Løsningsforslag til kontinuasjonseksamen ISTx100y 2023

Oppgave 1: Observasjoner fra normalfordeling

Vi har observert $x_1 = 6.6$, $x_2 = 3.8$ og $x_3 = 5.9$ fra en normalfordeling.

a) Gjennomsnittet av observasjonene blir:

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{6.6 + 3.8 + 5.9}{3} = \frac{16.3}{3} = 5.433333$$

og riktig svaralternativ (avrundet til 2 desimaler) blir 5.43.

b) Den empiriske variansen til observasjonene er gitt ved:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{3-1}\sum_{i=1}^{3}(x_i - \bar{x})^2 = \frac{(6.6 - 5.43)^2 + (3.8 - 5.43)^2 + (5.9 - 5.43)^2}{2}$$
$$= \frac{4.2467}{2} = 2.12335$$

som gir oss empirisk standardavvik lik $\sqrt{2.12335} = 1.457166$. Riktig svaralternativ blir derfor 1.46.

c) Et 95% konfidensintervall for μ med disse observasjonene finner vi ved å bruke formel for konfidensintervall for forventningsverdi med ukjent standardavvik:

$$\left[\bar{X} - t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1} \cdot \frac{S}{\sqrt{n}}\right]$$

Her er $t_{\alpha/2,n-1} = t_{0.05/2,2} = 4.303$, og observasjonene $\bar{x}$ og x har vi regnet ut over. Intervallet blir dermed:

$$\left[5.433333 - 4.303 \cdot \frac{1.457166}{\sqrt{3}}, 5.433333 + 4.303 \cdot \frac{1.457166}{\sqrt{3}}\right] = [1.81, 9.05]$$

der vi har rundet av til to desimaler etter utregning av intervallet.

Oppgave 2: Simultanfordeling

a) Marginalfordelingen til Y regnes ut fra simultanfordeling ved å summere sannsynlighetene for Y over alle mulige utfall for X:

$$P(Y = y) = \sum_{x=1}^{3} P(Y = y, X = x)$$

Da får vi at:

$$\begin{aligned} P(Y = 1) &= P(Y = 1, X = 1) + P(Y = 1, X = 2) + P(Y = 1, X = 3) = 0.2 + 0.1 + 0.3 \\ &= 0.6 \\ P(Y = 2) &= P(Y = 2, X = 1) + P(Y = 2, X = 2) + P(Y = 2, X = 3) = 0.1 + 0.2 + 0.1 \\ &= 0.4 \end{aligned}$$

b) Forventningsverdien til X finner vi ved å bruke marginalfordelingen (som står oppgitt i oppgaveteksten): $P(X = 1) = 0.3, P(X = 2) = 0.3, P(X = 3) = 0.4$. Forventningsverdi kan regnes ut ved hjelp av følgende formel:

$$\begin{aligned} E(X) &= \sum_{x} x \cdot P(X = x) = \sum_{x=1}^{3} x \cdot P(X = x) = 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) \\ &= 1 \cdot 0.3 + 2 \cdot 0.3 + 3 \cdot 0.4 = 0.3 + 0.6 + 1.2 = 2.1 \end{aligned}$$

c) Merk at $X < 3$ betyr $X \leq 2$. Vi skal altså se på tilfellene der X kan ta verdiene 1 og 2. Vi bruker formel for betinget sannsynlighet:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

og med våre tall får vi da (merk at vi bruker både $\cap$ og komma (,) for å beskrive "og"):

$$\begin{aligned} P(Y = 2|X < 3) &= \frac{P(Y = 2 \cap X < 3)}{P(X < 3)} = \frac{P(Y = 2, X = 1) + P(Y = 2, X = 2)}{P(X = 1) + P(X = 2)} \\ &= \frac{0.1 + 0.2}{0.3 + 0.3} = \frac{0.3}{0.6} = 0.5 \end{aligned}$$

