Tentamentsskrivning i **Matematisk Statistik TMA321**

Tid: den 7 oktober, 2017
Examinatorer: Erik Broman.
Jour: Sandra Eriksson Barman.
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
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. På ett barnkalas finns tre lådor med godis. Låda nummer 1 innehåller 6 goda och 3 äckliga godisbitar, låda nummer 2 innehåller 7 goda och 2 äckliga medan låda nummer 3 innehåller 8 goda och 1 äcklig.
Kalle är först i kön för att plocka godis. Han väljer först en låda slumpmässigt och därefter plockar han två slumpmässigt valda godisar ur den valda lådan.

   (a) Beräkna sannolikheten att Kalle får minst en äcklig godisbit. (3p)

   (b) Givet händelsen i (a), beräkna sannolikheten att Kalle får två äckliga godisbitar. (3p)

   (c) Om Kalle fick två äckliga godisbitar, vad är då sannolikheten att de kom från låda nummer 1? (2p)

   Lösning:

   (a) Låt X vara antalet äckliga godisar Kalle får, och låt $L \in \{1, 2, 3\}$ vara nummret på lådan han valde. Vi söker $\mathbb{P}(X \geq 1)$ och har att

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \sum_{i=1}^{3} \mathbb{P}(X = 0|L = i)\mathbb{P}(L = i).$$

   Vidare gäller att

$$\mathbb{P}(X = 0|L = 1) = \frac{30}{72},\ \mathbb{P}(X = 0|L = 2) = \frac{42}{72},\ \mathbb{P}(X = 0|L = 3) = \frac{56}{72},$$

   så att

$$\mathbb{P}(X \geq 1) = 1 - \frac{30 + 42 + 56}{3 * 72} = \frac{216 - 128}{216} = \frac{88}{216} \approx 0.4074.$$

(b) Vi söker

$$\mathbb{P}(X = 2|X \geq 1) = \frac{\mathbb{P}(X = 2)}{\mathbb{P}(X \geq 1)},$$

och har att

$$\mathbb{P}(X = 2) = \sum_{i=1}^{3} \mathbb{P}(X = 2|L = i)\mathbb{P}(L = i).$$

Vidare gäller att

$$\mathbb{P}(X = 2|L = 1) = \frac{6}{72}, \quad \mathbb{P}(X = 2|L = 2) = \frac{2}{72}, \quad \mathbb{P}(X = 2|L = 3) = 0$$

så att

$$\mathbb{P}(X = 2) = \frac{6 + 2}{3 * 72} = \frac{8}{216} = \frac{1}{27} \approx 0.037.$$

Vi får då

$$\mathbb{P}(X = 2|X \geq 1) = \frac{\mathbb{P}(X = 2)}{\mathbb{P}(X \geq 1)} = \frac{8}{88} = \frac{1}{11} \approx 0.0909.$$

(c) Vi söker

$$\mathbb{P}(L = 1|X = 2) = \mathbb{P}(X = 2|L = 1)\frac{\mathbb{P}(L = 1)}{\mathbb{P}(X = 2)} = \frac{6}{72}\frac{1}{3}\frac{3 * 72}{8} = \frac{6}{8} = 0.75.$$

2. En kontinuerlig slumpvariabel har följande täthetsfunktion:

$$f_X(x) = c\cos(x) \quad \text{för } x \in [-\pi/2, \pi/2]$$

(a) Bestäm c så att detta blir en täthetsfunktion. (2p)

(b) Beräkna fördelningsfunktionen $F_X(x)$. (2p)

(c) Låt $Y = X^2$, hitta täthetsfunktionen för Y. (3p)

Lösning:

(a) Vi har att

$$1 = \int_{-\pi/2}^{\pi/2} c\cos(x)dx = c\left[\sin(x)\right]_{-\pi/2}^{\pi/2} = 2c$$

så att $c = 1/2$.

