Tentamentsskrivning i **Matematisk Statistik MVE091**

Tid: den 5 januari, 2024, 08:30-13:30
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 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, motiverade och fullständiga. Svar skall ges på enklast möjliga form. Talen är ej ordnade efter svårighetsgrad.

1. Saba har en låda med fyra gula strumpor, tre svarta och tre blå. Saba tar två strumpor slumpmässigt.

   (a) Vad är sannolikheten att Saba drar två gula strumpor? (2p)

   (b) Vad är sannolikheten att Saba drar två strumpor av olika färg? (2p)

   (c) Givet att Saba inte drog någon gul strumpa, vad är sannolikheten att hon får två strumpor av samma färg? (3p)

   Lösning:

   (a) Antal sätt att välja två gula strumpor ur en samling av fyra är $\binom{4}{2}$ och det totala antalet sätt att välja två strumpor är $\binom{10}{2}$ så att den sökta sannolikheten blir

   $$\frac{\binom{4}{2}}{\binom{10}{2}} = \frac{6}{45} = \frac{2}{15} (\approx 0.1333).$$

   (b) Både antalet svarta och blå par är $\binom{3}{2}$. Vi får därför att den sökta sannolikheten blir

   $$1 - \frac{\binom{4}{2} + \binom{3}{2} + \binom{3}{2}}{\binom{10}{2}} = 1 - \frac{6+3+3}{45} = 1 - \frac{12}{45} = 1 - \frac{4}{15} = \frac{11}{15} (\approx 0.7333).$$

   (c) Låt P vara händelsen att Saba drar ett par, och låt G vara händelsen att Saba drar minst en gul strumpa. Vi söker då

   $$\mathbb{P}(P|G^c) = \frac{\mathbb{P}(P \cap G^c)}{\mathbb{P}(G^c)}.$$

Det finns sammanlagt 6 strumpor som inte är gula så att

$$\mathbb{P}(G^c) = \frac{\binom{6}{2}}{\binom{10}{2}} = \frac{15}{45} = \frac{1}{3}.$$

Storleken på (dvs antalet utfall i) händelsen $P \cap G^c$ är $\binom{3}{2} + \binom{3}{2}$ motsvarande antalet matchande par i svart respektive blått. Vi får att

$$\mathbb{P}(P \cap G^c) = \frac{\binom{3}{2} + \binom{3}{2}}{\binom{10}{2}} = \frac{6}{45} = \frac{2}{15}.$$

Till slut ser vi då att

$$\mathbb{P}(P|G^c) = \frac{\mathbb{P}(P \cap G^c)}{\mathbb{P}(G^c)} = \frac{\frac{2}{15}}{\frac{1}{3}} = \frac{6}{15} = \frac{2}{5}.$$

2. Låt X vara en slumpvariabel med täthetsfunktion

$$f_X(x) = Cx \text{ för } x \in [1, 5].$$

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

(b) Låt $Y = e^X$. Bestäm täthetsfunktionen för Y. (3p)

(c) Bestäm $\mathbb{E}[Y]$. (3p)

Lösning:

(a) Vi har att C måste uppfylla villkoret

$$1 = \int_{-\infty}^{\infty} f_X(x)dx = \int_1^5 Cxdx = C\left[\frac{x^2}{2}\right]_1^5 = C\frac{25-1}{2} = 12C$$

så att $C = 1/12$.

(b) Vi går som vanligt (när vi har kontinuerliga slumpvariabler) via fördelningsfunktionen. Vi ser då att

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(e^X \leq y) = \mathbb{P}(X \leq \log y) = F_X(\log y).$$

Vidare är (för $x \in [1, 5]$)

$$F_X(x) = \mathbb{P}(X \leq x) = \int_1^x Csds = C\left[\frac{s^2}{2}\right]_1^x = C\frac{x^2-1}{2} = \frac{x^2-1}{24}.$$

Vi observerar också att om $X \in [1, 5]$ så gäller att $Y \in [e^1, e^5]$. Därför ser vi att för $y \in [e^1, e^5]$

$$F_Y(y) = F_X(\log y) = \frac{(\log y)^2 - 1}{24}$$

så att

$$f_Y(y) = F_Y'(y) = \frac{d}{dy}\frac{(\log y)^2 - 1}{24} = \frac{1}{y}\frac{2\log y}{24} = \frac{\log y}{12y},$$

för $e \leq y \leq e^5$.

