Tentamentsskrivning i **Matematisk Statistik TMA321/MVE395/MVE302**

Tid: den 20 augusti, 2024, 08:30-12:30
Examinator och jour: Erik Broman, mob. 073 7320791,
Hjälpmedel: Typgodkänd miniräknare, 4 A4-sidor egenhändigt skrivna anteckningar (2 ark fram och bak eller 4 ark på en sida) samt utdelade tabeller.

Tentamen består av 6 (8) frågor om sammanlagt 40 (50) poäng. Preliminära betygsgränser är satta till:
F/Kf:
betyg 3: 16 till 23.5 poäng
betyg 4: 24 till 31.5 poäng
betyg 5: 32 eller fler poäng.
TM:
betyg 3: 20 till 29.5 poäng
betyg 4: 30 till 39.5 poäng
betyg 5: 40 eller fler poäng.

OBS! Alla lösningar skall vara väl redovisade, motiverade och fullständiga. Svar skall ges på enklast möjliga form. Talen är ej ordnade efter svårighetsgrad.

1. Låt X vara en slumpvariabel med sannolikhetsfunktion

$$f(k) = Ck \text{ för } k \in \{1, 3, 5, 7\}.$$

(a) Bestäm C så att detta verkligen är en sannolikhetsfunktion . (2p)
(b) Låt $Y = e^{-X}$. Bestäm sannolikhetsfunktionen för Y. (2p)
(c) Beräkna $\mathbb{E}[(\log Y)^2]$. (2p)

Lösning:

(a) Vi har att C måste uppfylla villkoret

$$1 = \sum_k p_X(k) = C(1 + 3 + 5 + 7) = 16C$$

så att $C = 1/16$.

(b) Y kan anta värden i mängden $\{e^{-1}, e^{-3}, e^{-5}, e^{-7}\}$ och vi har att

$$\mathbb{P}(Y = e^{-k}) = \mathbb{P}(X = k) = \frac{k}{16}$$

för $k \in \{1, 3, 5, 7\}$.

(c) Vi har att $\log Y = -X$ så att

$$\mathbb{E}[(\log Y)^2] = \mathbb{E}[(-X)^2] = \mathbb{E}[X^2]$$
$$= \sum_{k=1}^{4} (2k-1)^2 \mathbb{P}(X = 2k-1) = \frac{1 + 3^3 + 5^3 + 7^3}{16} = \frac{496}{16} = 31.$$

2. Herr K jobbar för säljföretaget SäljHets som säljer kompletta datasystem till en mycket hög kostnad. Säljarna ringer till många potentiella kunder, men det är sällan som faktiska kontrakt skrivs.

   SäljHets har olika nischer, och inom en specifik nisch har det sammanlagt skett 2 försäljningar de senaste 5 åren.

   (a) Välj en lämplig modell för att modellera försäljningarna inom ovan nämnda nisch. Skatta den ingående parametern (eller de ingående parametrarna om det finns fler än en) med hjälp av ovanstående data. (3p)

   (b) Herr K's chef har i sin oändliga vishet satt målet att Herr K skall genomföra minst 10 försäljningar inom denna nisch under nästkommande år. Beräkna (så gott du kan) sannolikheten för att detta skall ske. (3p)

   Lösning:

   (a) Vi har här att göra med en Poisson-process. Detta kan inses då vi betraktar antal händelser under en viss tidsrymd. Alternativt kan man förstå från texten att försäljarna gör många försök som alla har väldigt liten chans att lyckas. Dvs, man har en Binomialfördelning med stort n och litet p som man kan Poissonapproximera.

   Vi har alltså en Poisson-process med parameter α försäljningar per år. Antalet försäljningar under fem år blir därför Poisson-fördelat med parameter $\lambda = 5\alpha$. Data (x) ger skattningen

   $$\hat{\lambda}(x) = 2 \text{ så att } \hat{\alpha}(x) = \frac{2}{5}.$$

   Slutsatsen blir att vi modellerar att antalet försäljningarna sker enligt en Poisson-process med intensitet $\alpha = 2/5$ försäljningar per år.

   Sidokommentar: Man kan tycka att en enda datapunkt (2 försäljningar på 5 år) är för lite, men detta är missvisande. I själva verket har man ju samlat in data över en ganska lång period, nämligen 5 år. Det är alltså inte bara en datapunkt.