(b) Vi har att

$$F_X(x) = \frac{1}{2}\int_{-\pi/2}^{x} \cos(y)dy = \frac{1 + \sin(x)}{2} \quad \text{för } x \in [-\pi/2, \pi/2].$$

Dessutom är $F_X(x) = 0$ för $x \leq -\pi/2$ och $F_X(x) = 1$ för $x \geq \pi/2$.

(c) Vi börjar med att observera att $Y \geq 0$. Vi har att för varje $y \geq 0$ gäller

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = 2\mathbb{P}(0 \leq X \leq \sqrt{y}),$$

där den sista likheten följer av att X är symmetrisk kring 0. Vi får att

$$F_Y(y) = 2\int_0^{\sqrt{y}} \frac{1}{2}\cos(z)dz = [\sin(z)]_0^{\sqrt{y}} = \sin(\sqrt{y}) \text{ för } \sqrt{y} \leq \pi/2,$$

medan $F_Y(y) = 0$ för $y \leq 0$ och $F_Y(y) = 1$ för $y \geq \pi^2/4$. Därför blir

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{2\sqrt{y}}\cos(\sqrt{y}) \text{ för } y \in [0, \pi^2/4].$$

3. En liten solfångare kan ges två olika konfigurationer, A och B. För att utreda vilken som är effektivast inhandlades två identiska exemplar som ställdes på olika konfiguartioner.

Därefter mättes energiutbytet (i kWh) under 11 försöksdagar och detta blev som följer:

Dag nr	1	2	3	4	5	6	7	8	9	10	11
Konfig. A	9.4	7.7	13.6	15.7	12	8.3	13.2	9.2	9.9	10.3	8.1
Konfig. B	9.9	9.5	11.3	16.6	12.3	7	12.8	9.5	13.5	13.1	6.8

B:

Man kan anta att mätvärdena dras ur normalfördelningar.

(a) Bestäm ett 95% konfidensintervall för skillnaden i energiutbyte mellan konfiguration A och B. Redovisa noga alla eventuella antaganden som du gör. (4p)

(b) Testa hypotesen att det inte är någon skillnad mellan konfigurationerna med 5% felrisk. Gör ditt bästa för att hitta en ungefärligt p-värde för ditt test. Vilken felrisk skulle du behöva acceptera för att förkasta ditt H_0? (3p)

Lösning:

(a) Uppenbarligen har vi att göra med en situation med parade data. Då data antags komma från normalfördelningar gäller detta även skillnaderna. Vi får därför att

$$Z_i \sim N(\Delta, \sigma^2)$$

där σ^2 är okänd. Vi får då att

$$\bar{Z} \sim N\left(\Delta, \frac{\sigma^2}{11}\right) \Longrightarrow \frac{\bar{Z} - \Delta}{s/\sqrt{11}} \sim t_{10},$$

där så klart

$$s^2 = \frac{1}{10}\sum_{k=1}^{11}(Z_k - \bar{Z}).^2$$

Därmed ges ett 95% K.I. av

$$0.95 = \mathbb{P}\left(-2.228 \le \frac{\bar{Z}-\Delta}{s/\sqrt{11}} \le 2.228\right)$$

så att vi får intervallet

$$I_\Delta = \bar{Z} \pm 2.228\frac{s}{\sqrt{11}}.$$

Vi har följande tabell med uppmätta skillnader:

Dag nr	1	2	3	4	5	6	7	8	9	10	11
Skilln. (z)	-0.5	-1.8	2.3	-0.9	-0.3	1.3	0.4	-0.3	-3.6	-2.8	1.3

och därmed blir $s^2 \approx 3.193$, så att $s \approx 1.787$ och vårt numeriska K.I. blir därför

$$I_\Delta = \bar{z} \pm 2.228\frac{1.787}{\sqrt{11}} = -0.4455 \pm 2.228\frac{1.787}{\sqrt{11}} \approx [-1.646, 0.755].$$

