Tentamentsskrivning i **Matematisk Statistik TMA321**

Tid: den 23 augusti, 2017
Hjälpmedel: Typgodkänd miniräknare, egenhändigt skriven formelsamling om två A4 fram och bak (dvs 4 sidor), samt utdelade tabeller.

Tentamen består av 8 frågor om sammanlagt 50 poäng. Preliminära betygsgränser är satta till:
betyg "3": 20 till 29 poäng
betyg "4": 30 till 39 poäng
betyg "5": 40 eller fler poäng.

OBS! Alla lösningar skall vara väl redovisade och motiverade. Talen är ej ordnade efter svårighetsgrad.

1. Antag att (X, Y) är likformigt fördelad på enhetsdisken dvs området $\{x^2 + y^2 \leq 1\}$.

   (a) Ange den gemensamma sannolikhetstätheten för (X, Y). (2p)

   (b) Ange marginalfördelningarna för X och Y. (2p)

   (c) Beräkna $\mathbb{E}\left[X^2 Y I(Y \geq 0)\right]$. (3p)

   Lösning:

   (a) Enhetscirkeln har area π och därmed gäller att

$$f(x,y) = \begin{cases} \frac{1}{\pi} & \text{om } x^2 + y^2 \leq 1 \\ 0 & \text{annars.} \end{cases}$$

   (b) Vi har:

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} f_{X,Y}(x,y)dy = \begin{cases} \frac{2\sqrt{1-x^2}}{\pi} & \text{om } -1 \leq x \leq 1 \\ 0 & \text{annars.} \end{cases}$$

   På samma sätt:

$$f_Y(y) = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} f_{X,Y}(x,y)dx = \begin{cases} \frac{2\sqrt{1-y^2}}{\pi} & \text{om } -1 \leq y \leq 1 \\ 0 & \text{annars.} \end{cases}$$

(c) Vi får att

$$\mathbb{E}\left[X^2 Y I(Y \geq 0)\right] = \frac{1}{\pi} \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} x^2 y I(y \geq 0) dy dx$$
$$= \frac{1}{\pi} \int_{-1}^{1} \int_{0}^{\sqrt{1-x^2}} x^2 y dy dx = \frac{1}{\pi} \int_{-1}^{1} x^2 \left[\frac{y^2}{2}\right]_0^{\sqrt{1-x^2}} dx$$
$$= \frac{1}{2\pi} \int_{-1}^{1} x^2(1-x^2) dx = \frac{1}{\pi}\left[\frac{x^3}{6} - \frac{x^5}{10}\right]_{-1}^{1} = \frac{1}{\pi}\left(\frac{1}{3} - \frac{1}{5}\right) = \frac{2}{15\pi}.$$

2. Två kompisar, Kalle och Ada, står på ett hustak och kastar vattenballonger på förbipasserande. Sannolikheten att Kalle träffar kallar vi för p_K och sannolikheten att Ada träffar kallar vi för p_A. Vi antar också att sannolikheten för träff är oberoende mellan kasten. De kastar varannan gång och Ada får börja.

(a) De bestämmer sig för att tävla mot varandra. Tävlingen är slut så fort någon har träffat. Vad är sannolikheten att Ada vinner? Kom ihåg att Ada börjar. (3p)

(b) Vid ett senare tillfälle kastar de tills först Ada träffar och sedan Kalle träffar omedelbart därefter (så om Ada träffar och Kalle sedan missar fortsätter de). Vad är sannolikheten att de behöver exakt k kast tillsammans? (3p)

(c) Sista gången gör de på följande sätt. Ada kastar tills första gången hon träffar, och Kalle kastar tills första gången han träffar. Vad är sannolikheten att de behöver sammanlagt k kast? (3p)

Lösning:

(a) Sannolikheten att Ada vinner i omgång l är $p_A(1-p_A)^{l-1}(1-p_K)^{l-1}$. Vi får att sannolikheten att Ada vinner blir

$$\sum_{l=1}^{\infty} p_A(1-p_A)^{l-1}(1-p_K)^{l-1}$$
$$= p_A \sum_{l=0}^{\infty} (1-p_A)^l (1-p_K)^l = \frac{p_A}{1-(1-p_A)(1-p_K)}.$$

(b) Låt X vara antalet kast de behöver och låt l vara omgången de vinner i. Vi får att

$$\mathbb{P}(X = 2l) = p_A p_K (1 - p_A p_K)^{l-1},$$

där $1 - p_A p_K$ är sannolikheten att inte båda träffar i en fix omgång. Vi får att

$$\mathbb{P}(X = k) = \begin{cases} p_A p_K (1 - p_A p_K)^{k/2-1} & \text{för } k \geq 2 \text{ jämnt} \\ 0 & \text{för alla andra } k. \end{cases}$$