   (b) Om

   $$X = \text{antalet försäljningar per år}$$

   har vi att $X \sim$Poi$(2/5)$. Vi söker sannolikheten

   $$\mathbb{P}(X \geq 10) = \sum_{k=10}^{\infty} \frac{(2/5)^k}{k!} e^{-2/5} \approx \frac{(2/5)^{10}}{10!} e^{-2/5} (\approx 2 \cdot 10^{-11})$$

   där approximationen kan motiveras med att

   $$\frac{\frac{(2/5)^{11}}{11!} e^{-2/5}}{\frac{(2/5)^{10}}{10!} e^{-2/5}} = \frac{2}{5 \cdot 11} = \frac{2}{55}$$

   så att nästa term är mycket liten jämfört med den första.

3. Låt $X \sim \mathrm{U}([0,1])$ (dvs X är likformigt fördelat på $[0,1]$) och låt sedan $Y \sim \mathrm{Exp}(X)$ (dvs Y är exponentialfördelad med parameter X).

   (a) Bestäm den gemensamma täthetsfunktionen för X, Y. (2p)

   (b) Bestäm $f_X(x|y)$, dvs den betingade täthetsfunktionen för X givet $Y = y$. (3p)

   (c) Låt $Z = XY$, bestäm täthetsfunktionen för Z. (3p)

Lösning:

(a) Vi har att
$$f_X(x) = 1 \text{ för } 0 \le x \le 1$$
och att
$$f_Y(y|x) = xe^{-xy} \text{ för } 0 \le y < \infty$$
så att
$$f_{X,Y}(x,y) = f_Y(y|x)f_X(x) = xe^{-xy} \text{ för } 0 \le x \le 1 \text{ och } 0 \le y < \infty.$$

(b) Vi har att
$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \text{ för } 0 \le y < \infty \text{ om } 0 \le x \le 1.$$
Vidare är
$$\begin{aligned} f_Y(y) &= \int_0^1 xe^{-xy}dx = \left[-x\frac{e^{-xy}}{y}\right]_0^1 + \int_0^1 \frac{e^{-xy}}{y}dx \\ &= -\frac{e^{-y}}{y} + \left[-\frac{e^{-xy}}{y^2}\right]_0^1 = \frac{1-(1+y)e^{-y}}{y^2}, \end{aligned}$$
så att
$$f_X(x|y) = \frac{xe^{-xy}}{\frac{1-(1+y)e^{-y}}{y^2}} = \frac{xy^2e^{-xy}}{1-(1+y)e^{-y}} \text{ för } 0 \le y < \infty \text{ om } 0 \le x \le 1.$$

(c) Vi går via fördelningsfunktionen och observerar först att $0 \le Z < \infty$ och vidare att för $t > 0$,
$$\begin{aligned} F_Z(t) &= \mathbb{P}(Z \le t) = \mathbb{P}(XY \le t) \\ &= \int_0^1 \int_0^\infty I(xy \le t)xe^{-xy}dydx = \int_0^1 \int_0^{t/x} xe^{-xy}dydx \\ &= \int_0^1 \left[-e^{-xy}\right]_0^{t/x} dx = \int_0^1 1 - e^{-t}dx = 1 - e^{-t}, \end{aligned}$$
så att
$$f_Z(t) = e^{-t} \text{ för } 0 \le t < \infty.$$
Vi ser att $Z \sim$Exp(1).

4. **Enbart TM:** Låt $X_1, X_2, \ldots$ vara i.i.d. U($[-1,1]$) dvs likformig fördelade på intervallet $[-1,1]$ och låt

$$S_n = \sum_{k=1}^{n} X_k.$$

(a) Bestäm den momentgenererande funktionen för S_n. (2p)

(b) Använd svaret i (a) för att visa att $S_n/\sqrt{n}$ konvergerar i fördelning. Vilken fördelning konvergerar $S_n/\sqrt{n}$ mot? (3p)

Lösning:

(a) Vi har att

$$M_{S_n}(t) = \prod_{k=1}^{n} M_{X_k}(t) = (M_{X_1}(t))^n$$

och vidare är

$$M_{X_1}(t) = \mathbb{E}\left[e^{-tX_1}\right] = \frac{1}{2}\int_{-1}^{1} e^{-tx}dx = \frac{1}{2}\left[-\frac{e^{-tx}}{t}\right]_{-1}^{1} = \frac{e^t - e^{-t}}{2t}.$$