(b) Då $0 \in I_\Delta$ kan vi ej förkasta H_0 med felrisk 5%.
Låt nu

$$T = \frac{\bar{Z}-\Delta}{s/\sqrt{11}}$$

vara vår teststatistika, där $\Delta = 0$ under H_0. Vi frågar oss då vad sannolikheten är $|T|$ antar ett värde högre än

$$\frac{0.4455}{1.787/\sqrt(11)} \approx 0.827.$$

Vi får att

$$\mathbb{P}(|T| \ge 0.827) \approx \mathbb{P}(T \ge 0.879) + \mathbb{P}(T \le -0.879) = 0.4,$$

där 0.879 är det värdet i våra tabeller som är närmast 0.827. Det är rimligt att p-värdet ligger kring 0.4 till säg 0.45. Vi skulle alltså få ha en felrisk på ca 40-45% för att förkasta H_0.

4. Låt $X \sim \text{Geom}(p)$ (dvs geometrisk fördelning) och $T \sim \text{Exp}(\lambda)$ (dvs exponentialfördelning).

   (a) Beräkna de momentgenererande funktionerna för X och T. (3p)

(b) Låt $X_n \sim \text{Geom}(\lambda/n)$ och betrakta $T_n = X_n/n$. Visa med hjälp av momentgenererande funktioner att $T_n \xrightarrow{d} T$ (dvs att T_n konvergerar mot T i fördelning). (3p)

Lösning:

(a) Vi har att $\mathbb{P}(X = k) = (1-p)^{k-1}p$ så att

$$\begin{aligned} M_X(t) = \mathbb{E}[e^{Xt}] &= \sum_{k=1}^{\infty} e^{kt}(1-p)^{k-1}p \\ &= pe^t \sum_{k=1}^{\infty} (e^t(1-p))^{k-1} = pe^t \sum_{k=0}^{\infty} (e^t(1-p))^k = \frac{pe^t}{1-(1-p)e^t}, \end{aligned}$$

där det sista steget gäller för $t < -\log(1-p)$ så att summan blir konvergent.

Vidare blir

$$\begin{aligned} M_T(t) = \mathbb{E}\left[e^{Tt}\right] &= \int_0^{\infty} e^{st}\lambda e^{-\lambda s} ds \\ &= \lambda \int_0^{\infty} e^{(t-\lambda)s} ds = \lambda \left[-\frac{e^{(t-\lambda)s}}{t-\lambda}\right]_0^{\infty} = \frac{\lambda}{\lambda - t}, \end{aligned}$$

ifall $t < \lambda$.

(b) Vi har här att

$$\begin{aligned} M_{X_n/n}(t) &= \mathbb{E}[e^{(X_n/n)t}] = \mathbb{E}[e^{X_n(t/n)}] = M_{X_n}(t/n) \\ &= \frac{\frac{\lambda}{n}e^{t/n}}{1-(1-\frac{\lambda}{n})e^{t/n}} = \frac{\lambda e^{t/n}}{n-(n-\lambda)e^{t/n}} = \frac{\lambda e^{t/n}}{n(1-e^{t/n}) + \lambda e^{t/n}} \\ &= \frac{\lambda e^{t/n}}{n(1-(1+t/n+O(n^{-2}))) + \lambda e^{t/n}} = \frac{\lambda e^{t/n}}{-t+O(n^{-1})) + \lambda e^{t/n}} \\ &\longrightarrow \frac{\lambda}{\lambda - t}. \end{aligned}$$

5. I en storstad har man satt upp en mätstation för ozonhalten vid marknivån (Y). Man intresserar sig bl.a. för hur halten i fråga beror av lufttemperaturen (x). De första åtta mätdagarna låg i en period med likartade vindförhållanden men med stigande temperatur. Följande mätvärden (tagna kl. 17.00) erhölls:

Dag	1	2	3	4	5	6	7	8
Temperatur (°C)	9.3	10.2	12.6	15.5	13.3	16.6	19.7	22.7
Ozonhalt (mg/m^3)	31	34	41	50	46	52	64	73

Förutsätt nu att ozonhalterna beror linjärt av temperaturen, och att de uppmätts med oberoende slumpmässiga mätfel enligt modellen

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, 8$$

där $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$.

