UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1100 — Sannsynlighetsregning og statistisk modellering
Eksamensdag:	7. juni - 2024
Tid for eksamen:	09.00 – 13.00.
Oppgavesettet er på 7 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator Formelsamling for STK1100

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

Vi har at X er Poisson-fordelt, dvs.

$$P(X = x) = \frac{(\lambda v_0)^x}{x!} e^{-\lambda v_0}, \quad x = 0, 1, 2, ...$$

(a) $\mathsf{E}(X) = \lambda v_0$ og $\mathrm{V}(X) = \lambda v_0$. Vi kan tolke λ som forventet antall E.coli-bakterier per liter vann.

(b) Med $\lambda = 3$ og $v_0 = 1$ finner vi

$$P(X = 0) = e^{-3} = 0.0498.$$

Med $\lambda = 3$ skal vi finne v_0 slik at det er minst 0.9975 i sannsynlighet for at $P(X \geq 1)$:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-3v_0} \geq 0.9975$$

som gir $v_0 \geq 1.997$, dvs. prøven må være mer enn 1.997 liter.

(c) La X_i være antall bakterier i prøve i, $i = 1, 2, ..., 10$. Vi har

$$P(X_i = 0) = e^{-0.3} = 0.741.$$

Bruker uavhengigheten til å multiplisere sannsynlighetene i siste linje:

$$\begin{aligned} P(\text{alarm}) &= P(\text{minst én } X_i \geq 1) \\ &= 1 - P(\text{ingen } X_i \geq 1) \\ &= 1 - P(\text{alle } X_i = 0) \\ &= 1 - 0.741^{10} = 0.9502. \end{aligned}$$

(Fortsettes på side 2.)

Dette er egentlig identisk med situasjonen i a) der man ser etter minst én bakterie i en liter vann.

Også mulig å identifisere dette som en binomisk situasjon, $Y \sim \text{bin}(n, p)$ med $n = 10$, $p = 1 - P(X_i = 0) = 1 - e^{-0.3} = 0.259$, og finne $P(\text{alarm}) = P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1 - 0.259)^{10} = 0.9502$

(d) Har $\mathsf{E}(X_1) = \mathrm{V}(X_1) = \lambda v_1$ og $\mathsf{E}(X_2) = \mathrm{V}(X_2) = \lambda v_2$. Reglene for forventning og varians til lineærkomb. av uavhengige stokastiske variabler gir

$$\mathsf{E}(\hat{\lambda}) = \frac{\mathsf{E}(X_1 + X_2)}{v_1 + v_2} = \frac{\lambda(v_1 + v_2)}{v_1 + v_2} = \lambda$$

$$\mathrm{V}(\hat{\lambda}) = \frac{\mathrm{V}(X_1 + X_2)}{(v_1 + v_2)^2} = \frac{\lambda(v_1 + v_2)}{(v_1 + v_2)^2} = \frac{\lambda}{v_1 + v_2}$$

Standardfeilen til $\hat{\lambda}$ er derfor $\sqrt{\frac{\lambda}{v_1+v_2}}$.

Oppgave 2

La X og Y være kontinuerlige stokastiske variabler med simultantetthet

$$f(x, y) = \begin{cases} \frac{1}{17}x(x + y) & \text{når } 0 \leq x \leq 2 \text{ og } 0 \leq y \leq 3 \\ 0 & \text{ellers} \end{cases}$$

(a) Det enkleste er å beregne

$$P(Y \geq X) = \frac{1}{17}\int_0^2 \int_x^3 (x^2 + xy)dydx = \frac{11}{17}.$$

Kan også finnes ved

$$P(Y \geq X) = \frac{1}{17}\int_0^2 \int_0^y (x^2 + xy)dxdy + \frac{1}{17}\int_2^3 \int_0^2 (x^2 + xy)dxdy = \frac{11}{17}$$

(tegn figur over området for å se integrasjonsgrensene!).

