I}(z=(3,1))} \cdot p_{32}^{\mathbb{I}(z=(3,2))},$$

for $z \in \mathbb{Z}$ og hvor $\mathbb{I}(\cdot)$ som sædvanligt er indikatorfunktionen.

Hvis vi for eksempel har udfaldet $z = (1, 2)$ vil indikatorfunktioner gøre at $f_{Z_i} = p_{12}$, da $\mathbb{I}(z = (1, 2)) = 1$ og $\mathbb{I}(z = (x, y)) = 0$ for $(x, y) \neq (1, 2)$.

2. Da sandsynlighederne summerer til én lader vi $\theta = (p_{11}, p_{12}, p_{21}, p_{22}, p_{31}, p_{32})'$ sådan at $p_{32} = 1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}$.

Opskriv parameterrummet, Θ , sådan at $\theta \in \Theta$.

Parameterrummet:

$$\theta \in \Theta = \{\theta \in \mathbb{R}^5 | 0 < \theta_i < 1 \wedge \sum_{i=1}^5 \theta_i \leq 1, i = \{1, 2, \dots, 5\}\}$$

Vis, at sample log-likelihood funktionen kan skrives på formen

$$\begin{aligned} \log(L(\theta|z_1, \dots, z_n)) &= s_{11} \log(p_{11}) + s_{12} \log(p_{12}) + s_{21} \log(p_{21}) \\ &\quad + s_{22} \log(p_{22}) + s_{31} \log(p_{31}) \\ &\quad + (n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}) \log(1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}) \end{aligned}$$

hvor $s_{ab} = \sum_{i=1}^n \mathbb{I}(Z_i = (a, b))$

s_{11} angiver hvor mange gange udfaldet $z = (1, 1)$ er blevet observeret i samplen.

Likelihood funktionen som er produktet af sandsynlighederne af alle udfald er for sample-størrelse n :

$$L(\theta|z_1, \dots, z_n) = p_{11}^{s_{11}} \cdot p_{12}^{s_{12}} \cdot p_{21}^{s_{21}} \cdot p_{22}^{s_{22}} \cdot p_{31}^{s_{31}} \cdot (1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31})^{(n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31})}$$

Tager vi log af L får vi udtrykket ovenfor.

3. Find førsteordensbetingelserne for et maksimum

FOC:

$$\frac{\partial \log(L(\theta|z_1, \dots, z_n))}{\partial \theta} = \begin{pmatrix} \frac{s_{11}}{p_{11}} - \frac{n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}} \\ \frac{s_{12}}{p_{12}} - \frac{n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}} \\ \frac{s_{21}}{p_{21}} - \frac{n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}} \\ \frac{s_{22}}{p_{22}} - \frac{n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}} \\ \frac{s_{31}}{p_{31}} - \frac{n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \Leftrightarrow \begin{aligned} \frac{s_{11}}{p_{11}} &= \frac{s_{12}}{p_{12}} = \frac{s_{21}}{p_{21}} = \frac{s_{22}}{p_{22}} = \frac{s_{31}}{p_{31}} = \frac{n - s_{11} - s_{12} - s_{21} - s_{22} - s_{31}}{1 - p_{11} - p_{12} - p_{21} - p_{22} - p_{31}} \end{aligned}$$

Dette kan løses hvis man har (meget) tid.

4. Husk fra opgave 2 på ugeseddel 48, at maksimum likelihood estimatorerne i denne model er

$$\hat{p}_{11} = \frac{s_{11}}{n}, \quad \hat{p}_{12} = \frac{s_{12}}{n}, \quad \dots, \hat{p}_{31} = \frac{s_{31}}{n}$$

svarende til de relative frekvenser.

Find estimatorerne pba. tabellen ovenfor.

$$\begin{aligned} \hat{p}_{11} &= \frac{17}{111} = 0,15, & \hat{p}_{12} &= \frac{10}{111} = 0,09, & \hat{p}_{21} &= \frac{43}{111} = 0,39 \\ \hat{p}_{22} &= \frac{27}{111} = 0,24, & \hat{p}_{31} &= \frac{3}{111} = 0,03, & \hat{p}_{32} &= \frac{11}{111} = 0,10 \end{aligned}$$

Udregn den maksimale værdi af log-likelihood funktionen

Den maksimale værdi likelihood værdi fås når vi indsætter vores estimatorer i $\log L$:

$$\begin{aligned} \max_{\theta \in \Theta} \log L(\theta|z_1, \dots, z_n) &= 17 \cdot \log(0,15) + 10 \cdot \log(0,09) + 43 \cdot \log(0,39) \\ &\quad + 27 \cdot \log(0,24) + 3 \cdot \log(0,03) + 11 \cdot \log(0,10) \\ &= \underline{\underline{-171,2}} \end{aligned}$$

5. Betragt koden nedenfor og forklar hvad der foregår.

Brug data filen coffee_smoking sammen med STATA koden herunder til at estimere parametrene i modellen.

```

g i11 = (Y==1)*(X==1)
g i21 = (Y==2)*(X==1)
g i31 = (Y==3)*(X==1)
g i12 = (Y==1)*(X==2)
g i22 = (Y==2)*(X==2)
g i32 = (Y==3)*(X==2)

mat define init = J(1,5,0.1)

mlexp ( /**
    i11*log({p11}) + i21*log({p21}) + i31*log({p31}) /**
    + i12*log({p12}) + i22*log({p22}) /**
    + i32*log(1-{p11}-{p21}-{p31}-{p12}-{p22}) /**
    ), from(init)

```

Første del af koden genererer dummy-variable for de simultane udfald af Y og X .