Oppgave 3: Utregning av sannsynligheter med Python

a) Koden

```
from scipy import stats
print(stats.binom.pmf(4,8,0.4))
```

gir oss punktsannsynligheten for $X = 4$ der $X \sim \text{Binom}(n = 8, p = 0.4)$ (se også formelark). Denne sannsynligheten kan vi regne ut fra formelen for punktsannsynligheter i binomisk fordeling:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \ x = 0, 1, 2, \ldots, n$$

som gir oss:

$$P(X = 4) = \binom{8}{4} 0.4^4 (1 - 0.4)^{8-4} = 70 \cdot 0.4^4 \cdot 0.6^4 = 0.23$$

b) Kommandoen som gir oss $P(Z \leq 2)$ er kommandoen for kumulativ fordeling for en standard normalfordelt variabel (forventning 0 og standardavvik 1), se siste seksjon på formelarket. Dermed blir riktig svar:

```
stats.norm.cdf(2, 0, 1)
```

Oppgave 4: Drikkevann

a) I denne del-oppgaven får vi oppgitt at vannet er *ikke* forurenset i 40 dager. Det gjøres daglig uavhengige kontroller, der hver har en sannsynlighet på $p = 0.001$ for å føre til kokevarsel selv om vannet ikke er forurenset. La V være antall kokevarsler som sendes ut over de 40 dagene. V er da binomisk fordelt med $n = 40$ og $p = 0.001$, $V \sim \text{Binom}(40, 0.001)$. At minst ett kokevarsel sendes ut er hendelsen $V \geq 1$, og vi vil ha sannsynligheten for dette:

$$\begin{aligned} P(V \geq 1) &= 1 - P(V < 1) = 1 - P(V = 0) = 1 - \binom{40}{0} \cdot 0.001^0 \cdot (1 - 0.001)^{40-0} \\ &= 1 - 1 \cdot 1 \cdot 0.999^{40} = 1 - 0.9608 = 0.039 \end{aligned}$$

b) Nå antar vi at vannet faktisk er forurenset. Siden sjansen for å oppdage E. coli i forurenset vann er $p = 0.98$ hver dag er sjansen for å ikke oppdage E. coli (og dermed sende varsel):

$$1 - 0.98 = 0.020$$

c) I kontrollprøven kjenner vi den sanne konsentrasjonen, og derfor har vi $X \sim \text{N}(1, 0.01)$. Sannsynligheten for at én prøve gir et måleresultat på mellom 0.98 og 1.02 er gitt ved:

$$\begin{aligned} P(0.98 \leq X \leq 1.02) &= P(X \leq 1.02) - P(X \leq 0.98) = P\left(Z \leq \frac{1.02 - 1}{0.01}\right) - P\left(Z \leq \frac{0.98 - 1}{0.01}\right) \\ &= P\left(Z \leq \frac{0.02}{0.01}\right) - P\left(Z \leq \frac{-0.02}{0.01}\right) = P(Z \leq 2) - P(Z \leq -2) \\ &= 0.9772 - 0.0228 = 0.9544 \end{aligned}$$

Her har vi standardisert X så vi får en standard normalfordelt variabel og dermed kan bruke tabeller for å lete opp de kumulative sannsynlighetene.

d) Målingene som er gjort er: $x_1 = 0.964, x_2 = 1.001, x_3 = 0.948, x_4 = 0.954, x_5 = 1.021$. Vi skal teste: $H_0 : \mu = 1$ cfu/100 ml mot $H_1 : \mu < 1$ cfu/100ml. Dette er en venstresidig test (type 3 på formelarket), og vi har kjent standardavvik ($\sigma = 0.01$). For å utføre testen regner vi ut en verdi for testobservatoren:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Gjennomsnittet av de $n = 5$ målingene er 0.9776, $\mu_0 = 1$, og $\sigma = 0.01$ er oppgitt og kjent. Dermed får vi følgende verdi for testobservatoren:

$$z = \frac{0.9776 - 1}{0.01/\sqrt{5}} = -5.009$$

Den kritiske verdien i en venstresidig test er $-z_\alpha$. Ved signifikansnivå $\alpha = 0.05$ er altså den kritiske verdien $-z_\alpha = -1.645$. Siden $z < -z_\alpha$, forkaster vi H_0 til fordel for H_1 ved signifikansnivå 0.05. Det er dermed grunnlag for å tro at den nye metoden systematisk gir for lave målinger.