(c) Vi har att

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{e^1}^{e^5} y \frac{\log y}{12y} dy = \frac{1}{12} \int_{e^1}^{e^5} (\log y) dy \\
&= \left\{ z = \log y,\ dz = \frac{1}{y} dy \Rightarrow dy = y dz = e^z dz \right\} = \frac{1}{12} \int_1^5 z e^z dz
\end{aligned}$$

vilket man även kan se genom att observera att

$$\mathbb{E}[Y] = \mathbb{E}\left[e^X\right] = \frac{1}{12} \int_1^5 e^x x dx.$$

Vi fortsätter och ser att

$$\begin{aligned}
\mathbb{E}[Y] &= \frac{1}{12} \int_1^5 e^x x dx = \frac{1}{12} \left[x e^x\right]_1^5 - \frac{1}{12} \int_1^5 e^x dx \\
&= \frac{5e^5 - e^1}{12} - \frac{1}{12} \left[e^x\right]_1^5 = \frac{5e^5 - e^1}{12} - \frac{e^5 - e^1}{12} = \frac{4e^5}{12} = \frac{e^5}{3}.
\end{aligned}$$

3. Ottilia gillar att kasta snöbollar på sina lärare när de kommer till skolan. Hon står på skolans tak och kastar på dem när de går in genom huvudporten på morgonen.

   Antag att Ottilia har tio snöbollar och kastar en boll vardera mot tio olika lärare. Antag också att Ottilia träffar med sannolikhet $4/5$. Låt X vara antalet träffar en viss morgon.

   (a) Vilken fördelning har X? Vilka (om några) antaganden behöver du göra för att detta skall stämma? (2p)

   (b) För att Ottilia skall vara nöjd vill hon träffa minst nio av lärarna. Vad är sannolikheten att Ottilia träffar minst nio av lärarna en viss morgon? (2p)

   (c) Antag att Ottilia upprepar denna procedur under sammanlagt 30 mornar med snö. Vad är sannolikheten att hon blir nöjd minst 15 av dessa? Motivera ditt tillvägagångssätt. (3p)

   Lösning:

   (a) Vi måste anta (och det är fullt rimligt) att Ottilia träffar varje snöboll oberoende av varandra. Vi har då att $X \sim$Bin$(10, 4/5)$ dvs att X är binomialfördelad med parametrar $n = 10$ och $p = 4/5$.

   (b) Vi söker

$$\begin{aligned}
\mathbb{P}(X \geq 9) &= \mathbb{P}(X = 9) + \mathbb{P}(X = 10) \\
&= \binom{10}{9} p^9 (1-p) + \binom{10}{10} p^{10} (1-p)^0 \\
&= 10 \frac{4^9}{5^9} \frac{1}{5} + \frac{4^{10}}{5^{10}} = \frac{10 \cdot 4^9 + 4^{10}}{5^{10}} (\approx 0.3758).
\end{aligned}$$

(c) Här kan vi använda normalapproximation (dvs Centrala gränsvärdessatsen). Vi bör dock kolla ifall det är rimligt. Om Y är antalet dagar hon är nöjd så är $Y \sim$Bin$(30, q)$ där $q = \mathbb{P}(X \geq 9)$ och vi har att $30q > 11$ och våra vanliga tumregler ger då att vi kan använda normalapproximation.
Observera att $\mathbb{E}[Y] = 30q$ och att $\text{Var}(y) = 30q(1-q)$. Vi söker

$$\mathbb{P}(Y \geq 15) = \mathbb{P}\left(\frac{Y - 30q}{\sqrt{30q(1-q)}}\right) \geq \mathbb{P}\left(\frac{15 - 30q}{\sqrt{30q(1-q)}}\right)$$
$$\approx \mathbb{P}(Z \geq 1.4044) = \mathbb{P}(Z \leq -1.404) \approx 0.0749$$

där vi använde tabell och där som vanligt i dessa sammanhang $Z \sim$N$(0, 1)$.

4. Rebecca Schrödinger betraktar två atomer som hon väntar på skall sönderfalla. Deras sönderfallstider T_1 och T_2 anses vara exponentialfördelade med parametrar $\beta_1 = 1$ respektive $\beta_2 = 1/2$.