(a) Skatta parametrarna α och β. (2p)

(b) Beräkna ett 95% konfidensintervall för β. (2p)

(c) Vilken ozonhalt bör vi förvänta oss vid temperaturen 20°C? (2p)

Lösning:

(a) Minstakvadratskattningarna av α och β ges av

$$\alpha^* = \bar{y} - \beta^* \bar{x}, \quad \beta^* = \frac{S_{xy}}{S_{xx}}$$

Man finner $\alpha^* = 2.46$, $\beta^* = 3.10$, där vi utnyttjat $S_{xx} = 148.37$, $S_{xy} = 459.5$, $\bar{x} = 14.988$, $\bar{y} = 48.875$.

(b) Ett intervall ges av

$$I_\beta = \left[\beta^* \pm t_p(n-2)s/\sqrt{S_{xx}}\,\right].$$

Med

$$s = \sigma^* = \frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2} = 1.283$$

och $t_{0.025}(6) = 2.45$, $S_{yy} = 1432.88$ följer $I_\beta = [2.84,\ 3.36]$.

(c) Vår gissning av ozonhalten vid temperaturen 20°C ges av att stoppa in $x = 20$ i vårt skattade samband, dvs

$$Y(20) \approx \alpha^* + \beta^* \cdot 20 = 64.4.$$

6. Låt (X, Y) ha gemensam täthetsfunktion

$$f(x, y) = cxye^{xy^2} \text{ för } 0 \leq x \leq 1, \text{ och } x \leq y \leq 1,$$

där c är en konstant.

(a) Hitta marginalfördelningarna för X och Y. (3p)

(b) Hitta de betingade fördelningarna för X givet Y respektive för Y givet X. (2p)

(c) Bestäm $\mathbb{E}[X|Y]$. (3p)

Lösning:

(a) Vi har att

$$f_X(x) = \int_x^1 cxye^{xy^2} dy = c\left[\frac{e^{xy^2}}{2}\right]_x^1 = c\frac{e^x - e^{x^3}}{2} \text{ för } 0 \leq x \leq 1.$$

Vidare gäller att

$$\begin{aligned}
f_Y(y) &= \int_0^y cxye^{xy^2} dx = c\left(\left[x\frac{e^{xy^2}}{y}\right]_0^y - \int_0^y \frac{e^{xy^2}}{y} dx\right) \\
&= ce^{y^3} - c\left[\frac{e^{xy^2}}{y^3}\right]_0^y = ce^{y^3} - c\frac{e^{y^3}-1}{y^3} = \frac{c}{y^3}\left(y^3e^{y^3} - e^{y^3} + 1\right) \text{ för } 0 \leq y \leq 1.
\end{aligned}$$

(b) De betingade fördelningarna ges av

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{cxye^{xy^2}}{\frac{c}{y^3}\left(y^3e^{y^3} - e^{y^3} + 1\right)} = \frac{xy^4e^{xy^2}}{\left(y^3e^{y^3} - e^{y^3} + 1\right)} \text{ för } 0 \leq x \leq y \leq 1,$$

och

$$f_{Y|X}(y|y) = \frac{f(x,y)}{f_X(x)} = \frac{cxye^{xy^2}}{c\frac{e^x - e^{x^3}}{2}} = \frac{2xye^{xy^2}}{e^x - e^{x^3}} \text{ för } 0 \leq x \leq y \leq 1.$$