Oppgave 5: Elektrisk tungtransport

El-lastebilene ankommer ifølge en Poissonprosess med rate $\lambda = 3/\text{time}$.

a) La X være antall el-lastebiler som ankommer. Det betyr at $X \sim \text{Poisson}(\lambda t) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$. Sannsynligheten for at det i løpet av én time ($t = 1$) ankommer nøyaktig 2 lastebiler er da:

$$P(X = 2) = \frac{(3 \cdot 1)^2}{2!} e^{-3 \cdot 1} = \frac{3^2}{2} e^{-3} = 0.2240418 \approx 0.22$$

b) La X være som i forrige deloppgave. Sannsynligheten for at det i løpet av én time ankommer mer enn 3 el-lastebiler er:

$$P(X > 3) = 1 - P(X \leq 3) = 1 - 0.6472 = 0.3528 \approx 0.35$$

der vi har brukt tabell for Poisson kumulativfordeling for å finne $P(X \leq 3)$ (med $\lambda t = 3$).

c) I en Poissonprosess er tid mellom hendelser eksponensialfordelt. La T betegne tiden mellom ankomst av el-lastebiler. Da er $T \sim \text{Eksponensial}(\lambda)$, og $P(T \leq t) = 1 - e^{-\lambda t}$. Sannsynligheten for at den neste el-lastebilen ankommer i løpet av 30 minutter (= 0.5 time) etter den forrige blir da:

$$P(T \leq t) = 1 - e^{-3 \cdot 0.5} = 1 - e^{-1.5} = 0.7768698 \approx 0.78$$

d) Kjørelengden L på et oppdrag er Weibullfordelt slik at $L \sim \text{Weibull}(\alpha = 1.5, \lambda = 1/100)$. Merk at dette er lengden fra firmaets sentral til leveringsstedet (og ikke medregnet retur!). Fra formelark har vi da at $P(L \leq l) = 1 - e^{-(\lambda l)^\alpha}$. Sannsynligheten for at et oppdrag har kjørelengde høyere enn 200 km er da:

$$P(L > 200) = 1 - P(L \leq 200) = 1 - (1 - e^{-(\frac{1}{100} \cdot 200)^{1.5}}) = e^{-(\frac{1}{100} \cdot 200)^{1.5}} = e^{-(2^{1.5})} = 0.05910575 \approx 0.06$$

Figuren hjelper oss å dobbelsjekke svaret vårt siden arealet fra 200 km og ut i høyre hale er svært lite (totalt areal = 1).

e) En fulladet el-lastebil har en rekkevidde på 250 km, som betyr at for å kjøre tur-retur må oppdragets lengde være maks $250/2 = 125$ km for at bilen ikke skal trenge lading før den er tilbake ved sentralen. Dermed skal vi regne ut:

$$P(L \leq 125) = 1 - e^{-((\frac{125}{100})^{1.5})} = 0.7527963 \approx 0.75$$

Oppgave 6: Enkel lineær regresjon

a) Kryssplottet har ikke en lineær trend, men heller noe som ser kvadratisk ut. Dermed er det "Det er en lineær sammenheng mellom x og y" som er den brutte antagelsen.

b) Ved å heller bruke $v = x^2$ får vi en lineær trend mellom v og y, og kan bruke enkel lineær regresjon. Stigningstallet β_1 estimeres med følgende formel (fra formelark):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(v_i - \bar{v})(y_i - \bar{y})}{\sum_{i=1}^{n}(v_i - \bar{v})^2} = \frac{40472.6}{20824.3} = 1.94$$

der vi har brukt de oppgitte størrelsene i oppgaveteksten for å fylle inn tall.

c) $x = 5$ betyr at $v = x^2 = 5^2 = 25$. Vi trenger estimatet til skjæringspunktet β_0 for å predikere y, som vi finner med formelen (formel fra formelark, tall fra oppgavetekst):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{v} = 72.3 - 1.94 \cdot 34.5 = 5.37$$

Dermed blir den estimerte regresjonslinja:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 v_i = 5.37 + 1.94 v_i = 5.37 + 1.94 x_i^2$$

og predikert y for $x = 5$ blir da

$$5.37 + 1.94 \cdot 25 = 53.9$$