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 5 sider.	
Vedlegg:	Formelsamling for STK1100
Tillatte hjelpemidler:	Godkjent kalkulator

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

Tabell over øvre persentiler (kvantiler) for standard normalfordeling:

α	0.1	0.05	0.025	0.01	0.005	0.0025
z_α	1.282	1.645	1.960	2.326	2.576	2.807

Følgende resultat kan brukes i deler av oppgavene:

$$\int_0^\infty x^{a-1}e^{-x/b}dx = b^a\Gamma(a).$$

der $\Gamma(a) = (a-1)!$ for a heltall.

Oppgave 1

Med jevne mellomrom kommer det melding om funn av E.coli-bakterier i prøver av drikkevannet i Oslo, som medfører at befolkningen må koke vannet før bruk. La X betegne antall E.coli-bakterier i v_0 liter vann fra en bestemt drikkevannskilde. Det kan da antas 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) Hva blir $\mathsf{E}(X)$ og $\mathrm{V}(X)$? Gi en enkel tolkning av parameteren λ.

(b) Anta at $\lambda = 3$. Finn sannsynligheten for at en tilfeldig valgt liter av drikkevannet er fri for E.coli-bakterier. Hvor stor prøve (v_0 liter) må man ta for at det skal være minst 0.9975 i sannsynlighet for at prøven skal inneholde minst én E.coli-bakterie?

(Fortsettes på side 2.)

(c) Noen foreslår at man i stedet for én stor prøve, tar 10 mindre prøver, hver på 0.1 liter, og sender ut alarm om E.coli dersom minst én av disse er positiv (positiv betyr funn av minst én bakterie i prøven). Anta som i b) at $\lambda = 3$, og at prøvene er uavhengige av hverandre. Hva er sannsynligheten for at alarmen går?

(d) Egentlig er λ ukjent og må estimeres. Det tas nå to tilfeldige vannprøver, en på v_1 liter og en på v_2 liter. Kall antall E.coli-bakterier i prøvene hhv. X_1 og X_2. Vi kan anta at X_1 og X_2 er uavhengige stokastiske variabler. En mulig estimator er

$$\hat{\lambda} = \frac{X_1 + X_2}{v_1 + v_2}$$

Finn denne estimatorens forventning og standardfeil.

Oppgave 2

La X og Y være kontinuerlige stokastiske variabler med simultantetthet

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

(a) Bruk simultanfordelingen ovenfor til å finne $P(Y \ge X)$.

(b) Vis at marginalfordelingen til Y er

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

og finn marginalfordelingen til X. Er X og Y uavhengige?

(c) Finn den kumulative fordelingen til Y. Hvordan kan du finne medianen til Y fra den kumulative fordelingen? I siste spørsmål trenger du kun skissere fremgangsmåten, ikke løse ligningen som fremkommer.

Oppgave 3

Anta at $X_1, X_2, ..., X_n$ er uavhengige og identisk fordelte stokastiske variable med sannsynlighetstetthet

$$f(x;\sigma) = \frac{1}{2\sigma} e^{-|x|/\sigma}, \quad -\infty < x < \infty.$$

(a) Vis at maksimum likelihoodestimatoren for σ, $\hat{\sigma}$, er gitt ved

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

(Fortsettes på side 3.)

(b) Finn $\mathsf{E}[\hat{\sigma}]$ og vis at $\mathrm{V}[\hat{\sigma}] = \sigma^2/n$.

Hint: Bruk at $|x|$ er symmetrisk om 0.

(c) Argumenter hvorfor

$$Z = \frac{\hat{\sigma} - \sigma}{\sigma/\sqrt{n}} \approx N(0,1).$$

Bruk dette til å vise at intervallet

$$\left[\frac{\hat{\sigma}}{1 + z_{\alpha/2}/\sqrt{n}}, \frac{\hat{\sigma}}{1 - z_{\alpha/2}/\sqrt{n}}\right] \tag{ki1}$$

er et konfidensintervall med tilnærmet konfidensnivå $100(1-\alpha)\%$ for σ.

(d) Argumenter hvorfor også

$$Z' = \frac{\hat{\sigma} - \sigma}{\hat{\sigma}/\sqrt{n}} \approx N(0,1).$$

og bruk dette til å konstruere et alternativt (tilnærmet) konfidensintervall for σ:

$$\left[\hat{\sigma} - z_{\alpha/2}\hat{\sigma}/\sqrt{n}, \hat{\sigma} + z_{\alpha/2}\hat{\sigma}/\sqrt{n}\right]. \tag{ki2}$$

I en simuleringsstudie ble $n = 10$ observasjoner simulert fra $f(x;\sigma)$ og de to konfidensintervallene (ki1) og (ki2) ble beregnet for $\alpha = 0.05$. Denne prosedyren ble repetert 1000 ganger. For intervall (ki1) dekket intervallet den sanne σ 959 ganger mens for intervallet (ki2) dekket intervallet den sanne σ 911 ganger.

