LINKÖPINGS UNIVERSITET

Matematiska institutionen

20 oktober 2020, kl. 14.00-18.00
Examinator: Jörg-Uwe Löbus

Tillåtna hjälpmedel är en räknare, formelsamling i matematisk statistisk utgiven av MAI, och ett ytterligare formelblad (framsida och baksida).

Betygsgränser 3: minst 8 poäng, 4: minst 11.5 poäng, 5: minst 15 poäng (av 18)

(1) Låt $X_1, \ldots, X_{100}$ vara oberoende och binomialfördelade slumpvariabler där varje X_i är $Bin(10, 0.4)$-fördelad. Låt vidare $Y_1, \ldots, Y_{70}$ vara oberoende och Poissonfördelade slumpvariabler där varje Y_i är $Po(6)$-fördelad. Sätt X och Y till

$$X = \sum_{i=1}^{100} X_i \quad \text{and} \quad Y = \sum_{i=1}^{70} Y_i \,.$$

(a) Använd centrala gränsvärdessatsen för att bestämma de approximativa fördelningarna av X och Y. (1.5p)

(b) Använd resultatet av (a)-delen för att beräkna approximativt $P(Y \geq X)$. Om du inte kunde lösa (a)-delen använd istället $X \sim N(390, 15)$ och $Y \sim N(390, 20)$. (1.5p)
Ledning: Beakta $P(Y \geq X) = P(Y - X \geq 0)$. Bestäm fördelningen av $Y - X$.

(2) År 2008 hade i Sverige 20% av befolkningen en årsinkomst som översteg 350.000 kronor och 10% av befolkningen hade en årsinkomst som översteg 550.000 kronor.

(a) Om man fick veta (2008) att en svensk person tjänade mer än 350.000 kronor om året, vad är sannolikheten att personen tjänade mer än 550.000 kronor? (1.5p)

(b) Bland svenska män hade 30% en årsinkomst som översteg 350.000 kronor. Antag att det bodde lika många kvinnor som män i Sverige. Om man fick veta att en svensk person tjänade mer än 350.000 kronor om året, vad är sannolikheten att det var en man? (1.5p)

(3) En person åker först med buss 1, sedan med buss 2. Väntetiderna X och Y är oberoende och likformigt fördelade över intervallet $(0, 10)$ respektive $(0, 8)$, där enheten är minuter. Bestäm sannolikheten att personen får vänta sammanlagt minst 16 minuter på de båda bussarna. (3p)
Ledning: Den stokastiska vektorn (X, Y) är likformigt fördelad över ett rektangelformat område.

(4) Ett företag som tillverkar skruvar vill se om man håller samma kvalitet som förra året. Då valde man slumpmässigt ut 1000 skruvar och fann 16 defekta. I år var man

mer ambitiös och valde ut 2000 skruvar, varvid 45 befanns vara defekta. Undersök med ett approximativt 95% konfidensintervall om andelar defekta skruvar (p_1 resp p_2) kan anses vara lika vid de två tillfällena.

(a) Ange det generella (tvåsidiga) konfidensintervallet $I_{p_2-p_1}$ för skillnaden i dessa andelar p_1 och p_2. (1p)

(b) Beräkna ett tvåsidigt 95%-konfidensintervall för skillnaden i andelarna p_1 och p_2 baserat på uppgifterna i texten. (1.5p)

(c) Får man hävda att det är statistiskt säkerställt att det finns en skillnad mellan de båda andelarna, på konfidensnivå 95%? (0.5p)

(5) Vid studier av syrsor har man funnit att frekvensen av vingrörelser växer med temperaturen. En modell för detta är:

$$y = \alpha + \beta \cdot x + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

där x = temperatur i $^\circ F$ och y = antal svängningar per sekund. I femton temperaturkammare med skilda temperaturer x uppmättes vingfrekvensen y.

(a) Bestäm den skattade regressionslinjen $y = \alpha^*_{obs} + \beta^*_{obs} \cdot x$. (1p)

(b) Pröva hypotesen vid signifikansnivån $\alpha = 0.05$,

$$H_0 : \beta = 0 \quad \text{mot} \quad H_1 : \beta \neq 0\,.$$

(1p)

(c) Bilda ett 95% konfidensintervall för β. (1p)

Räknehjälp:

$$\begin{aligned} \bar{x} &= 16.65, \quad \bar{y} = 80.04 \\ S_{xx} &= 40.558 \\ S_{yy} &= 629.832 \\ S_{xy} &= 133.476 \end{aligned}$$

(6) Låt X vara en stokastisk variabel med täthetsfunktion

$$f_X(x) = \begin{cases} \theta \cdot x^{\theta-1} & \text{för} \quad 0 \le x \le 1 \\ 0 & \text{annars} \end{cases}.$$

(a) Bestäm väntevärdet och variansen för X. (1p)

(b) Härled maximumlikelihood-skattningen av θ baserad på oberoende observationer $x_1, \ldots, x_n$ av X. (2p)