(a) Rebeccas katt påstår att den gemensamma täthetsfunktionen för T_1 och T_2 ges av

$$f_{T_1,T_2}(t, s) = 2tse^{-2ts} \text{ för } t, s > 0.$$

Rebecca har inte för vana att ta råd av sin katt, men undrar om det överhuvud taget är möjligt det katten påstår? (3p)

(b) Rebecca känner ändå att det är mer rimligt att sönderfallstiderna är oberoende. Under detta antagande, beräkna sannolikheten att den första sönderfallstiden (T_1) är högst hälften av den andra (T_2). (3p)

Lösning

(a) Nej, och detta kan man se på fler olika sätt. T.ex. så har vi att marginaltäthetsfunktionen för T_1 i sådana fall skulle bli

$$\begin{aligned} f_{T_1}(t) &= \int_0^\infty f_{T_1,T_2}(t, s)ds \\ &= \int_0^\infty 2tse^{-2ts}ds = \left[s\left(-e^{-2ts}\right)\right]_0^\infty + \int_0^\infty e^{-2ts}ds \\ &= 0 + \left[-\frac{1}{2t}e^{-2ts}\right]_0^\infty = \frac{1}{2t} \text{ för } t > 0. \end{aligned}$$

Detta är dock ingen täthetsfunktion då vi uppenbarligen har att

$$\int_0^\infty \frac{1}{2t}dt = \infty.$$

(b) Om de är oberoende gäller att

$$f_{T_1,T_2}(t, s) = f_{T_1}(t)f_{T_2}(s) = \frac{1}{\beta_1}e^{-t/\beta_1}\frac{1}{\beta_2}e^{-s/\beta_2} = 2e^{-t-2s}.$$

Vi söker $\mathbb{P}(T_1 \leq T_2/2)$ och har då att

$$\begin{aligned}
\mathbb{P}(T_1 \leq T_2/2) &= \int\int_{t \leq s/2} f_{T_1,T_2}(t,s)dtds \\
&= \int_0^\infty \int_0^{s/2} 2e^{-t-2s}dtds = \int_0^\infty \left[-2e^{-t-2s}\right]_0^{s/2} ds \\
&= \int_0^\infty 2e^{-2s} - 2e^{-s/2-2s}ds = \int_0^\infty 2e^{-2s} - 2e^{-5s/2}ds \\
&= \left[-e^{-2s} + \frac{4}{5}e^{-5s/2}\right]_0^\infty = 1 - \frac{4}{5} = \frac{1}{5}.
\end{aligned}$$

5. Låt X vara en slumpvariabel med täthetsfunktion

$$f_X(x) = \theta e^{-\theta} x^{\theta-1} \text{ för } 0 \leq x \leq e,$$

där $\theta > 0$ är en parameter som skall skattas. Använd maximum likelihood metoden för att hitta en skattare för θ. (5p)

Lösning: Vi har att likelihooden ges av

$$\begin{aligned}
L(\theta) = f(x_1, \ldots, x_n|\theta) &= \prod_{k=1}^n f(x_k|\theta) \\
&= \prod_{k=1}^n \theta e^{-\theta} x_n^{\theta-1} = \theta^n e^{-n\theta} \prod_{k=1}^n x_k^{\theta-1}.
\end{aligned}$$

och log-likelihooden blir därför

$$l(\theta) = \log L(\theta) = n\log\theta - n\theta + (\theta - 1)\sum_{k=1}^n \log x_k.$$

Vi vill maximera denna med avseende på θ, och söker därför nollställen för derivatan. Vi har att

$$l'(\theta) = \frac{n}{\theta} - n + \sum_{k=1}^n \log x_k,$$

och villkoret $0 = l'(\theta)$ ger oss att

$$\frac{1}{\theta} = 1 - \frac{1}{n}\sum_{k=1}^n \log x_k \text{ så att } \theta = \frac{1}{1 - \frac{1}{n}\sum_{k=1}^n \log x_k}.$$

För att kontrollera att detta är ett minimum observerar vi att

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

eftersom $\theta > 0$.

