Tentamentsskrivning i **Matematisk Statistik TMA321**

Dag: Onsdagen den 29 augusti, 2018
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. Alexandra har en byrålåda med 6 röda, 4 blå och 4 göna strumpor.

   (a) Beräkna sannolikheten att Alexandra får två strumpor av samma färg om hon drar två strumpor slumpmässigt. (2p)

   (b) Om hon i uppgift a drar två strumpor av samma färg, vad är sannolikheten att dessa är röda? (2p)

   (c) Vad är sannolikheten att hon får ett färgmatchande par om hon istället drar tre strumpor? (2p)

   Lösning:

   (a) Låt G_1, G_2 vara händelserna att den första strumpan är grön respektive att den andra strumpan är grön. Låt B_1, B_2, R_1, R_2 vara de analoga händelserna för blått och rött. Vi söker då

   $$\begin{aligned}&\mathbb{P}((G_1,G_2),(R_1,R_2),(B_1,B_2))\\&\quad=\mathbb{P}(G_1,G_2)+\mathbb{P}(R_1,R_2)+\mathbb{P}(B_1,B_2)\\&\quad=\mathbb{P}(G_2|G_1)\mathbb{P}(G_1)+\mathbb{P}(B_2|B_1)\mathbb{P}(B_1)+\mathbb{P}(R_2|R_1)\mathbb{P}(R_1)\\&\quad=\frac{3}{13}\frac{4}{14}+\frac{3}{13}\frac{4}{14}+\frac{5}{13}\frac{6}{14}=\frac{54}{182}=\frac{27}{91}\approx 0.2967\end{aligned}$$

   (b) Vi söker nu

   $$\begin{aligned}&\mathbb{P}((R_1,R_2)|(G_1,G_2),(R_1,R_2),(B_1,B_2))\\&\quad=\frac{\mathbb{P}((R_1,R_2))}{\mathbb{P}((G_1,G_2),(R_1,R_2),(B_1,B_2))}=\frac{\frac{30}{182}}{\frac{54}{182}}=\frac{15}{27}=\frac{5}{9}.\end{aligned}$$

   (c) Det är enklare att räkna ut sannolikheten att hon drar tre strumpor av olika färger. Vi har att (med uppenbar notation)

   $$\mathbb{P}(G_1,B_2,R_3)=\frac{4}{14}\frac{4}{13}\frac{6}{12}=\frac{96}{2184}=\frac{4}{91}\approx 0.044.$$

Det finns sex olika ordningar man kan få de tre missmatchade strumporna på, så den sökta sannolikheten blir då

$$1 - 6 * \frac{4}{91} = \frac{67}{91} \approx 0.736.$$

2. Carl för statistik över hur långt han går varje dag. Carl kommer fram till att sträckan kan betraktas som en kontinuerlig slumpvariabel X med täthetsfunktion

$$f(x) = \frac{30x^2 - 3x^3}{2500} \text{ för } 0 \leq x \leq 10.$$

Enheten är sjömil.

(a) Beräkna väntevärde och standardavvikelse för X. (2p)

(b) Beräkna sannolikheten att Carl sammanlagt går mindre än eller lika-med 2200 sjömil under en tidsperiod om 365 dagar. Vi kan anta att gångsträckorna är oberoende mellan dagarna. (2p)

(c) Betrakta återigen 365 dagar, och låt Y vara antalet av dessa då Carl går längre än 8 sjömil. Beräkna sannolikheten att Y är större än eller lika med 45. (2p)

Lösning:

(a) Vi har att

$$\mathbb{E}[X] = \int_0^{10} x \frac{30x^2 - 3x^3}{2500} dx = \frac{1}{2500} \left[\frac{30}{4} x^4 - \frac{3}{5} x^5 \right]_0^{10} = 6,$$

och att

$$\mathbb{E}[X^2] = \int_0^{10} x^2 \frac{30x^2 - 3x^3}{2500} dx = \frac{1}{2500} \left[\frac{30}{5} x^5 - \frac{3}{6} x^6 \right]_0^{10} = 40,$$

så att

$$\sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2} = \sqrt{40 - 36} = 2.$$

(b) Låt $X_1, \ldots, X_{365}$ beteckna steglängderna för de 365 dagarna. Vi har enligt centrala gränsvärdessatsen att

$$W = \sum_{i=1}^{365} X_i \approx N(6 * 365, 4 * 365) = N(2190, 1460).$$

Därför får vi att

$$\mathbb{P}(Y \leq 2200) = \mathbb{P}\left(\frac{Y - 2190}{\sqrt{730}} \leq \frac{2200 - 2190}{\sqrt{1460}} \right) \approx \mathbb{P}(Z \leq 0.262) \approx 0.6.$$

