

Løsningsforslag eksamen TMA4245, mai 2024

Oppgave 1 Hendelser og sannsynlighet

a)

Området viser hendelsen C , og siden C er en delmengde av B så er det riktige svaret $B \cap C$.

b)

For å løse oppgaven bruker vi varianter av regnereglene $P(A|B) = P(A \cap B)/P(B)$ og $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Det er også svært nyttig å skissere et venndiagram. Vi finner at

$$P(A'|B) = \frac{P(A' \cap B)}{P(B)} = \frac{P(B) - P(A \cap B)}{P(B)} = \frac{P(A \cup B) - P(A)}{P(B)} = \frac{0.88 - 0.25}{0.70} = 0.90.$$

Oppgave 2 Simultanfordeling

a)

$$P(Y = 2) = 0.10 + 0.02 + 0.09 + 0.07 = 0.28$$

b)

Vi finner først marginalfordelingen til X ; $P(X = -1) = 0.25$, $P(X = 0) = 0.10$, $P(X = 1) = 0.30$ og $P(X = 2) = 0.35$. Deretter regner vi ut forventningsverdien:

$$E(X) = \sum_x x \cdot P(X = x) = -1 \cdot 0.25 + 0 \cdot 0.1 + 1 \cdot 0.3 + 2 \cdot 0.35 = 0.75.$$

Oppgave 3 Poissonfordeling

a)

Siden $X + Y \sim \text{Poisson}(5)$ kan vi finne fra tabell for kumulative sannsynligheter i Poissonfordelingen at $P(X + Y \leq 5) = 0.6160$.

b)

Vi finner forventningsverdien til alle de seks uttrykkene:

$$E\left(\frac{X_1}{t_1} + \frac{X_2}{t_2} - \frac{X_3}{t_3}\right) = \frac{\lambda t_1}{t_1} + \frac{\lambda t_2}{t_2} - \frac{\lambda t_3}{t_3} = 2\lambda - \lambda = \lambda$$

$$E\left(\frac{X_1}{t_1}\right) = \frac{\lambda t_1}{t_1} = \lambda$$

$$E\left(\sum_{i=1}^3 \frac{X_i}{t_i}\right) = \frac{\lambda t_1}{t_1} + \frac{\lambda t_2}{t_2} + \frac{\lambda t_3}{t_3} = 3\lambda$$

$$E\left(0.5 \cdot \left(\frac{X_1}{t_1} + \frac{X_2}{t_2}\right)\right) = 0.5 \cdot 2\lambda = \lambda$$

$$E\left(\frac{X_1 + X_2}{t_1 + t_2}\right) = \frac{\lambda t_1 + \lambda t_2}{t_1 + t_2} = \lambda$$

$$E\left(\frac{\sum_{i=1}^3 X_i}{\sum_{i=1}^3 t_i}\right) = \frac{\lambda t_1 + \lambda t_2 + \lambda t_3}{t_1 + t_2 + t_3} = \lambda$$

Det er dermed $E\left(\sum_{i=1}^3 \frac{X_i}{t_i}\right)$ som *ikke* er en forventningsrett estimator for λ .

Oppgave 4 Feilforplantning

(i) Målefunksjonen er $g(V, D) = l \cdot f_d \cdot \frac{\rho}{2} \cdot \frac{V^2}{D}$ der $E(V) = v$, $E(D) = d$, $\text{Var}(V) = \sigma_V^2$, $\text{Var}(D) = \sigma_D^2$ og $\text{Cov}(V, D) = 0$. Videre er

$$\frac{\partial g(v, d)}{\partial v} = l \cdot f_d \cdot \frac{\rho}{2} \cdot \frac{2v}{d} \quad \text{og} \quad \frac{\partial g(v, d)}{\partial d} = -l \cdot f_d \cdot \frac{\rho}{2} \cdot \frac{v^2}{d^2}.$$