6. Ett speditionsföretag får en massa paket med mystiskt innehåll addresserat till en Dr Fronkensteen. Enligt beskrivningen på paketen skall de väga i genomsnitt 9 kilo, men det står ingenting i beskrivningen vilken varians vikterna har. Speditionsföretaget bestämmer sig för att göra en kontroll och väger därför sex paket med följande resultat.

Paket nr:	1	2	3	4	5	6
Vikt (kg):	7.8	10	13.1	8.5	10.7	9.6

Bestäm ett 90% konfidensintervall för σ^2. (5p)

Lösning:

Vi ser att (med $n = 6$)

$$\hat{\mu} = \overline{x} = 9.93 \text{ och } s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2 \approx 3.45.$$

Enligt föreläsningen så är

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) = \chi^2(5).$$

Tabellerna ger att $\chi^2_{0.95}(5) \approx 1.15$ och att $\chi^2_{0.05}(5) \approx 11.1$. Vi ser därför att

$$\begin{aligned} 0.9 &= \mathbb{P}\left(\chi^2_{0.95}(5) \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{0.05}(5)\right) \\ &= \mathbb{P}\left(\frac{(n-1)s^2}{\chi^2_{0.05}(5)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{0.95}(5)}\right), \end{aligned}$$

så att

$$I_{\sigma^2} = \left[\frac{(n-1)s^2}{\chi^2_{0.05}(5)}, \frac{(n-1)s^2}{\chi^2_{0.95}(5)}\right] \approx [1.56, 15.07],$$

blir ett 90% konfidensintervall för σ^2.

7. Företaget Zkan tillverkar en mängd olika matprodukter och måste alltid deklarera innehållet i dessa. För att mäta halten (i %) av mättade fetter har de använt en traditionell metod som eventuellt inte var så exakt. De testar därför en ny metod som kanske gör precisare mätningar.

   För att ta reda på om det finns någon skillnad mellan metoderna tar de åtta olika matprodukter, och för varje produkt mäter de med båda metoderna. Resultatet blev som följer.

Matprodukt nr:	1	2	3	4	5	6	7	8
% Mättat fett (Metod 1):	2.2	13.1	1.4	6.4	8.9	19.4	12.0	3.9
% Mättat fett (Metod 2):	2.5	13.3	2.1	7.5	10	19.1	12.4	3.4

Zkan undrar ifall metoderna systematiskt producerar olika resultat. Ange därför ett 95% konfidensintervall för skillnaderna. (5p)

Lösning: Det viktiga här är att observera att det rör sig om en situation med parade stickprov. Vi börjar därför med att göra en tabell över dessa skillnader:

Matprodukt nr:	1	2	3	4	5	6	7	8
Skillnad (z):	-0.3	-0.2	-0.7	-1.1	-1.1	0.3	-0.4	0.5

Vår punktskattning av skillnaden blir

$$\hat{\Delta} = \frac{1}{8}\sum_{k=1}^{8} z_k \approx -0.375.$$

Vi får här använda referensvariabeln

$$R = \frac{\hat{\Delta} - \Delta}{s(z)/\sqrt{n}} \sim t(n-1)$$

där $n = 8$ och (som vanligt)

$$s^2(z) = \frac{1}{n-1}\sum_{k=1}^{8}(z_k - \overline{z})^2 = 0.345.$$

Kvantilerna blir här $t_{0.025}(7) \approx 2.365$ så att vi får

$$\begin{aligned}
0.95 &= \mathbb{P}\left(-t_{0.025}(7) \le R \le t_{0.025}(7)\right) \\
&= \mathbb{P}\left(-t_{0.025}(7) \le \frac{\hat{\Delta} - \Delta}{s/\sqrt{8}} \le t_{0.025}(7)\right) \\
&= \mathbb{P}\left(\hat{\Delta} - t_{0.025}(7)\frac{s}{\sqrt{8}} \le \Delta \le \hat{\Delta} + t_{0.025}(7)\frac{s}{\sqrt{8}}\right),
\end{aligned}$$

så ett 95% numeriskt konfidensintervall blir

$$I_\Delta = \hat{\Delta} \pm t_{0.025}(7)\frac{s}{\sqrt{8}} = -0.375 \pm 2.365\frac{\sqrt{0.345}}{\sqrt{8}} \approx [-0.867, 0.116].$$

8. Ett företag IsoHeatPressure (IHP) håller på att ta fram ett nytt värmeisolerande material. De har tagit patent på en ny typ av fiber (kallat ArmoFib$^{\text{TM}}$) som de blandar in i materialet. Företaget hoppas att ArmoFib$^{\text{TM}}$skall öka resistensen mot slag (värmeisolerande material är notoriskt spröda vilket är ett stort problem).