(c) För att de sammanlagt skall behöva k kast behöver Ada l och Kalle $k-l$ kast för något $l = 1, \ldots, k-1$. Vi summerar över de olika fallen och får

$$
\begin{aligned}
\mathbb{P}(X=k) &= \sum_{l=1}^{k-1} \mathbb{P}(X_A = l, X_K = k-l) \\
&= \sum_{l=1}^{k-1} p_A(1-p_A)^{l-1} p_K(1-p_K)^{k-l-1} \\
&= \frac{p_A p_K}{(1-p_A)(1-p_K)}(1-p_K)^k \sum_{l=1}^{k-1}(1-p_A)^l(1-p_K)^{-l} \\
&= \frac{p_A p_K}{(1-p_A)(1-p_K)}(1-p_K)^k \sum_{l=1}^{k-1}\left(\frac{1-p_A}{1-p_K}\right)^l \\
&= \frac{p_A p_K}{(1-p_A)(1-p_K)}(1-p_K)^k \sum_{l=0}^{k-2}\left(\frac{1-p_A}{1-p_K}\right)^{l+1} \\
&= p_A p_K (1-p_K)^{k-2} \sum_{l=0}^{k-2}\left(\frac{1-p_A}{1-p_K}\right)^l \\
&= p_A p_K (1-p_K)^{k-2} \frac{1-\left(\frac{1-p_A}{1-p_K}\right)^{k-1}}{1-\left(\frac{1-p_A}{1-p_K}\right)} \\
&= p_A p_K \frac{(1-p_K)^{k-1}-(1-p_A)^{k-1}}{p_A - p_K}.
\end{aligned}
$$

3. Låt $T_i \sim$Exp(λ_i) för $i = 1, 2$.

(a) Beräkna mgf för T_1. (2p)

(b) Använd svaret i uppgift (a) för att bestämma $\mathbb{E}[T_1^n]$ för alla $n \geq 1$. (2p)

(c) Låt X vara sådan att $\mathbb{P}(X=0) = \mathbb{P}(X=1) = 1/2$ och X är oberoende av T_1, T_2. Beräkna mgf för Z där

$$Z = XT_1 + (1-X)T_2.$$

(2p)

Lösning:

(a) Vi har att för $t < \lambda_1$,

$$
\begin{aligned}
M_{T_1}(t) = \mathbb{E}[e^{T_1 t}] &= \int_0^\infty e^{st}\lambda_1 e^{-\lambda_1 s} ds \\
&= \lambda_1 \int_0^\infty e^{(t-\lambda_1)s} ds = \left[\frac{e^{(t-\lambda_1)s}}{t-\lambda_1}\right]_0^\infty = \frac{\lambda_1}{\lambda_1 - t}.
\end{aligned}
$$

(b) Det enklaste är att Taylorutvecka mgf:en.

$$M_{T_1}(t) = \frac{1}{1 - t/\lambda_1} = \sum_{k=0}^{\infty} \left(\frac{t}{\lambda_1} \right)^k = \sum_{k=0}^{\infty} \frac{t^k}{k!} \frac{k!}{\lambda_1^k},$$

så vi kan avläsa koefficienterna till

$$\mathbb{E}[T_1^n] = M_{T_1}^{(n)}(0) = \frac{k!}{\lambda_1^k}.$$

(c) Vi får att för $t < \min(\lambda_1, \lambda_2)$

$$\begin{aligned}
&\mathbb{E}[e^{Zt}] \\
&\quad = \mathbb{E}[e^{Zt}|X = 1]\mathbb{P}(X = 1) + \mathbb{E}[e^{Zt}|X = 0]\mathbb{P}(X = 0) \\
&\quad = \frac{1}{2}\left(\mathbb{E}[e^{T_1 t}] + \mathbb{E}[e^{T_2 t}]\right) = \frac{1}{2}\left(\frac{\lambda_1}{\lambda_1 - t} + \frac{\lambda_2}{\lambda_2 - t}\right).
\end{aligned}$$

4. (a) Låt $X_1, \ldots, X_n$ vara i.i.d. med täthetsfunktion

$$f(x) = \frac{\alpha}{2} \sin(\alpha x) \text{ för } 0 \leq x \leq \pi/\alpha$$

där $\alpha > 0$ är en okänd parameter. Hitta MME (momentskattaren) för α. (3p)