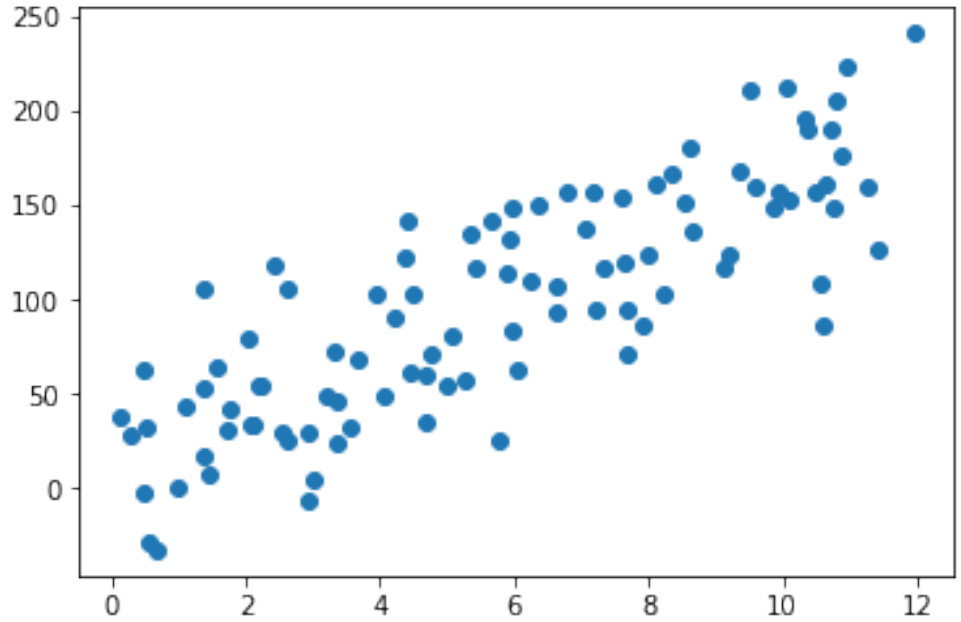
En tilsvarende simuleringstudie med $n = 100$ førte til at intervallene dekket σ 943 og 941 ganger for (ki1) og (ki2), henholdsvis.

Diskutér disse resultatene.

(Fortsettes på side 4.)

Oppgave 4

Figuren nedenfor viser data $\{(x_i, y_i), i = 1, ..., n\}$ for $n = 100$.



Vi vil anta at x_i-ene er kjente tall, mens Y_i-ene følger sammenhengen

$$Y_i = f(x_i) + \varepsilon_i$$

der ε_i-ene er uavhengige og identisk fordelte med $\mathsf{E}[\varepsilon_i] = 0$ og $\mathrm{V}[\varepsilon_i] = \sigma^2$. For $f(x)$, så vil vi se på to mulige modeller:

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x && \text{Modell 1;} \\ f(x) &= \gamma_1 x && \text{Modell 2.} \end{aligned}$$

For Modell 1 så vet vi at minste kvadraters estimater for β_0 og β_1 er

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

der $\bar{x}$ og $\bar{y}$ er gjennomsnittet av observasjonene. Dette trenger du ikke å vise!

(a) For Modell 2, vis at minste kvadraters estimat for γ_1 er

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

(b) Beregn forventningen til $\hat{\gamma}_1$ under forutsetning av at Modell 2 er riktig.

Beregn også varians til $\hat{\gamma}_1$, igjen under forutsetning av at Modell 2 er riktig.

Hvordan kan dette benyttes for å estimere standardfeilen til $\hat{\gamma}_1$?

(Fortsettes på side 5.)

Tabellen nedenfor viser estimater for β_0, β_1 og γ_1 basert på data i figuren ovenfor. Også estimert standardfeil er oppgitt, basert på tilsvarende argumenter som for forrige deloppgave.

Parameter	Estimat	Standardfeil
β_0	12.064	6.801
β_1	14.895	1.022
γ_1	16.460	0.522

(c) Basert på en antagelse om at alle 3 estimatorer er tilnærmet normalfordelte, lag 95% konfidensintervaller for de tre parametrene.

Basert på konfidensintervallene; hvilken modell ville du valgt?

I argumentasjonen for det neste punktet så kan du bruke at hvis Modell 1 er riktig så er variansen til $\hat{\beta}_1$ gitt ved

$$\mathrm{V}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Dette trenger du ikke å vise!

(d) Beregn forventningen til $\hat{\gamma}_1$ under forutsetning av at Modell 1 er riktig.

Beregn også varians til $\hat{\gamma}_1$, igjen under forutsetning av at Modell 1 er riktig.

Diskuter hvorfor det i noen tilfeller kan være fornuftig å bruke $\hat{\gamma}_1$ som estimat for β_1 selv om Modell 1 er riktig.

SLUTT