   IHP gjorde 7 blandningar med olika halter av ArmoFib$^{\text{TM}}$och formade dessa till plattor. Sedan slog de på plattorna tills dessa krackelerade, och noterade anslagsenergin på det slag som fick plattorna att krackelera.

De fick följande data:

x: Halt ArmoFib$^{\text{TM}}$(%):	2	2.5	3	3.5	4	4.5	5
y: Energi (kJ):	1.8	1.8	2	2.4	2.6	2.7	2.7

och dessa data kan sammanfattas i att

$$\sum_{k=1}^{7}(x_k - \overline{x}).^2 = 7, \quad \sum_{k=1}^{7}(x_k - \overline{x})(y_k - \overline{y}) \approx 2.55 \quad \sum_{k=1}^{7}(y_k - \overline{y}).^2 \approx 1.01.$$

(a) Använd data för att anpassa bästa möjliga regressionslinje. (1p)

(b) Gör ett hypotestest för att undersöka huruvida motståndskraften mot slag påverkas av halten ArmoFib$^{\text{TM}}$. Använd signifikansnivån 0.01. (4p)

(c) Antag att man vill ha en platta som kan motså anslagsenergin 6 kJ. Hur mycket ArmoFib$^{\text{TM}}$säger din modell att man skall ha? Reflektera över ditt svar. (2p)

Lösning:

(a) Vi ansätter $y = \beta_0 + \beta_1 x$ och har att

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx 0.364 \text{ och } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \approx 1.01$$

(b) Vi skall alltså testa

$$\begin{aligned} H_0 &: \quad \beta_1 = 0 \\ H_1 &: \quad \beta_1 \neq 0 \end{aligned}$$

på $\alpha = 0.01$. Vi vet att $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$ så att under H_0 (dvs om $\beta_1 = 0$) är

$$\hat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{S_{xx}}\right).$$

Vi väljer därför test-statistikan

$$T(E) = \frac{\hat{\beta}_1(E)}{S_r(E)/\sqrt{S_{xx}}} \sim t(n-2) = t(5),$$

där $t(5)$ är noll-fördelningen och

$$S_r^2 = \frac{1}{n-2}\sum_{k=1}^{n}(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k)^2 (\approx 0.016).$$

Om H_1 stämmer så bör $T(E)$ avvika från 0 med ganska mycket, antingen i positiv eller negativ riktning. P-värdet ges därför av

$$p\text{-värde} = 2\mathbb{P}(T(E) \geq |T(\epsilon)|)$$

och här blir

$$T(\epsilon) = \frac{\hat{\beta}_1(\epsilon)}{S_r(\epsilon)/\sqrt{S_{xx}}} \approx \frac{0.364}{\sqrt{0.016}/\sqrt{7}} \approx 7.64.$$

Tabell ger att

$$\begin{aligned} p\text{-värde} &= 2\mathbb{P}(T(E) \geq |T(\epsilon)|) \\ &\approx \mathbb{P}(T(E) \geq 7.64) \leq \mathbb{P}(T(E) \geq 5.959) = 0.0005. \end{aligned}$$

Då p-värdet är mindre än α förkastar vi H_0 på signifikansnivån 0.01.

(c) Vi vill alltså hitta halten x så att

$$\beta_0 + \beta_1 x = y = 6.$$

Detta innebär att x skall väljas så att

$$x = \frac{6 - \beta_0}{\beta_1} \approx \frac{6 - \hat{\beta}_0}{\hat{\beta}_1} = \frac{6 - 1.01}{0.364} \approx 13.7.$$

Reflektion: Vårt svar (13.7) ligger mycket långt utanför det område där data som modellen är baserad på finns. Det finns en överhängande risk att lineariteten inte är "giltig" när man går så långt utanför området, och därmed bör vi tillskriva denna förutseelse mycket låg tillförsikt. I princip bör vi ignorera det helt och istället göra fler experiment.