(b) Låt $Y_1, \ldots, Y_n$ vara i.i.d. med täthetsfunktion

$$f(x) = \frac{\beta}{e - 1} e^{\beta x} \text{ för } 0 \leq x \leq 1/\beta.$$

Hitta MLE (maximum likelihood skattaren) för β. (3p)

Lösning:

(a) Vi har att

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{\pi/\alpha} x \frac{\alpha}{2} \sin(\alpha x) = \left[x \frac{-\cos(\alpha x)}{2} \right]_0^{\pi/\alpha} + \int_0^{\pi/\alpha} \frac{\cos(\alpha x)}{2} dx \\
&= \frac{\pi}{2\alpha}(-\cos(\pi)) + \frac{1}{2\alpha}[\sin(\alpha x)]_0^{\pi/\alpha} = \frac{\pi}{2\alpha}.
\end{aligned}$$

Momentmetoden ger att vi skall lösa ekvationen

$$\bar{X} = \frac{\pi}{2\hat{\alpha}} \quad \Leftrightarrow \quad \hat{\alpha} = \frac{\pi}{2\bar{X}}.$$

(b) Likelihooden är

$$\begin{aligned}
&\prod_{k=1}^{n} f(Y_k) = \prod_{k=1}^{n} \frac{\beta}{e - 1} e^{\beta Y_k} I(0 \leq Y_k \leq 1/\beta) \\
&\quad = \frac{\beta^n}{(e - 1)^n} e^{\beta \sum_{k=1}^{n} Y_k} I(0 \leq \min(Y_1, \ldots, Y_n)) I(\max(Y_1, \ldots, Y_n \leq 1/\beta)).
\end{aligned}$$

Vi ser att

$$\frac{\beta^n}{(e-1)^n} e^{\beta \sum_{k=1}^n Y_k} I(0 \leq \min(Y_1, \ldots, Y_n))$$

växer då β växer medans

$$I(\max(Y_1, \ldots, Y_n \leq 1/\beta))$$

blir 0 om $\beta \geq 1/\max(Y_1, \ldots, Y_n)$ så att

$$\hat{\beta} = \frac{1}{\max(Y_1, \ldots, Y_n)}.$$

5. I en studie ville man jämföra två olika processer för tillverkning av formalin. I studien ville man speciellt jämföra halten av toxiska biprodukter och fick därför följande två mätserier över det toxiska innehållet (angett i ppm):

Metod nr 1 (x)	4.21	5.33	5.85	6.40	6.84	5.39	3.81	4.56
	4.24	7.65	4.68	6.05				
Metod nr 2 (y)	4.91	6.39	5.74	7.10	5.08	5.99	7.32	7.3
	6.92	5.79	5.60	5.59				

Vid alla mätningarna användes samma utrustning. Vi har även att

$$\frac{1}{12}\sum_{k=1}^{12}(x_k - \bar{x})^2 \approx 1.273 \qquad \frac{1}{12}\sum_{k=1}^{12}(y_k - \bar{y})^2 \approx 0.6555.$$

(a) Sätt upp ett test för att se om det finns en systematisk skillnad mellan metoderna. Välj signifikansnivån 5%. Antag att alla data är oberoende och kommer från normalfördelningar. (3p)

(b) Upprepa proceduren i (a) men nu under antagandet att data *ej* kommer från en normalfördelning. Hur skall du göra nu? Reflektera över din metod och ditt resultat. Jämför med vad du fick i uppgift (a). (3p)

Lösning:

(a) Vi har två stickprov som uppenbarligen inte är parade. Då det är samma utrustning som används för båda mätserierna följer det att $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Vi får att

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma^2}{12}\right) \text{ och } \bar{Y} \sim N\left(\mu_y, \frac{\sigma^2}{12}\right),$$

så att

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma^2}{6}\right).$$

Därmed blir
$$\frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{s_p/\sqrt{6}} \sim t(22)$$
där
$$\begin{aligned} s_p^2 &= \frac{(12-1)s_X^2+(12-1)s_Y^2}{12+12-2} \\ &= \frac{\sum_{k=1}^{12}(x_k-\bar{x})^2+\sum_{k=1}^{12}(y_k-\bar{y})^2}{22} \approx 1.0519. \end{aligned}$$
Vi använder här att
Om vi väljer signifikansnivå 5% får vi att
$$\begin{aligned} 0.95 &= \mathbb{P}(-2.074 \le T \le 2.074) \\ &= \mathbb{P}\left(-2.074 \le \frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{s_p/\sqrt{6}} \le 2.074\right) \\ &= \mathbb{P}\left(\bar{X}-\bar{Y}-2.074\frac{s_p}{\sqrt{6}} \le \mu_x-\mu_y \le \bar{X}-\bar{Y}+2.074\frac{s_p}{\sqrt{6}}\right) \end{aligned}$$
så att ett 95% K.I. blir
$$I_{\mu_x-\mu_y} = \bar{x}-\bar{y} \pm 2.074\frac{s_p}{\sqrt{6}} \approx [-1.595, 0.142].$$
Då $0 \in I_{\mu_x-\mu_y}$ kan vi därmed inte dra slutsatsen att det är en systematisk skillnad på signifikansnivån 5%.