(b) Når $0 \leq y \leq 3$ har vi at

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx = \frac{1}{17}\int_0^2 (x^2 + xy)dx = \frac{1}{17}\left[\frac{1}{3}x^3 + \frac{1}{2}x^2 y\right]_{x=0}^2 = \frac{1}{17}(\frac{8}{3} + 2y)$$

slik at vi får

$$f_Y(y) = \begin{cases} \frac{2}{17}(\frac{4}{3} + y) & \text{når } 0 \leq y \leq 3 \\ 0 & \text{ellers} \end{cases}$$

(Fortsettes på side 3.)

Når $0 \leq x \leq 2$ har vi at

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{17}\int_0^3 (x^2+xy)dy = \frac{1}{17}\left[x^2y + \frac{1}{2}xy^2\right]_{y=0}^3 = \frac{1}{17}(3x^2+\frac{9}{2}x),$$

slik at vi får

$$f_X(x) = \begin{cases} \frac{3}{17}x(x+\frac{3}{2}) & \text{når } 0 \leq x \leq 2 \\ 0 & \text{ellers} \end{cases}$$

X og Y kan ikke være uavhengige, siden vi enkelt ser fra formlene at

$$f(x,y) \neq f_X(x) \cdot f_Y(y).$$

(c) Når $0 \leq y \leq 3$ har vi at

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \frac{2}{17}\int_0^y (\frac{4}{3}+t)dt = \frac{2}{17}\left[\frac{4}{3}t + \frac{1}{2}t^2\right]_0^y$$

slik at vi får

$$F_Y(y) = \begin{cases} 0 & \text{når } y < 0 \\ \frac{2}{17}(\frac{4}{3}y + \frac{1}{2}y^2) & \text{når } 0 \leq y \leq 3 \\ 1 & \text{når } y > 3 \end{cases}$$

Vi finner medianen ν ved å løse andregradsligningen som oppstår når vi setter $F_Y(\nu) = \frac{1}{2}$.

Oppgave 3

(a) Vi har at

$$L(\sigma) = \prod_{i=1}^n f(x_i|\sigma) = \prod_{i=1}^n \frac{1}{2\sigma}\exp(-|x_i|/\sigma) = \frac{1}{2^n\sigma^n}\exp(-\sum_{i=1}^n |x_i|/\sigma)$$

og

$$\ell(\sigma) = -n\log(2) - n\log(\sigma) - \frac{1}{\sigma}\sum_{i=1}^n |x_i|.$$

Dermed er

$$\frac{\partial}{\partial\sigma}\ell(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^2}\sum_{i=1}^n |x_i|$$

og setter vi lik 0, får vi

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^n |x_i|.$$

(Fortsettes på side 4.)

(b) Vi har at

$$\begin{aligned} E[\hat{\sigma}] =& E[|X_i|] = \int_{-\infty}^{\infty} \frac{1}{2\sigma}|x| \exp(-|x|/\sigma)dx \\ =& \int_0^{\infty} \frac{1}{\sigma} x \exp(-x/\sigma)dx = \frac{1}{\sigma}\sigma^2\Gamma(2) = \sigma \end{aligned}$$

dvs forventningsrett.

Vi har videre at

$$\begin{aligned} E[|X|^2] =& \int_{-\infty}^{\infty} \frac{1}{2\sigma} x^2 \exp(-|x|/\sigma)dx = \int_0^{\infty} \frac{1}{\sigma} x^2 \exp(-x/\sigma)dx \\ =& \frac{1}{\sigma}\sigma^3\Gamma(3) = 2\sigma^2 \end{aligned}$$

Dermed er

$$V[|X|] = E[|X|^2] - E[|X|]^2 = 2\sigma^2 - \sigma^2 = \sigma^2$$

(c) Vi har at $\hat{\sigma}$ er et gjennomsnitt, og ifølge sentralgrenseteoremet vil da $\hat{\sigma}$ være tilnærmet normalfordelt når n er stor.