Vi ser därför att

$$M_{S_n}(t) = \left(\frac{e^t - e^{-t}}{2t}\right)^n.$$

(b) Vi har att

$$M_{S_n/\sqrt{n}}(t) = \mathbb{E}\left[e^{tS_n/\sqrt{n}}\right] = M_{S_n}(t/\sqrt{n}) = \left(\frac{e^{t/\sqrt{n}} - e^{-t/\sqrt{n}}}{2t/\sqrt{n}}\right)^n$$

$$= \left(\frac{(1 + t/\sqrt{n} + t^2/(2n) + t^3/(6n^{3/2})) - (1 - t/\sqrt{n} + t^2/(2n) - t^3/(6n^{3/2})) + O(n^{-5/2})}{2t/\sqrt{n}}\right)^n$$

$$= \left(\frac{2t/\sqrt{n} + t^3/(3n^{3/2}) + O(n^{-5/2})}{2t/\sqrt{n}}\right)^n = \left(1 + \frac{t^2}{6n} + O(n^{-2})\right)^n \overset{n\to\infty}{\longrightarrow} e^{t^2/6},$$

vilket är mgf:en för en normalfördelad slumpvariabel med parametrar 0 och 1/3, dvs

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} N(0, 1/3).$$

Detta kan också inses genom att observera att

$$\mathrm{Var}(X_1) = \frac{1}{3}$$

och sedan att CGS ger

$$\frac{S_n}{\sqrt{n}/\sqrt{3}} \xrightarrow{d} N(0, 1).$$

5. En rektangel har sidor X, Y där $X, Y \sim U([0, \sqrt{\theta}])$ och X, Y är oberoende. För att skatta parametern θ tas ett stickprov bestående av n stycken rektanglar, där den första har sidlängder X_1 och Y_1 och så vidare fram till rektangel nummer n som har sidlängder X_n och Y_n. Alice vill använda dessa stickprov för att skatta parametern θ. Efter en lång betänketid kommer hon fram till två kandidater för skattare. Dessa är

$$\hat{\theta}_1 = 4\overline{X} \cdot \overline{Y} = 4\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)$$

och

$$\hat{\theta}_2 = 4\overline{XY} = \frac{4}{n}\sum_{k=1}^{n} X_k Y_k.$$

Alice antar att $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ alla är oberoende (och så klart med ovan angivna fördelning $U([0, \sqrt{\theta}])$.)

(a) Hjälp Alice att kolla huruvida hennes skattare är väntevärdesriktiga (VVR) skattare av θ. (3p)

(b) Alice undrar om skattarna är konsistenta. Hjälp henne med att kolla om $\hat{\theta}_2$ är konsistent. (3p)

Lösning:

(a) Vi har att

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_1\right] &= 4\mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right)\right] \\ &= 4\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right]\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} Y_k\right] = 4\mathbb{E}[X_1]\mathbb{E}[Y_1] = 4\frac{\sqrt{\theta}}{2}\frac{\sqrt{\theta}}{2} = \theta,\end{aligned}$$

så att $\hat{\theta}_1$ är VVR. Här använde vi oberoende i den andra likheten och standardräkningar för den tredje och fjärde.

Vidare gäller att

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_2\right] &= \mathbb{E}\left[\frac{4}{n}\sum_{k=1}^{n} X_k Y_k\right] = \frac{4}{n}\sum_{k=1}^{n}\mathbb{E}[X_k Y_k] \\ &= \frac{4}{n}\sum_{k=1}^{n}\mathbb{E}[X_k]\mathbb{E}[Y_k] = \frac{4}{n}\sum_{k=1}^{n}\frac{\sqrt{\theta}}{2}\frac{\sqrt{\theta}}{2} = \theta,\end{aligned}$$

så att även $\hat{\theta}_2$ är VVR. Här använde vi oberoende i den tredje likheten.

(b) För att visa konsistens måste vi visa att $\hat{\theta}_2 \xrightarrow{P} \theta$, dvs att för varje $\epsilon > 0$ så gäller att

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_2 - \theta| \geq \epsilon) = 0.$$

Chebyshev's olikhet ger att

$$\mathbb{P}(|\hat{\theta}_2 - \theta| \geq \epsilon) \leq \frac{\mathbb{E}[(\hat{\theta}_2 - \theta)^2]}{\epsilon^2} = \frac{\mathrm{Var}(\hat{\theta}_2)}{\epsilon^2},$$

(där den sista likheten gällde då $\hat{\theta}_2$ var VVR) och vidare så är

$$\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}\left(\frac{4}{n}\sum_{k=1}^{n} X_k Y_k\right) = \frac{16}{n^2}\sum_{k=1}^{n} \mathrm{Var}(X_k Y_k) = \frac{16}{n}\mathrm{Var}(X_1 Y_1).$$

Vi ser alltså att

$$\mathbb{P}(|\hat{\theta}_2 - \theta| \geq \epsilon) \leq \frac{16\mathrm{Var}(X_1 Y_1)}{\epsilon^2 n} \to 0, \text{ då } n \to \infty.$$

Därmed är alltså $\hat{\theta}_2$ konsistent.