(b) Vi får använda oss av normalapproximation, dvs centrala gränsvärdessatsen. Detta är egentligen olämpligt då antalet mätdata är för få. Men då det är vårt enda verktyg har vi inget val. Vi bör dock vara noggranna med att påpeka att vårt resultat därmed kommer vara behäftad med en större grad av osäkerhet kring vilka slutsatser vi kan dra.

Den enda skillnaden rent räknemässigt blir att vi antar (på osäkra grunder) att
$$\frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{s_p/\sqrt{6}} \sim N(0,1)$$
så att
$$I_{\mu_x-\mu_y} = \bar{x}-\bar{y} \pm 1.96\frac{s_p}{\sqrt{6}} \approx [-1.547, 0.094].$$
Då $0 \in I_{\mu_x-\mu_y}$ drar vi samma slutsats som innan. Men, då detta grundar sig i ett tvivelaktigt användande av centrala gränsvärdessatsen bör vi förhålla oss skeptiska och rekommendera fortsatta studier.

6. Låt $X_1, \ldots, X_n$ vara i.i.d. exponentialfördelade slumpvariabler. Betrakta hypoteserna
$$H_0 : \lambda = 1 \text{ och } H_1 : \lambda \neq 1.$$

(a) Bestäm den generaliserade likelihood ration. Bestäm även motsvarande förkastningsregion. (3p)

(b) Vad förändras ifall vi byter ut H_1 mot

$$H_1' : \lambda = 2?$$

(2p)

Lösning:

(a) Vi har att

$$\begin{aligned}\text{lik}(\lambda) &= f(X_1, \dots, X_n|\lambda) \\ &= \prod_{k=1}^{n} f(X_k|\lambda) = \prod_{k=1}^{n} \lambda e^{-\lambda X_k} = \lambda^n e^{-\lambda \sum_{k=1}^{n} X_k}.\end{aligned}$$

Vi betraktar den generaliserade LR-kvoten

$$\Lambda = \frac{\text{lik}(1)}{\max_{\lambda \geq 0} \text{lik}(\lambda)} = \frac{\text{lik}(1)}{\text{lik}(\hat{\lambda})},$$

där $\hat{\lambda}$ är MLE:n för λ. Vi har att

$$l(\lambda) = \log(\text{lik}(\lambda)) = n \log \lambda - \lambda \sum_{k=1}^{n} X_k$$

så att

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{k=1}^{n} X_k = 0$$

och därmed blir

$$\hat{\lambda} = \frac{n}{\sum_{k=1}^{n} X_k} = \frac{1}{\bar{X}}$$

ty $l''(\lambda) < 0$. Den sökta LR kvoten blir

$$\Lambda = \frac{\text{lik}(1)}{\text{lik}(\hat{\lambda})} = \frac{e^{-\sum_{k=1}^{n} X_k}}{\bar{X}^{-n} e^{-\frac{1}{\bar{X}} \sum_{k=1}^{n} X_k}} = \left(e\bar{X}e^{-\bar{X}}\right)^n.$$

Vi ser att kvoten är liten om $\bar{X}e^{-\bar{X}}$ är liten. Detta ger oss förkastningsregionen

$$\bar{X} \in [0, c_0] \cup [c_1, \infty)$$

där c_0, c_1 är sådana att

$$\mathbb{P}(\bar{X} \leq c_0) = \mathbb{P}(\bar{X} \geq c_1) = \alpha/2,$$

om vi vill ha signifikansnivån α.

(b) Istället får vi nu

$$\frac{\text{lik}(1)}{\text{lik}(2)} = \frac{e^{-\sum_{k=1}^{n} X_k}}{2^n e^{-2\sum_{k=1}^{n} X_k}} = \left(\frac{e^{\sum_{k=1}^{n} X_k}}{2}\right)^n$$

som är liten om $\sum_{k=1}^{n} X_k$ är liten. Vår förkastningsregion blir nu $[0, c]$ där $\alpha = \mathbb{P}(0 \leq \bar{X} \leq c)$.