Dermed er

$$\text{Var}(g(V, D)) \approx \left(l \cdot f_d \cdot \frac{\rho}{2}\right)^2 \cdot \left(\left(\frac{2v}{d}\right)^2 \cdot \sigma_V^2 + \left(\frac{v^2}{d^2}\right)^2 \cdot \sigma_D^2\right).$$

(ii) De manglende linjene med kode er

```
simD = np.random.normal(d, sigma_D, size = n)
mux = np.sum(simX)/n
```

eller

```
mux = np.mean(simX)
```

Oppgave 5 Binomisk fordeling

a)

For uavhengige stokastiske variabler $X_1, \dots, X_n$ gjelder $M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t)$. I vårt tilfelle er $M_{X_i}(t) = (pe^t + 1 - p)$ for alle $i = 1, \dots, n$ og dermed er

$$M_Y(t) = (pe^t + 1 - p)^n.$$

Dette kjenner vi igjen som den momentgenererende funksjonen til en binomisk fordelt stokastisk variabel med parametere n og p , og dermed er $Y \sim \text{Binom}(n, p)$. Videre finner vi forventningsverdien til Y ved å derivere den momentgenererende funksjonen mhp t , før vi setter $t = 0$. Den deriverte er

$$\frac{d}{dt}M_Y(t) = \frac{d}{dt}(pe^t + 1 - p)^n = n(pe^t + 1 - p)^{n-1}pe^t.$$

Ved å sette $t = 0$ finner vi dermed at

$$E(Y) = n(p + 1 - p)^{n-1}p = np.$$

b)

Rimelighetsfunksjonen for p er

$$L(p; y, w) = f(y; n, p)f(w; n, 1 - p) = \binom{n}{y}p^y(1 - p)^{n-y}\binom{n}{w}(1 - p)^w p^{n-w}.$$

Log-rimelighetsfunksjonen er

$$l(p; y, w) = \ln \binom{n}{y} + \ln \binom{n}{w} + y \ln p + (n - y) \ln (1 - p) + w \ln (1 - p) + (n - w) \ln p.$$

Den deriverte av log-rimelighetsfunksjonen med hensyn på p er

$$\frac{dl(p; y, w)}{dp} = \frac{y}{p} - \frac{n - y}{1 - p} - \frac{w}{1 - p} + \frac{m - w}{p} = \frac{y + m - w}{p} - \frac{n - y + w}{1 - p}.$$

Vi setter den deriverte lik null og løser for p :

$$\begin{aligned} \frac{y + m - w}{p} - \frac{n - y + w}{1 - p} &= 0 \Rightarrow \\ (y + m - w)(1 - p) - (n - y + w)p &= 0 \Rightarrow \\ y + m - w &= (n + m)p \Rightarrow \\ p &= \frac{y + m - w}{n + m} \end{aligned}$$

Vi finner dermed at sannsynlighetsmaksimeringsestimatoren for p er

$$\hat{p} = \frac{Y + m - W}{n + m}.$$

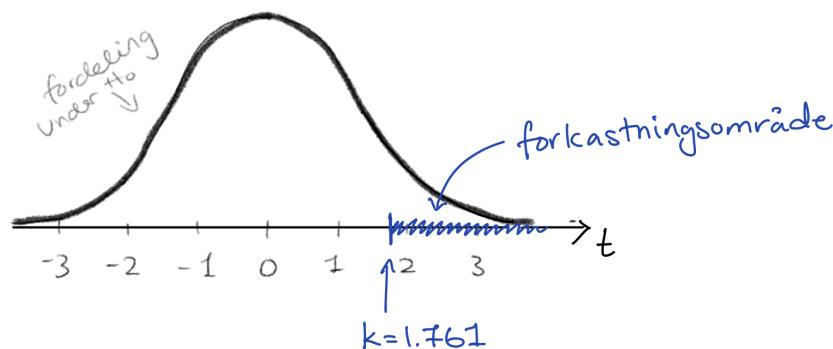


Figure 1: Skisse av fordeling til T under H_0 , inkludert kritisk verdi og forkastningsområde ved signifikansnivå $\alpha = 0.05$.