6. Sönderfallstiden för en viss radioaktiv isotop anses vara exponentialfördelad med parameter λ. Vid ett experiment vill man försöka hitta värdet på λ och i detta experiment gav $n = 500$ mätpunkter att den genomsnittliga sönderfallstiden var 2.9 ms (millisekunder).

   (a) Hitta ett 99% tvåsidigt konfidensintervall för λ. (6p)

   (b) Man önskar veta vad λ är med en felosäkerhet på högst 1 μs (mikrosekund). Baserat på vad du gjorde i (*a*), hur många mätvärden skulle du tro behövs för att få fram en sådan felosäkerhet (med samma konfidensgrad som i (a))? (2p)

Lösning:

(a) Låt X_k=sönderfallstiden för atomkärna nummer k. Då n är stort kan vi använda CGS (Centrala GränsvärdesSatsen) för att se att

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \overset{d}{\approx} N\left(\mathbb{E}[X_1], \frac{\mathrm{Var}(X_1)}{n}\right).$$

Vidare har vi att då $X_1 \sim \mathrm{Exp}(\lambda)$ så gäller att

$$\mathbb{E}[X_1] = \frac{1}{\lambda} \text{ och att } \mathrm{Var}(X_1) = \frac{1}{\lambda^2}$$

så att

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \overset{d}{\approx} N\left(\frac{1}{\lambda}, \frac{1}{n\lambda^2}\right).$$

Vi använder därför referensvariabeln

$$R = \frac{\overline{X} - \frac{1}{\lambda}}{\frac{1}{\sqrt{n}\lambda}} = \sqrt{n}\left(\lambda\overline{X} - 1\right) \overset{d}{\approx} N(0,1).$$

Vi ser då att (med $Z \sim N(0,1)$)

$$\begin{aligned} 0.99 &= \mathbb{P}(-2.58 \leq Z \leq 2.58) = \mathbb{P}(-2.58 \leq \sqrt{n}\left(\lambda\overline{X} - 1\right) \leq 2.58) \\ &= \mathbb{P}\left(1 - \frac{2.58}{\sqrt{n}} \leq \lambda\overline{X} \leq 1 + \frac{2.58}{\sqrt{n}}\right) = \mathbb{P}\left(\frac{1}{\overline{X}} - \frac{2.58}{\sqrt{n}\cdot\overline{X}} \leq \lambda \leq \frac{1}{\overline{X}} + \frac{2.58}{\sqrt{n}\cdot\overline{X}}\right) \end{aligned}$$

så ett 99% numeriskt K.I. för λ blir då

$$I_\lambda = \frac{1}{\overline{x}} \pm \frac{2.58}{\sqrt{n}\cdot\overline{x}} \approx \frac{1}{2.9} \pm \frac{2.58}{\sqrt{500}\cdot 2.9} \approx [0.305, 0.385]$$

där enheten är 1/ms.

(b) Här vill vi alltså ha ett 99% K.I. med bredd högst 2 μs (1 μs "åt vardera håll"). Enligt (a) blir halva bredden ca

$$\frac{2.58}{\sqrt{n}\cdot 2.9}$$

och om man antar att ett nytt medelvärde vid nya mätningar inte dramatiskt kommer avvika från det vi fick i vår första mätserie skall, vi lösa ekvationen

$$\frac{2.58}{\sqrt{n}\cdot 2.9} = 0.001 \Rightarrow n = \left(\frac{2.58}{2.9\cdot 0.001}\right)^2 \approx 791486.$$

7. **Enbart TM:** Carl är lite av en pedant som anser att om det i affären står att bullen väger 60 g, så skall den också göra det. Den skall inte väga mer, och inte heller mindre. Åtminstone inte i genomsnitt. Carl antar att vikterna är normalfördelade med parametrar μ, σ^2 och vill nu se om $\mu = 60$, dvs han vill testa

$$\begin{aligned} \mathrm{H}_0: &\quad \mu = 60 \\ \mathrm{H}_A: &\quad \mu \neq 60. \end{aligned}$$

Carl köpte därför 100 bullar och vägde dem. Datan visade att samplingsmedlet $\overline{x}$ blev 59.6 (g) och att samplingsvariansen s_x^2 blev 2.9 (g^2).

