Løsningsskisse eksamen
TMA4240 Statistikk, 20 desember 2024

Oppgave 1 Pinnekjøtt

a) Fra tabell

$$P(X \leq 3) = F(3) = 0.647.$$

(Alternativt kan man summere sannsynligheter oppgitt i spørsmål for $x = 0, 1, 2$ og 3.)

$$P(X \leq 3 | X > 0) = \frac{P(X \leq 3 \cap X > 0)}{P(X > 0)} = \frac{F(3) - F(0)}{1 - F(0)} = \frac{0.647 - 0.0498}{1 - 0.0498} = 0.63$$

b) For $y = 1, \ldots$:

$$f_Y(y) = P(X = y | X > 0) = \frac{P(X = y \cap X > 0)}{P(X > 0)} = \frac{P(X = y)}{1 - F(0)} = \frac{\mu^y e^{-\mu}/y!}{(1 - e^{-\mu})}$$

Her er $F(0) = P(X = 0) = \frac{\mu^0 e^{-\mu}}{0!} = e^{-\mu}$.

$$E(Y) = \sum_{y=1}^{\infty} y f_Y(y) = \sum_{y=1}^{\infty} y \frac{\mu^y e^{-\mu}/y!}{(1 - e^{-\mu})} = \frac{1}{(1 - e^{-\mu})} \mu \sum_{z=0}^{\infty} \mu^z e^{-\mu}/z! = \frac{\mu}{(1 - e^{-\mu})}.$$

der $z = y - 1$ og vi gjenkjenner summen av sannsynligheter over alle utfall i en Poissonfordeling: $\sum_{z=0}^{\infty} \mu^z e^{-\mu}/z! = 1$.

Her blir da $E(Y) = 3/(1 - 0.0498) = 3.16$.

c) Likelihood er

$$L(\mu) = \prod_{i=1}^{21} f(y_i; \mu) = \prod_{i=1}^{21} \frac{\mu^{y_i} e^{-\mu}}{y_i!(1 - e^{-\mu})}$$

Log-likelihood er

$$\begin{aligned} l(\mu) = \ln L(\mu) &= \sum_{i=1}^{21} \ln f(y_i; \mu) = \sum_{i=1}^{21} (y_i \ln \mu - \mu - \ln y_i! - \ln(1 - e^{-\mu})) \\ &= -21(\mu + \ln(1 - e^{-\mu})) + \sum_{i=1}^{21} (y_i \ln \mu - \ln y_i!) \end{aligned}$$

Ved Taylorutvikling har vi approksimasjon

$$\hat{l}(\mu) = -21(\mu + \ln(1 - e^{-3}) + \frac{e^{-3}}{1 - e^{-3}}(\mu - 3)) + \sum_{i=1}^{21}(y_i \ln \mu - \ln y_i!)$$

Derivering gir

$$\hat{l}'(\mu) = -21(1 + \frac{e^{-3}}{1 - e^{-3}}) + \sum_{i=1}^{21} \frac{y_i}{\mu}$$

Løsning er gitt ved $\hat{l}'(\hat{\mu}) = 0$, som gir

$$\hat{\mu} = \frac{\sum_{i=1}^{21} y_i}{21(1 + \frac{e^{-3}}{1-e^{-3}})}$$

Innsetting av tall gir estimatet $\hat{\mu} = 74/(21(1 + 0.0498/(1 - 0.0498))) = 3.35$.

(Utdypende svar, ikke krevd: Merk at en Taylorutvikling nok en gang, om 3.35, gir $\hat{\mu} = 3.40$, og videre iterering gir 3.406, 3.40691, 3.40702, etc. Følgen konvergerer nokså raskt til 3 desimalers nøyaktighet.)

Det observerte antallet O_i i hver søyle $i = 1, 2, \ldots$ av histogrammet er: $O_1 = 2, O_2 = 4, O_3 = 6, O_4 = 3, O_5 = 3, O_6 = 2, O_7 = 1, O_8 = O_9 = \ldots = 0$.