Oppgave 6 Velferd i lakseoppdrett

a)

(i) Under antagelsen om at H_0 er sann er testobservatoren T t -fordelt med parameter $\nu = 14$. Kritisk verdi ved signifikansnivå $\alpha = 0.05$ er $k = 1.761$. Sannsynlighetsfordelingen og forkastningsområdet er skissert i Figur 1.

(ii) Vi regner ut $\bar{d} = \frac{8.55}{15} = 0.57$ og $s^2 = \frac{17.64}{14} = 1.26$. Dermed blir observasjonen vår $t_{\text{obs}} = \frac{0.57}{\sqrt{\frac{1.26}{15}}} = 1.966687 \approx 1.97$. Siden $t_{\text{obs}} > k$ forkaster vi null-hypotesen. De 15 observasjonene som er gjort indikerer at behandling er forbundet med dårlige fiskevelferd.

(ii) Resultatet av testen er skissert i Figur 2, og p -verdien er gitt av arealet under kurven fra t_{obs} og ut i høyre hale, altså $P(T \geq t_{\text{obs}})$ under antagelsen om at H_0 er sann. Siden $P(T \geq k) = \alpha$, og $t_{\text{obs}} > k$ følger det at p -verdien er mindre enn α .

b)

(i) I en fortegnstest vil vi her telle vi opp antall observasjoner (observerte differanser) som er større enn null. Dette antallet bruker vi som vår testobservator. Testobservatoren U kan dermed formuleres som $U = \sum_{i=1}^{15} \mathbb{1}(D_i > 0)$. Under H_0 vil U være binomisk fordelt med parametre $n = 15$ og $p = 0.5$. Under H_1 vil $p > 0.5$. Med antagelsen om at H_0 er sann ser vi etter en høyest mulig kritisk verdi k_u som oppfyller $P(U \geq k_u) \leq 0.05$. Tilsvarende må vi finne lavest mulig k_u slik at $P(U \leq k_u - 1) \geq 0.95$. Fra tabeller ser vi at $P(X \leq 11) = 0.982$ for $X \sim \text{Binom}(15, 0.5)$ og vi velger derfor $k_u = 12$ som kritisk verdi. Vi forkaster H_0 dersom vi observerer $U \geq 12$. (Også korrekte forkastningsregler; $U > 11$, og p -verdi mindre eller lik 0.05 dersom p -verdien er definert som $P(U \geq u)$ for $U \sim \text{binom}(15, 0.05)$ og u er observasjonen).

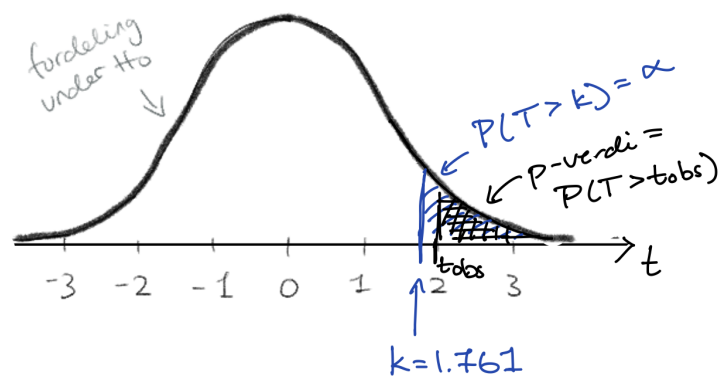


Figure 2: Skisse av fordeling til T under H_0 , inkludert kritisk verdi og forkastningsområde ved signifikansnivå $\alpha = 0.05$, observasjonen t_{obs} og tilhørende p -verdi.