(a) Hitta en lämplig test-statistika och använd data ovan för att beräkna p-värdet av testet. Genomför testet på signifikansnivån $\alpha = 0.01$. (4p)

(b) Vad ändras om Carl från början istället hade misstänkt att bagarna försökte spara lite av mjölet för att på så sätt tjäna en extra slant? (1p)

Lösning:

(a) Detta är en standardmässig situation och test-statisktikan vi skall använda är

$$T = \frac{\overline{X} - 60}{s/\sqrt{100}} \overset{H_0}{\sim} t(99).$$

Antalet frihetsgrader är här så högt att vi kan approximera fördelningen med $N(0,1)$. p-värdet ges av

$$\text{p-värde} = 2\mathbb{P}(T(X) \geq |T(x)| \, |\mathrm{H}_0 \text{ sann})$$

och vi har i detta fall att

$$T(x) = \frac{\overline{x} - 60}{s/\sqrt{100}} = \frac{59.6 - 60}{\sqrt{2.9}/10} \approx -2.35.$$

Vi får alltså att (med $Z \sim N(0,1)$)

$$\text{p-värde} = 2\mathbb{P}(Z \geq 2.35) \approx 2(1 - 0.991) = 0.018.$$

Då p-värdet överstiger signifikansnivån förkaster vi inte H_0.

(b) Om Carl istället hade haft H_A: $\mu < 60$ så hade

$$\text{p-värde} = \mathbb{P}(T(X) \leq T(x) \, |\mathrm{H}_0 \text{ sann}) = \mathbb{P}(Z \leq -2.35) \approx 0.009$$

och han hade då förkastat H_0 till förmån för H_A.

8. Några forskare vill undersöka huruvida halten av Thorium ökar konduktiviteten i en kopparledning. De blandade därför ihop några legeringar av koppar och Thorium med följande resultat:

Halt Thorium (ppm):	10	20	40	60	80	120
Konduktivitet (S/m):	170	127	172	305	212	402

Forskarna antar ett linjärt samband mellan halten Thorium och konduktiviteten på formen $y = a + bx$. Data kan sammanfattas med att

$$S_{xx} = 8350, S_{xy} = 18280 \text{ och att } S_{yy} = 53095$$

(a) Skatta parametrarna a, b. (1p)

(b) Om man istället hade blandat i motsvarande mängd plutonium så vet man att parametern b i det fallet är 2. Testa på signifikansnivån 5% huruvida (i fallet med Thorium) man får en legering som leder bättre än om man använt plutonium. (4p)

(c) Beräkna förklaringsgraden. Verkar modellen rimlig? (1p)

Lösning:

(a) Vi har att

$$\hat{b} = \frac{S_{xy}}{S_{xx}} \approx 2.19 \text{ och att } \hat{a} = \overline{y} - \hat{b}\overline{x} \approx 111.$$

(b) Vi skall alltså testa

$$\begin{aligned} \mathrm{H}_0: &\quad b = 2 \\ \mathrm{H}_A: &\quad b > 2 \end{aligned}$$

Vi har att

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right) \overset{H_0}{=} N\left(b_0, \frac{\sigma^2}{S_{xx}}\right)$$

där $b_0 = 2$. Vi använder därför test-statistikan

$$T = \frac{\hat{b} - 2}{S_r/\sqrt{S_{xx}}} \overset{H_0}{\sim} t(4),$$

där

$$S_r^2 = \frac{1}{n-2}\sum_{k=1}^{n}(y_k - \hat{y})^2 \approx 3269$$

så att $S_r \approx 57.2$.

Vi förkastar H_0 till förmån för H_A ifall test-statistikan antar stora, positiva värden. Vårt kritiska område för T blir därför $[t_{0.05}(4), \infty)$ där $t_{0.05}(4) \approx 2.13$ enligt tabell. Insatta data ger

$$T(y) = \frac{2.19 - 2}{57.2/\sqrt{8350}} \approx 0.3.$$

Vi kan därför inte förkasta H_0 på signifikansnivån 0.05.

(c) Förklaringsgraden blir

$$R^2 = \frac{Sxy^2}{Sxx \cdot Syy} \approx 0.754$$

vilket inte är jättebra, men inte heller någon katastrof. Modellen verkar hygglig, men så klart behövs mer data.