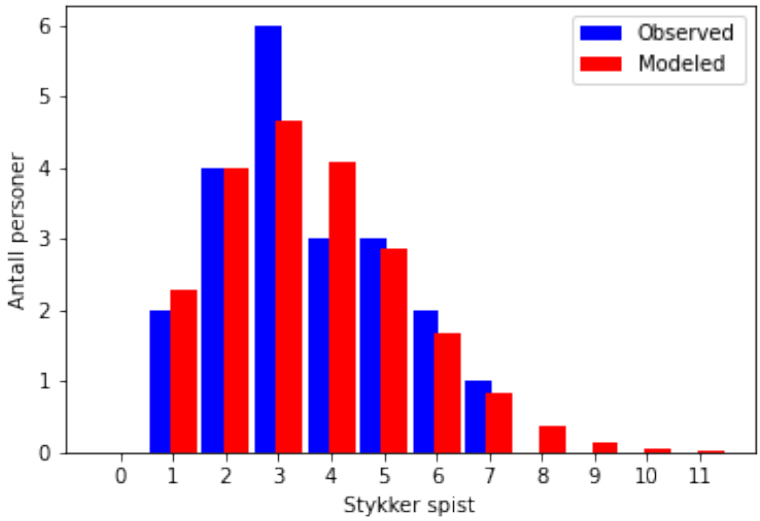
Hva sier modellen? Om vi ser på ett utfall i, er antallet observasjoner lik det utfallet binomisk fordelt med 21 forsøk og $p = f_Y(i)$ som suksesssannsynlighet. Da er forventet antall $E_i = 21 f_Y(i)$.

Dersom vi setter $\mu = 3.5$ (tabellert verdi nærmest 3.35) og deler på $(1-e^{-3.5})$ i punktsannsynligheter så får vi $f_Y(y)$, $y = 1, 2, \ldots$. Vi har $f_Y(1) = 0.11$, $f_Y(2) = 0.19$, $f_Y(3) = 0.22$, $f_Y(4) = 0.19$, $f_Y(5) = 0.14$, $f_Y(6) = 0.08$, $f_Y(7) = 0.04$, $f_Y(8) = 0.02$, $f_Y(9) = 0.007$, etc. Da blir forventet antall observasjoner i hvert utfall lik: $E_1 = 2.3$, $E_2 = 4.0$, $E_3 = 4.7$, $E_4 = 4.1$, $E_5 = 2.9$, $E_6 = 1.7$, $E_7 = 0.8$, $E_8 = 0.4$, $E_9 = 0.14$, $E_{10} = 0.04$, etc.

Tallene basert på modellen ser ut til å følge modellen nokså bra. Tallmaterialet er nokså lavt - kun 21 gjester. Avvik kan trolig forklares med tilfeldig variasjon.

(Utdypende svar, ikke krevd: For å undersøke dette nærmere, kan vi se på variansen i modell: Variansen i en binomisk fordeling er $21 f_Y(i)[1 - f_Y(i)]$. For utfall $i = 3$ vil standardavviket være $\sqrt{21 \cdot 0.22(1 - 0.22)} = 1.9$. Det er da ikke veldig usannsynlig med 6 observerte mens modell forventer 4.7. Det samme gjelder for andre enkeltutfall.)

Figur 1 viser forventet antall i hvert utfall (rød) og observasjoner (blå).



Figur 1: Observerte antall i hvert utfall (blå) og antall modell forventer i hvert utfall (rød).

(Utdypende svar, ikke krevd: Resultat for $\mu = 3$ (lineariseringspunkt) er nokså likt, men gir litt flytting av sannsynlighet mot lavere utfall. Avvik mellom de observerte utfall i histogram og det som modellen predikerer er mindre for 3.5 enn for 3.)

Oppgave 2 Lyspærer

a)

$$P(X < 3) = \int_0^3 e^{-x} dx = [-e^{-x}]_0^3 = 1 - e^{-3} = 1 - 0.0498 = 0.95$$

$$P(X < m) = \int_m^\infty e^{-x} dx = e^{-m} = 0.5, \quad -m = \ln(0.5), \quad m = \ln 2 = 0.69$$

b) Y er den største verdien av X_1, X_2 og X_3. Dersom denne er mindre enn y, så må alle enkeltvariable være mindre:

$$Y = \max\{X_1, X_2, X_3\} \leq y \rightarrow X_1 \leq y \cap X_2 < y \cap X_3 < y$$

Ved uavhengighet:

$$P(Y < y) = P(X_1 < y \cap X_2 < y \cap X_3 < y) = P(X_1 < y)P(X_2 < y)P(X_3 < y).$$

De tre er identisk fordelte og eksponensialfordelte: $P(X_i < y) = 1 - e^{-y}$ for $i = 1, 2, 3$. Da er:

$$F(y) = P(Y < y) = (1 - e^{-y})^3.$$

Sannsynlighetstettheten:

$$f(y) = F'(y) = 3e^{-y}(1 - e^{-y})^2.$$

c) Monte Carlo metoder genererer stokastiske variable på datamaskinen og bruker statistikk over mange slike simuleringer til å forstå ulike egenskaper med lys i toalettet. Her simulerer man 10000 ganger fra de aktuelle fordelingene med parametre $\beta = 1$, $\beta = 2$ og $\beta = 3$. For hver av de 10000 simuleringene kan man notere hvilken lyspære som lever lengst (lyspære 1, 2 eller 3), hvor lenge dette er ($Y = \max\{X_1, X_2, X_3\}$), og om denne er større enn verdien 10.