Opgave 6

Fortsæt analysen ovenfor, og betragt tilfældet med sandsynligheder givet ved

Kaffe	Rygning		Sum
	Nej	Ja	
0	p_{11}	p_{12}	$p_{1\cdot}$
1-5 kopper	p_{21}	p_{22}	$p_{2\cdot}$
Mere	p_{31}	p_{32}	$p_{3\cdot}$
Sum	$p_{\cdot 1}$	$p_{\cdot 2}$	1

hvor notationen p_x og p_y angiver de marginale sandsynligheder.

1. Argumenter for, at kaffeforbrug og rygevaner er stokastisk uafhængige, hvis det gælder at

$$p_{yx} = p_y \cdot p_x$$

Dette følger af definitionen på uafhængighed (Se ”Sørensen, side 58”).

2. Modificér den stokastiske model ovenfor sådan at kaffeindtag og rygevaner er stokastisk uafhængige.

Før kunne vi skrive de simultane sandsynligheder således:

$$P(Z_i = (y, x)) = p_{ab} \quad \text{hvis } (y, x) = (a, b)$$

Nu ved vi, at de simultane sandsynligheder er produkter af de marginale sandsynligheder:

$$P(Z_i = (y, x)) = \begin{cases} p_{(1\cdot)}p_{(\cdot 1)} \\ p_{(1\cdot)}p_{(\cdot 2)} \\ p_{(2\cdot)}p_{(\cdot 1)} \\ p_{(2\cdot)}p_{(\cdot 2)} \\ p_{(3\cdot)}p_{(\cdot 1)} \\ p_{(3\cdot)}p_{(\cdot 2)} \end{cases}$$

Så der er 5 parametre $(p_{(1\cdot)}, p_{(2\cdot)}, p_{(3\cdot)}, p_{(\cdot 1)}, p_{(\cdot 2)})$.

Af disse er kun 3 frie, da $p_{(3\cdot)} = 1 - p_{(1\cdot)} - p_{(2\cdot)}$ og $p_{(\cdot 2)} = 1 - p_{(\cdot 1)}$.

3. *Modificér koden ovenfor og estimér parametrene under antagelse af uafhængighed.*

Estimaterne bliver:

$$\begin{aligned} p_{(1\cdot)} &= 0,24 \\ p_{(2\cdot)} &= 0,63 \\ p_{(\cdot 1)} &= 0,57 \end{aligned}$$

Find også den maksimale værdi af log-likelihood funktionen under restriktionen

$$\log L(\tilde{\theta}|z_1, \dots, z_2)$$

Log-likelihood = -175,35

4. *Udregn likelihood ratio test-størrelsen for hypotesen on uafhængighed,*

$$\mathcal{H}_0 : Y_i \text{ og } X_i \text{ er uafhængige.}$$

$$\begin{aligned} LR_n &= 2(\log L(\theta) - \hat{\log L}(\tilde{\theta})) \\ &= 2(-171,2 - (-175,35)) \\ &= 8,36 \end{aligned}$$

Angiv den asymptotiske fordeling, antallet af frihedsgrader, den kritiske værdi og p-værdien.

Hvis \mathcal{H}_0 er sand så er likelihood ratio test-statistikken er $\chi^2(v)$ -fordelt, hvor antallet af frihedsgrader v er lig med antallet af parameterrestriktioner i den restrikterede model.

Da antallet af parametre i den urestrikterede model er fem og antallet i den restrikterede model er tre, gælder:

$$LR_n \sim \chi^2(2)$$

Den kritiske værdi er 95%-fraktilen og fås i excel ved at taste CHI2.INV(0,95;2)=5,99.

Da test-statistikken er over den kritiske værdi forkaster vi nulhypotesen. Der lader ikke til at være uafhængighed mellem kaffe og rygning.

p-værdien er $1 - (CHI2.FORDELING(8, 36; 2; 1)) \approx 0,015$. Dette tal betegner sandsynligheden for at få 8,36 eller højere i en $\chi^2(2)$ -fordeling, hvis nulhypotesen er sand.

5. Forklar hvorfor den statistiske model ovenfor ikke kan fortolkes som en betinget model for kaffeindtag givet rygning, $Y_i|X_i$.

En betinget model kræver *eksogenitet*, dvs. at der er en klar kausal sammenhæng fra den ene hændelse til den anden. Den ene tilstand skal så at sige være bestemt før den anden.

I eksemplet med forskelle mellem køn i overlevelseschance om bord på Titanic var dette selvfølgelig opfyldt, da overlevelse ikke influerer hvilket køn man har, men ens køn i høj grad influerede ens overlevelse.

I eksemplet med kaffe og rygning kan man ikke lave en entydig bestemmelse af den kausale rækkefølge. Vi har blot statistik på en række simultane hændelser.