(c) Vi har att $Y \sim \text{Bin}(365, p)$ där

$$p = \mathbb{P}(X \geq 8) = \int_8^{10} \frac{30x^2 - 3x^3}{2500} dx = \cdots = \frac{113}{625}.$$

Vi använder CGS för att se att $Y \approx N(365p, 365p(1-p)) \approx N(66, 54.1)$ så att

$$\mathbb{P}(Y \geq 45) = \mathbb{P}\left(\frac{Y - 66}{\sqrt{54.1}} \geq \frac{45 - 66}{\sqrt{54.1}}\right) \approx \mathbb{P}(Z \geq -2.855) = \mathbb{P}(Z \leq 2.855) \approx 0.99785.$$

3. I ett experiment mäts drogkoncentration i blodet som en funktion av läkemedelsdosen. De uppmätta data blev som följer:

dos (mg):	50	100	200	300	400	600	800
konc (μg/liter):	54.5	62.9	132.2	266.8	291.7	427.9	623.5

Data kan sammanfattas med att $S_{xx} = 445000, S_{yy} \approx 258562$ och $S_{xy} = 336890$.

Forskarna ansatta en linjär regressionsmodell $y = \beta_0 + \beta_1 x$ där x är dosen (i mg) och y är blodkoncentrationen (i μg per liter blod). Lös följande uppgifter:

(a) Skatta β_0 och β_1. (2p)

(b) Skapa ett 95% konfidensintervall för β_1 och testa huruvida $\beta_1 = 0.5$ på 95%-nivån. (2p)

(c) Ange förklaringsgraden och beräkna residualerna. Verkar modellen rimlig? (2p)

Lösning

(a) Vi har att

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx 0.757 \text{ och } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \approx 0.673.$$

(b) Vi använder att

$$\frac{\hat{\beta}_1 - \beta_1}{s_r / \sqrt{S_{xx}}} \sim t(5)$$

och får med hjälp av tabell att $t_{0.025}(5) \approx 2.571$ så att

$$\begin{aligned} 0.99 &= \mathbb{P}\left(-2.571 \leq \frac{\hat{\beta}_1 - \beta_1}{s_r / \sqrt{S_{xx}}} \leq 2.571\right) \\ &= \mathbb{P}\left(\hat{\beta}_1 - 2.571 \frac{s_r}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + 2.571 \frac{s_r}{\sqrt{S_{xx}}}\right). \end{aligned}$$

Vi har att

$$s_r^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) \approx 586.3$$

så att $s_r = 24.2$. Ett 95% numeriskt K.I. för β_1 blir då

$$I_{\beta_1} = \hat{\beta}_1 \pm 2.571\frac{s_r}{\sqrt{S_{xx}}} \approx [0.664, 0.850].$$

Då $0.5 \notin I_{\beta_1}$ förkastar vi H_0 på 95% nivån.

(c) Förklaringsgraden blir

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \approx 0.986$$

vilket är extremt bra. Man undrar nästan om de har forskningsfuskat. Residualerna blir

e_1	e_2	e_3	e_4	e_5	e_6	e_7
16	−13.5	−19.9	39	−11.8	−27	17.2

Inga mönster kan ses och vi får anse att modellen är bra. Man bör dock vara något försiktig då antalet mätpunkter är så lågt.

4. Låt (X, Y) vara likformigt fördelade på området

$$A = \{0 \le x \le 2, x \le y \le 2x\}.$$

(a) Hitta den gemensamma täthetsfunktionen för (X, Y) och även de två marginaltäthetsfunktionerna. (3p)

(b) Bestäm de betingade täthetsfunktionerna. (2p)

(c) Ange explicita uttryck för slumpvariablerna $\mathbb{E}[Y|X]$ och $\mathbb{E}[X|Y]$. (2p)

Lösning

(a) Arean av A är

$$\int_0^2 \int_x^{2x} dydx = \int_0^2 xdx = 2.$$

Den gemensamma täthetsfunktionen blir därför

$$f(x, y) = \frac{1}{2} \text{ för } (x, y) \in A.$$

Vi får sedan att

$$f(x) = \int_x^{2x} f(x, y)dy = \frac{x}{2} \text{ för } 0 \le x \le 2$$

och att

$$f(y) = \begin{cases} \int_{y/2}^{y} f(x, y)dx = \frac{y}{4} & \text{för } 0 \le y \le 2, \\ \int_{y/2}^{2} f(x, y)dx = \frac{4-y}{4} & \text{för } 2 \le y \le 4. \end{cases}$$