(c) Vi har att

$$\begin{aligned}
\mathbb{E}[X|Y=y] &= \int_0^y x f_{X|Y}(x|y) dx = \int_0^y \frac{x^2y^4e^{xy^2}}{\left(y^3e^{y^3} - e^{y^3} + 1\right)} dx \\
&= \frac{1}{y^3e^{y^3} - e^{y^3} + 1} \int_0^y x^2y^4e^{xy^2} dx \\
&= \frac{1}{y^3e^{y^3} - e^{y^3} + 1} \left(\left[x^2y^2e^{xy^2}\right]_0^y - 2\int_0^y xy^2e^{xy^2} dx\right) \\
&= \frac{1}{y^3e^{y^3} - e^{y^3} + 1} \left(y^4e^{y^3} - 2\left[xe^{xy^2}\right]_0^y + 2\int_0^y e^{xy^2} dx\right) \\
&= \frac{1}{y^3e^{y^3} - e^{y^3} + 1} \left(y^4e^{y^3} - 2ye^{y^3} + 2\left[\frac{e^{xy^2}}{y^2}\right]_0^y\right) \\
&= \frac{1}{y^3e^{y^3} - e^{y^3} + 1} \left(y^4e^{y^3} - 2ye^{y^3} + 2\frac{e^{y^3}-1}{y^2}\right) \\
&= \frac{e^{y^3}(y^6 - 2y^3 + 2) - 2}{y^2(y^3e^{y^3} - e^{y^3} + 1)}
\end{aligned}$$

så därmed blir

$$\mathbb{E}[X|Y] = \frac{e^{Y^3}(Y^6 - 2Y^3 + 2) - 2}{Y^2(Y^3e^{Y^3} - e^{Y^3} + 1)}.$$

7. Låt X vara en diskret slumpvariabel med sannolikhetsfunktion

$$\mathbb{P}(X = k) = \frac{1}{N} \text{ för } k = 2, 4, \ldots, 2N$$

där N är en heltalsvärd parameter som vi vill skatta.

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

(b) Hitta maximum likelihood skattaren (MLE) för X. (2p)

(c) Antag att vi har en oberoende mätserie bestående av 6 slumptal dragna ur fördelningen ovan som ser ut som följer:

Tal nr	1	2	3	4	5	6
Resultat	4	8	2	24	10	14

Vad blir nu de numeriska värdena på era skattare? Diskutera och analysera ditt resultat. (2p)

Lösning:

(a) Vi har att

$$\mathbb{E}[X] = \sum_{k=1}^{N} 2k\mathbb{P}(X = 2k) = 2\sum_{k=1}^{N} k\frac{1}{N} = \frac{2}{N}\frac{N(N+1)}{2} = N + 1$$

så vår momentskattare ges av sambandet

$$\bar{X} = \hat{N} + 1 \text{ så att } \hat{N} = \bar{X} - 1.$$

(b) Likelihood funktionen blir nu

$$\begin{aligned} L(X_1, \ldots, X_n) &= \prod_{l=1}^{n} \frac{1}{N} I(X_l \in \{2, 4, \ldots, 2N\}) \\ &= N^{-n} I(X_1, \ldots, X_n \in \{2, 4, \ldots, 2N\}), \end{aligned}$$

och vi ser att detta maximeras (som funktion av N) då N är så liten som möjligt utan att något $X_l > 2N$. Därmed upppnås maximum av

$$\hat{N} := \max(X_1, \ldots, X_n)/2.$$

(c) Vi har att $\bar{x} = 62/6 = 31/3$ så att

$$\hat{N} = \frac{31}{3} - 1 = \frac{28}{3} \approx 9.33333.$$

Detta är inte ett speciellt bra resultat då vi dels vet att N måste vara ett heltal, och dels vet att N måste vara minst 12.
Å andra sidan ger vår MLE att

$$\hat{N} = 24/2 = 12,$$

vilket åtminstone är en rimlig gissning.

8. Beskriv kortfattat men kärnfullt idén bakom ett likelihood ratio test. (2p)