Vi har da at

$$\Pr(-z_{\alpha/2} < \frac{\hat{\sigma} - \sigma}{\sigma/\sqrt{n}} < z_{\alpha/2}) \approx 1 - \alpha$$

Nå er

$$\begin{aligned} \frac{\hat{\sigma} - \sigma}{\sigma/\sqrt{n}} <& z_{\alpha/2} \\ \Updownarrow& \\ \hat{\sigma} - \sigma <& z_{\alpha/2}\sigma/\sqrt{n} \\ \Updownarrow& \\ \hat{\sigma} <& (1 + z_{\alpha/2}/\sqrt{n})\sigma \\ \Updownarrow& \\ \frac{\hat{\sigma}}{1 + z_{\alpha/2}/\sqrt{n}} <& \sigma \end{aligned}$$

Tilsvarende vil den andre ulikheten gi

$$\frac{\hat{\sigma}}{1 - z_{\alpha/2}/\sqrt{n}} > \sigma$$

som da gir det ønskede intervallet.

(Fortsettes på side 5.)

(d) Vi kan også bruke at siden $\hat{\sigma}$ er konsistent, så kan argumentere med at

$$\frac{\hat{\sigma}-\sigma}{\sigma/\sqrt{n}} \approx \frac{\hat{\sigma}-\sigma}{\hat{\sigma}/\sqrt{n}}$$

som da også vil være tilnærmet normalfordelt. Da har vi

$$\begin{aligned}
\Pr(-z_{\alpha/2} < \frac{\hat{\sigma}-\sigma}{\hat{\sigma}/\sqrt{n}} < z_{\alpha/2}) \approx & 1-\alpha \\
& \Updownarrow \\
\Pr(-z_{\alpha/2}\hat{\sigma}/\sqrt{n} < \hat{\sigma}-\sigma < z_{\alpha/2}\hat{\sigma}/\sqrt{n}) \approx & 1-\alpha \\
& \Updownarrow \\
\Pr(-\hat{\sigma}-z_{\alpha/2}\hat{\sigma}/\sqrt{n} < -\sigma < -\hat{\sigma}+z_{\alpha/2}\hat{\sigma}/\sqrt{n}) \approx & 1-\alpha \\
& \Updownarrow \\
\Pr(\hat{\sigma}+z_{\alpha/2}\hat{\sigma}/\sqrt{n} > \sigma > \hat{\sigma}-z_{\alpha/2}\hat{\sigma}/\sqrt{n}) \approx & 1-\alpha
\end{aligned}$$

som da gir det alternative intervallet.

Siden vi har $\alpha = 0.05$ så ønsker vi at konfidensintervallet skal dekke den sanne parameter 95% av gangene vi bruker dette intervallet. For $n = 10$ så vil det første intervallet dekke det 95.9% av gangene mens det andre intervallet kun dekker det 91.1% av gangene, som antyder at det første intervallet er best.

Når $n = 100$ så oppfører begge intervallene seg svært nærme målnivået.

Oppgave 4

(a) Under Modell 2 så ønsker vi å minimere

$$g(b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - bx_i)^2$$

Vi har

$$\frac{\partial}{\partial b}g(b) = -2\sum_{i=1}^{n}(y_i - bx_i)x_i = -2[\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2]$$

og setter vi det lik 0, får vi

$$b = \hat{\gamma}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(b) Under Modell 2 har vi at

$$E\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right] = \frac{\sum_{i=1}^{n} x_i \gamma_1 x_i}{\sum_{i=1}^{n} x_i^2} = \gamma_1$$

(Fortsettes på side 6.)

Videre er

$$V[\hat{\gamma}_1] = V\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{\sum_{i=1}^n x_i^2\sigma^2}{[\sum_{i=1}^n x_i^2]^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

Ved å bruke et estimat for σ, kan vi da estimere standardfeilen.

(c) Vi kan da bruke den generelle formelen $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$ og får da

Parameter	L	U
β_0	-1.266	25.393
β_1	12.891	16.898
γ_1	15.437	17.483

Da intervallet for β_0 inneholder 0, er dette en mulig verdi, så Modell 2 er ikke helt urimelig her.

(d) Under Modell 1 har vi at

$$E\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{\sum_{i=1}^n x_i(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n x_i^2} = \beta_0\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_1$$

Videre er

$$V[\hat{\gamma}_1] = V\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{\sum_{i=1}^n x_i^2\sigma^2}{[\sum_{i=1}^n x_i^2]^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

Vi har da at $V[\hat{\gamma}_1] < V[\hat{\beta}_1]$ slik at vi kan få redusert varians ved å godta noe forventningsskjevhet.

(Fortsettes på side 7.)