(b) Denna ges av

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f(x)} = \frac{1/2}{x/2} = \frac{1}{x} \text{ för } x \leq y \leq 2x \text{ då } 0 \leq x \leq 2$$

och

$$f_{X|Y}(x|y) = \frac{f(y,x)}{f(y)} = \begin{cases} \frac{1/2}{y/4} = \frac{2}{y} & \text{för } y/2 \leq x \leq y \text{ då } 0 \leq y \leq 2 \\ \frac{1/2}{(4-y)/4} = \frac{2}{4-y} & \text{för } y/2 \leq x \leq 2 \text{ då } 2 \leq y \leq 4. \end{cases}$$

(c) Vi har att

$$\mathbb{E}[Y|X=x] = \int_x^{2x} y f_{Y|X}(y|x)dy = \int_x^{2x} y\frac{1}{x}dy = \frac{3x}{2} \text{ om } 0 \leq x \leq 2,$$

så att $\mathbb{E}[Y|X] = \frac{3X}{2}$. Vidare blir

$$\mathbb{E}[X|Y=y] = \begin{cases} \int_{y/2}^{y} x\frac{2}{y}dx = \frac{3y}{4} \text{ om } 0 \leq y \leq 2 \\ \int_{y/2}^{2} x\frac{2}{4-y}dx = 1 + \frac{y}{4} \text{ om } 2 \leq y \leq 4. \end{cases}$$

Vi får därför att

$$\mathbb{E}[X|Y] = \frac{3Y}{4}I(0 \leq Y \leq 2) + \left(1 + \frac{Y}{4}\right)I(2 \leq Y \leq 4).$$

5. Låt X_n vara en diskret slumpvariabel med sannolikhetsfunktion

$$\mathbb{P}(X_n = k/n) = \begin{cases} \frac{1}{3n} & \text{om } k = 0, \ldots, n-1 \\ \frac{2}{3n} & \text{om } k = n, \ldots, 2n-1, \end{cases}$$

och låt X vara en kontinuerlig slumpvariabel med täthetsfunktion

$$f(x) = \begin{cases} \frac{1}{3} & \text{om } 0 \leq x \leq 1 \\ \frac{2}{3} & \text{om } 1 < x \leq 2. \end{cases}$$

(a) Hitta den momentgenererande funktionen $M_{X_n}(t)$ för X_n. Skriv den på enklaste form. (2p)

(b) Hitta den momentgenererande funktionen $M_X(t)$ för X. Skriv den på enklaste form. (2p)

(c) Visa att $M_{X_n}(t) \to M_X(t)$ (och därmed att $X_n \xrightarrow{d} X$). (2p)

Lösning:

(a) Vi har att

$$\begin{aligned}
M_{X_n}(t) = \mathbb{E}\left[e^{X_n t}\right] &= \sum_{k=0}^{n-1} e^{kt/n}\frac{1}{3n} + \sum_{k=n}^{2n-1} e^{kt/n}\frac{2}{3n} \\
&= \frac{1}{3n}\sum_{k=0}^{n-1}\left(e^{t/n}\right)^k + \frac{2}{3n}\sum_{k=n}^{2n-1}\left(e^{t/n}\right)^k \\
&= \frac{1}{3n}\frac{1-e^{tn/n}}{1-e^{t/n}} + \frac{2}{3n}e^t\sum_{k=n}^{2n-1}\left(e^{t/n}\right)^{k-n} \\
&= \frac{1}{3n}\frac{1-e^t}{1-e^{t/n}} + \frac{2}{3n}e^t\sum_{k=0}^{n-1}\left(e^{t/n}\right)^k \\
&= \frac{1}{3n}\frac{1-e^t}{1-e^{t/n}} + \frac{2}{3n}e^t\frac{1-e^t}{1-e^{t/n}} = \frac{\left(1-e^t\right)\left(1+2e^t\right)}{3n\left(1-e^{t/n}\right)}.
\end{aligned}$$

(b) Vi har att

$$\begin{aligned}
M_X(t) = \mathbb{E}\left[e^{Xt}\right] &= \int_0^1 e^{xt}\frac{1}{3}dx + \int_1^2 e^{xt}\frac{2}{3}dx \\
&= \frac{1}{3}\left[\frac{e^{xt}}{t}\right]_0^1 + \frac{2}{3}\left[\frac{e^{xt}}{t}\right]_1^2 = \frac{e^t - 1 + 2(e^{2t}-e^t)}{3t} = \frac{2e^{2t}-e^t-1}{3t}.
\end{aligned}$$