(ii) Det faktiske signifikansnivået blir $1 - 0.0982 = 0.018$, som er en del lavere enn 0.05. At signifikansnivået ikke er nøyaktig lik 0.05 skyldes at testobservatoren har en diskret fordeling.

(iii) Resultatet av testen er at vi ikke kan forkaste H_0 siden observasjonen vår $u = 10$ er mindre enn kritisk verdi $k_u = 12$. Også korrekt svar vil være å regne ut p -verdien $P(U \geq 10) = 1 - P(U \leq 9) = 1 - 0.849 = 0.151$. Siden p -verdien er høyere enn ønsket signifikansnivå α kan vi ikke forkaste H_0 .

Oppgave 7 93CARS

a)

(i) I boksplottet ser vi en boks som strekker seg fra 0.25-kvantilen til 0.75-kvantilen. Dersom vi sorterer alle observerte priser på biler med 0 kollisjonsputer etter hverandre så vil vil 25% av observasjonene ha en pris som er lavere enn nederste linje i boksen (ca 9 000 USD), mens de neste 50% av observasjonene i den sorterte rekka vil ha en pris som ligger innenfor boksen (mellom ca 9 000 og 17 000 USD). De 25% dyreste bilene med 0 kollisjonsputer har en pris over (omlag) 17 000 USD. Den tykke svarte linja som ligger inne i boksen representerer medianen. Her er medianen ca 12 000 USD. Dermed har halvparten av bilene uten kollisjonspute en pris under 12 000 USD, mens den andre halvparten har en pris over 12 000 USD.

(ii) I figuren ser vi en økende trend - jo fler kollisjonsputer jo dyrere bil. Det er imidlertid stor variasjon i observasjonene i pris for biler med 2 kollisjonsputer, og det er usikkert om det er en reell prisforskjell mellom biler med 1 og 2 kollisjonsputer. Vi har derimot grunn til å tro at biler uten kollisjonsputer gjerne er rimeligere enn biler med kollisjonsputer.

b)

(i) For å regne ut estimatet for stigningstallet og skjæringspunkt trenger vi gjennomsnittlig antall kollisjonsputer

$$\bar{x} = \frac{0 \cdot 34 + 1 \cdot 43 + 2 \cdot 16}{93} = \frac{75}{93} \approx 0.806,$$

og gjennomsnittspris

$$\bar{y} = \frac{13.17 \cdot 34 + 21.22 \cdot 43 + 28.37 \cdot 16}{93} = \frac{1814.16}{93} = 19.5071.$$

Vi regner ut

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{34 \cdot (0 - \bar{x}) \cdot 13.17 + 43 \cdot (1 - \bar{x}) \cdot 21.22 + 16 \cdot (2 - \bar{x}) \cdot 28.37}{34 \cdot (0 - \bar{x})^2 + 43 \cdot (1 - \bar{x})^2 + 16 \cdot (2 - \bar{x})^2} \\ &= 7.6805 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 13.31315\end{aligned}$$

Regresjonslinja er dermed estimert til å være $y = 13.31 + 7.68 \cdot x$, der y er pris i 1000\$.

(ii) Predikert pris for en bil med en kollisjonspute blir i følge modellen fra (i) $13.31 + 7.68 = 20.99$ USD. Dette stemmer godt med figuren, der vi kan se at linja krysser $x = 1$ omtrent ved $y = 21$ (1000 \$). Figuren viser et 95% konfidensintervall som reflekterer usikkerhet i estimering av forventet pris som en funksjon av antall kollisjonsputer. Et 95% prediksjonsintervall vil være videre enn konfidensintervallet fordi prediksjonsintervallet reflekterer både usikkerhet i estimering og usikkerhet forbundet med den nye observasjonen.

(iii) Vi observerer noe mindre spredning rundt regresjonslinja ved $x = 0$ enn ved $x = 1$ og $x = 2$. Dette kan tyde på at antagelsen om konstant variasjon $\varepsilon \sim N(0, \sigma^2)$ er brutt.