Basert på Monte Carlo simuleringen anslås ulike resultatet: sannsynlighet for at lyspære 1 lever lengst: 0.1278, at lyspære 2 lever lengst: 0.3393 og at lyspære 3 lever lengst 0.5329. Fordelingen til maksimumsverdien har forventning 3.9291 og median 3.2104. Sannsynligheten for å se levetid lengre enn 10 år er 0.0405. Histogram til simuleringene av Y er vist i figuren.

Oppgave 3 Skigåing

a) En standardisering gir at $Z = \frac{X-10}{0.5}$ er standard normalfordelt. Den sine percentiler kan slås opp i tabell.

$$P(X < 9) = P(\frac{X - 10}{0.5} < \frac{9 - 10}{0.5}) = P(Z < -2) = 0.023$$

For kritisk grense c:

$$P(X > c) = 0.99 \rightarrow P(\frac{X - 10}{0.5} > \frac{c - 10}{0.5}) = 0.99$$

Det betyr at

$$\frac{c - 10}{0.5} = -2.33 \rightarrow c = -2.33 \cdot 0.5 + 10 = 8.84$$

b) Hypotesetest $H_0 : \mu = 10$, mot alternativ H_1: $\mu < 10$.

Her er gjennomsnittet $\bar{x} = 97.55/10 = 9.755$ et estimat på μ. Vi forkaster H_0 dersom $\bar{x}$ er signifikant mindre enn 10. Spesielt er $T = \frac{\bar{X}-10}{s/\sqrt{10}}$ t-fordelt med $10 - 1 = 9$ frihetsgrader, og vi forkaster på signifikansnivå 0.05 dersom $T < t_{9,0.05} = -1.83$.

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{9}(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2) = \frac{1}{9}(952.3 - 10 \cdot 9.755^2) = 0.28^2$$

Vi observerer $T = \frac{9.755-10}{0.28/\sqrt{10}} = -2.77$. Det vil si at hypotese H_0 forkastes på dette signifikansnivået. (p-verdi er ca 0.011.)

Fortegnstesten teller opp hvor mange data som er på hver side av $\mu = 10$ som gjelder under H_0. Her er hypotesen at medianen i fordelingen er 10, så antall data som er lavere enn 10 er binomisk fordelt med $n = 10$ forsøk og sannsynlighet 0.5. Dersom vi observerer signifikant mange lave tider, så forkastes H_0.

Fra data ser vi at 8 tider som er mindre enn 10. Under H_0 er antallet under medianen fordelt som $V \sim \text{Bin}(10, 0.5)$. Da er $P(V \geq 8) = 1 - P(V \leq 7) = 1 - 0.9453 = 0.0547$. Ettersom dette gir en p-verdi høyere enn 0.05, så forkastes ikke H_0, men det er nær kritisk grense.

Oppgave 4 Saltmålinger

a) Vanlige formler for regresjonskoeffisientene gir:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{77.2}{64.6} = 1.195$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 29.6 - 1.195 \cdot 28.0 = -3.86$$

Vi vet at $T = \frac{\hat{\beta}-\beta}{s/\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}} \sim t_{n-2}$. Her er $n = 20$. Percentil for $0.10/2 = 0.05$ er $t_{18,-0.05} = 1.734$.

$$P(-1.734 < \frac{\hat{\beta} - \beta}{s/\sqrt{\sum_{i=1}^{20}(x_i - \bar{x})^2}} < 1.734) = 0.9$$

$$P(\hat{\beta} - 1.734s/\sqrt{\sum_{i=1}^{20}(x_i - \bar{x})^2} < \beta < \hat{\beta} + 1.734s/\sqrt{\sum_{i=1}^{20}(x_i - \bar{x})^2}) = 0.9$$

Innsatte verdier gir $\hat{\beta} \pm 1.734 \cdot 0.48/\sqrt{64.6}$ og 90 % intervall $(1.09, 1.30)$.

b) Prediksjonen er $\hat{y}_0 = \hat{\alpha} + x_0\hat{\beta} = -3.86 + 1.195 \cdot 30 = 32.0$.

Vi kan skrive $\hat{y}_0 = \bar{y} + (x_0 - \bar{x})\hat{\beta}$, og $\bar{y}$ og $\hat{\beta}$ er uavhengige. Variansen til $\hat{y}_0$ er da $\sigma^2(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})$. Usikkerhet er minst for $x_0 = \bar{x}$.

Verdien $x_0 = 20$ er langt fra $\bar{x}$. Variansen blir større på grunn av avstanden til $\bar{x}$, men i tillegg ligger 20 utenfor de fysiske verdiene vi har i datamaterialet. Der er kanskje ikke modellen gyldig. Ekstrapolering utenfor område der man har data er ofte vanskelig å stole på.