(c) Betrakta $M_{X_n}(t)$ och låt $n \to \infty$,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{\left(1-e^t\right)\left(1+2e^t\right)}{3n\left(1-e^{t/n}\right)} \\
&\quad= \lim_{n\to\infty}\frac{\left(1-e^t\right)\left(1+2e^t\right)}{3n\left(1-\left(1+t/n+O(n^{-2})\right)\right)} \\
&\quad= \lim_{n\to\infty}\frac{\left(1-e^t\right)\left(1+2e^t\right)}{3(-t+O(n^{-1}))} = \frac{\left(1-e^t\right)\left(1+2e^t\right)}{3(-t)} \\
&\quad= \frac{1+2e^t-e^t-2e^{2t}}{3(-t)} = \frac{2e^{2t}-e^t-1}{3t}.
\end{aligned}$$

6. Låt X vara en slumpvariabel med täthetsfunktion

$$f(x) = (2+\theta)x^{1+\theta} \text{ för } 0 \le x \le 1,$$

där $\theta > -2$ är parametern vars värde skall skattas.

(a) Hitta momentskattaren (MME) för θ. (2p)

(b) Hitta maximum likelihoodskattaren (MLE) för θ. (3p)

Lösning:

(a) Enligt momentmetoden skall vi börja med att beräkna väntevärdet:

$$\mathbb{E}[X] = \int_0^1 x(2+\theta)x^{1+\theta}dx = (2+\theta)\left[\frac{x^{3+\theta}}{3+\theta}\right]_0^1 = \frac{2+\theta}{3+\theta}.$$

Vi löser sedan ut $\hat{\theta}$ ur ekvationen

$$\frac{2+\hat{\theta}}{3+\hat{\theta}} = \bar{X}$$

vilket ger

$$\hat{\theta} = \frac{3\bar{X}-2}{1-\bar{X}}.$$

(b) Vi börjar med att plocka fram likelihooden

$$L(\theta) = (2+\theta)^n \prod_{k=1}^n X_k^{1+\theta}$$

så att log-likelihooden blir

$$l(\theta) = n\log(2+\theta) + (1+\theta)\sum_{k=1}^n \log(X_k).$$

Derivering ger då

$$l'(\theta) = \frac{n}{2+\theta} + \sum_{k=1}^n \log(X_k) = 0,$$

ur vilket vi får

$$2+\theta = -\frac{n}{\sum_{k=1}^n \log(X_k)} \text{ så att } \hat{\theta} = -2 - \frac{n}{\sum_{k=1}^n \log(X_k)}.$$

Dessutom är

$$l''(\theta) = -\frac{n}{(2+\theta)^2} < 0$$

så vi har hittat ett maximum.

7. Den galna kattladyn Cat har jättemånga katter. Tyvärr har de blivit feta av all utfodring, så Cat bestämmer sig för att sätta dem på diet. Cat testar dieten på nio av sina katter. Hon väger dem innan dieten börjar och sedan igen efter två månader. Hon fick följande data

katt nr:	1	2	3	4	5	6	7	8	9
vikt innan:	4.64	5.12	5.67	5.31	4.68	4.53	5.05	5.12	5.89
vikt efter:	4.3	5.19	5.43	4.99	4.42	4.16	4.63	5.43	6.02

(a) Skatta effekten av Cats diet, och skatta även standardavvikelsen för densamma. (2p)

(b) Ange ett 95% ensidigt konfidensintervall för viktskillnaden. Formulera lämpligt hypotestest motsvarande detta konfidensintervall. Verkar Cats diet ge effekt? (2p)

(c) Vad är p-värdet av ditt test? (2p)

(d) Vilka antaganden måste du göra för att kunna genomföra din analys? (1p)

Lösning:

(a) Data är uppenbarligen parade, och skillnaderna d_k är

$$-0.34 \quad 0.07 \quad -0.24 \quad -0.32 \quad -0.26 \quad -0.37 \quad -0.42 \quad 0.31 \quad 0.13$$