7. Karl-Johan vill undersöka viskositeten i en svampstuvning som funktion av mängden maizena han har i stuvningen. Han fick följande data:

Halt:	10	12	14	16	18	20	22	24
Viskositet:	12.4	2.9	7.4	10.8	17.3	43.2	54.2	85.8

Data kan sammanfattas med $S_{xx} = 168, S_{yy} = 5953.88$, och $S_{xy} = 884.2$.

a. Ansätt en linjär regressionsmodell och skatta β_0, β_1. (2p)

b. Bestäm förklaringsgraden och beräkna residualerna. Verkar modellen rimlig? Verkar data rimliga? (3p)

Lösning:

a. Vi har att

$$\beta_1^* = \frac{S_{xy}}{S_{xx}} \approx 5.263, \qquad \beta_0^* = \bar{y} - \beta_1^* \bar{x} \approx -60.223.$$

b. Vi får att förklaringsgraden blir

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} \approx 0.7816128,$$

och vi beräknar residualerna till

datapunkt nr:	1	2	3	4	5	6	7	8
residual:	20	-0.03	-6.06	-13.19	-17.21	-1.84	-1.37	19.71.

Den låga förklaringsgraden och det märkliga utseendet på residualerna får en att misstänka att det snarare föreligger ett annat samband. Kanske exponentiellt eller polynomiellt växande med en högre potens än ett. Vid närmare betraktande verkar dessutom den första mätpunkten vara felaktig då viskositeten bör växa som funktion av mängden maizena.

8. Låt X_1, X_2, X_3 vara oberoende, alla med samma (okända) väntevärde μ. Låt vidare varianserna av X_1, X_2, X_3 vara $\sigma^2, 2\sigma^2$ respektive $3\sigma^2$ där $\sigma^2 > 0$. Låt vidare

$$\hat{\mu}_1 := \frac{X_1 + X_2 + X_3}{3}$$

$$\hat{\mu}_2 := (X_1 X_2)^{1/2}$$

och

$$\hat{\mu}_3 := \frac{X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3}{\frac{11}{6}}$$

vara tre skattningar av μ.

(a) Vilken eller vilka av dessa tre skattningarna är väntevärdesriktiga? (3p)

(b) Vilken av de väntevärdesriktiga skattningarna i (a) har minst varians? (3p)

Lösning:

(a) Vi har att

$$\mathbb{E}[\hat{\mu}_1] = \mathbb{E}\left[\frac{X_1 + X_2 + X_3}{3}\right] = \frac{\mu + \mu + \mu}{3} = \mu,$$

medans

$$\mathbb{E}\left[(X_1X_2)^{1/2}\right] = \mathbb{E}\left[X_1^{1/2}\right]\mathbb{E}\left[X_2^{1/2}\right] < \mathbb{E}[X_1]^{1/2}\mathbb{E}[X_2]^{1/2} = \mathbb{E}[X_1] = \mu,$$

där olikheten kommer sig av att

$$0 < \mathrm{Var}(X_1^{1/2}) = \mathbb{E}[X_1] - \mathbb{E}[X_1^{1/2}]^2,$$

så att $\mathbb{E}[X_1^{1/2}] < \mathbb{E}[X_1]^{1/2}$. Vidare har vi att

$$\mathbb{E}[\hat{\mu}_3] = \mathbb{E}\left[\frac{X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3}{\frac{11}{6}}\right] = \frac{\mu + \frac{1}{2}\mu + \frac{1}{3}\mu}{\frac{11}{6}} = \mu,$$

så $\hat{\mu}_1$ och $\hat{\mu}_3$ båda är väntevärdesriktiga.

(b) Vi beräknar nu varianserna. Vi har att

$$\mathrm{Var}(\hat{\mu}_1) = \frac{1}{9}(\mathrm{Var}(X_1)+\mathrm{Var}(X_2)+\mathrm{Var}(X_3)) = \frac{1}{9}(\sigma^2+2\sigma^2+3\sigma^2) = \frac{2\sigma^2}{3},$$

medan

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_3) &= \frac{6^2}{11^2}(\mathrm{Var}(X_1) + \frac{1}{4}\mathrm{Var}(X_2) + \frac{1}{9}\mathrm{Var}(X_3)) \\ &= \frac{6^2}{11^2}(\sigma^2 + \frac{1}{2}\sigma^2 + \frac{1}{3}\sigma^2) = \frac{6\sigma^2}{11}.\end{aligned}$$

Vi ser att μ_3 har minst varians.