Om skillnaden i vikt betecknas med Δ så kan vi skatta denna med medelvärdet av skillnaderna. Vi får då att

$$\bar{d} = -0.16$$

medan

$$s_\Delta = \sqrt{\sum_{k=1}^{9}(d_k - \bar{d})^2} \approx 0.2608.$$

(b) Vi gör följande hypotestest:

$$H_0 : \Delta = 0 \qquad H_1 : \Delta < 0$$

Vi antar att viktskillnaderna är oberoende och normalfördelade. Då blir

$$\frac{\bar{d} - \Delta}{s_\Delta/\sqrt{9}} \sim t(8).$$

Enligt tabell blir

$$0.95 \approx \mathbb{P}\left(\frac{\bar{d} - \Delta}{s_\Delta/3} \geq -1.86\right) = \mathbb{P}\left(\Delta \leq \bar{d} + 1.86\frac{s_\Delta}{3}\right)$$

så ett numeriskt 95% ensidigt K.I. blir därför

$$I_\Delta = \left(-\infty, -0.16 + 1.86\frac{0.2608}{3}\right] \approx (-\infty, 0.0017].$$

Då $0 \in I_\Delta$ kan vi ej förkasta H_0.

(c) För att beräkna p-värdet observerar vi att under H_0 så är $\Delta = 0$ och

$$\frac{\bar{d} - \Delta_0}{s_\Delta/\sqrt{9}} = \frac{\bar{d}}{s_\Delta/3} \approx -1.8405.$$

Om $T \sim t(8)$ får vi att

$$\mathbb{P}(T \leq -1.8405) = 1 - \mathbb{P}(T \geq -1.8405) = 1 - \mathbb{P}(T \leq 1.8405) \approx 0.05+.$$

p-värdet är mycket nära 0.05, men lite större.

(d) Vi antar att katterna påverkas oberoende av varandra och att viktnedgången är normalfördelad. Båda dessa antaganden kan kritiseras, speciellt den sista om normalfördelning. Det är dock svårt att räkna på metoden utan att använda dessa antaganden (men det finns sätt som inte ingår i kursen).

8. Antag att vi har tre slumpvariabler $X \sim N(\mu, \sigma^2), Y \sim N(\gamma, \sigma^2)$ och $Z \sim N(\mu + \gamma, \sigma^2)$, och låt x, y, z vara oberoende observationer från respektive slumpvariabler. Om vi vill skatta μ kan vi t.ex. använda $\mu^* = x$, eller $\hat{\mu} = (x - y + z)/2$.

   (a) Visa att båda skattarna är väntvärdesriktiga (man får självklart använda att väntevärdet av en normalfördelad slumpvariabel är känd). (2p)

   (b) Vilken av skattarna är effektivast? (2p)

   (c) Kan du hitta en egen väntevärdesriktig skattare som är effektivare än båda de två föreslagna? (3p)

   Lösning:

   (a) Vi har att $\mathbb{E}[\mu^*(X)] = \mathbb{E}[X] = \mu$ samt att

   $$\mathbb{E}[\hat{\mu}(X, Y, Z)] = \frac{\mathbb{E}[(X - Y + Z)]}{2} = \frac{\mu - \gamma + (\mu + \gamma)}{2} = \mu,$$

   så båda skattarna är väntevärdesriktiga.

   (b) Vi skall nu beräkna varianserna. Vi har att $\mathrm{Var}(\mu^*(X)) = \mathrm{Var}(X) = \sigma^2$ medan

   $$\begin{aligned}\mathrm{Var}(\hat{\mu}(X, Y, Z)) &= \frac{\mathrm{Var}(X - Y + Z)}{4} \\ &= \frac{\mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z)}{4} = \frac{3\sigma^2}{4}.\end{aligned}$$

   Därmed är $\hat{\mu}$ effektivare än μ^*.

   (c) Vi kan ansätta $W = \alpha X + (1 - \alpha)(Z - Y)$ så att

   $$\mathbb{E}[W] = \alpha\mu + (1 - \alpha)(\mu + \gamma - \gamma) = \mu,$$

   och därmed är W väntevärdesriktig. Vidare har vi att

   $$\begin{aligned}\mathrm{Var}(W) &= \mathrm{Var}(\alpha X + (1 - \alpha)(Z - Y)) \\ &= \alpha^2 \mathrm{Var}(X) + (1 - \alpha)^2 \mathrm{Var}(Z - Y) \\ &= \alpha^2\sigma^2 + (1 - \alpha)^2 2\sigma^2 = \sigma^2(3\alpha^2 - 4\alpha + 2).\end{aligned}$$

   Detta uttryck är enkelt att minimera, och vi ser att minimat inträffar då $\alpha = 2/3$. Om vi alltså låter

   $$W = \frac{2X}{3} + \frac{Z - Y}{3}$$

får vi att

$$\mathrm{Var}(W) = \sigma^2 \left(3\frac{4}{9} - 4\frac{2}{3} + 2 \right) = \sigma^2 \frac{2}{3}.$$

Skattaren W är alltså mer effektiv.