Lösningar

(1) (a) Vi har att

$$\begin{array}{llll} X_i \text{ är Bin}(10, 0.4) & \Rightarrow & E(X_i) = 10 \cdot 0.4 = 4 & V(X_i) = 10 \cdot 0.4 \cdot 0.6 = 2.4 \\ Y_i \text{ är } Po(6) & \Rightarrow & E(Y_i) = 6 & V(Y_i) = 6 \end{array}$$

Centrala gränsvärdessatsen ger att X är approximativt

$$N(100 \cdot 4, \sqrt{100 \cdot 2.4}) = N(400, 15.49) - \text{fördelad}$$

och Y är approximativt

$$N(70 \cdot 6, \sqrt{70 \cdot 6}) = N(420, 20.49) - \text{fördelad}\,.$$

(b) Vi skriver om den sökta sannolikheten: $P(Y \geq X) = P(Y - X \geq 0) = 1 - P(Y - X < 0)$ Vad har $Y - X$ för fördelning? Väntevärde och varians blir

$$\begin{array}{l} E(Y - X) = E(Y) - E(X) = 420 - 400 = 20 = \mu \\ V(Y - X) = V(Y) + V(X) = 420 + 240 = 660 = \sigma^2 \end{array}$$

Normalisering till $N(0, 1)$ ger

$$P(Y - X < 0) = P\left(\frac{Y - X - \mu}{\sigma} < \frac{0 - \mu}{\sigma}\right) = P\left(Z < \frac{-20}{\sqrt{660}}\right)$$

$$\approx \Phi(-0.78) = 1 - \Phi(0.78) = 1 - 0.7823$$

Därmed blir $P(Y \geq X) \approx 1 - (1 - 0.7823) = 0.7823$

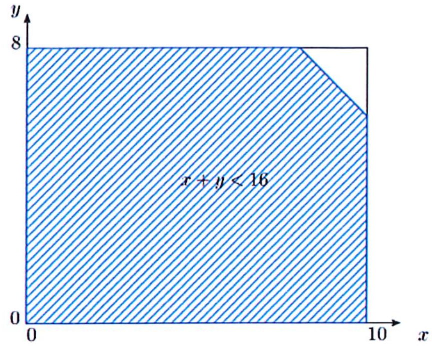
(2) (a) Låt G beteckna händelsen att personens årsinkomst översteg 350.000 kronor och H händelsen att årsinkomsten översteg 550.000 kronor. Vi har då att $P(G) = 0.2$ och $P(H) = 0.1$. Sökt: $P(H|G)$. Eftersom $H \subseteq G$, så får vi

$$P(H|G) = \frac{P(G \cap H)}{P(G)} = \frac{P(H)}{P(G)} = \frac{0.1}{0.2} = \frac{1}{2}.$$

(b) Låt M beteckna händelsen att personen är man. Vi har då att $P(M) = 0.5$ och $P(G|M) = 0.3$. Sökt: $P(M|G)$. Med hjälp av Bayes sats fås

$$P(M|G) = \frac{P(M \cap G)}{P(G)} = \frac{P(G|M)P(M)}{P(G)} = \frac{0.3 \cdot 0.5}{0.2} = 0.75.$$

(3) Sökt är $P(X + Y \geq 16)$. Den vita arean motsvarar $\{X + Y \geq 16\}$.



Paret av väntetider (X, Y) för de två bussarna är likformigt fördelat över rektangeln $[0, 10] \times [0, 8]$.

Vi får

$$P(X + Y \geq 16) = \frac{\text{vit area}}{\text{total area}} = \frac{2 \cdot 2/2}{80} = \frac{1}{40}.$$

(4) Låt $X_i \sim Bin(n_i, p_i)$, $i = 1, 2$.

(a) Konfidensintervallet för $p_2 - p_1$ är

$$I_{p_2 - p_1} = \left(p^*_{2\,obs} - p^*_{1\,obs} \pm \lambda_{\alpha/2} \sqrt{\frac{p^*_{1\,obs}(1 - p^*_{1\,obs})}{n_1} + \frac{p^*_{2\,obs}(1 - p^*_{2\,obs})}{n_2}} \right)$$

där för $i = 1, 2$

x_i betecknar utfallet av X_i i en konkret undersökning,

$$p^*_{i\,obs} = \frac{x_i}{n_i}.$$

(b)

$$\begin{aligned}
I_{p_2 - p_1} &= \left(p^*_{2\,obs} - p^*_{1\,obs} \pm \lambda_{\alpha/2} \sqrt{\frac{p^*_{1\,obs}(1 - p^*_{1\,obs})}{n_1} + \frac{p^*_{2\,obs}(1 - p^*_{2\,obs})}{n_2}} \right) \\
&= \left(\frac{45}{2000} - \frac{16}{2000} \pm 1.96 \cdot \sqrt{\frac{0.016 \cdot (1 - 0.016)}{1000} + \frac{0.0225 \cdot (1 - 0.0225)}{2000}} \right) \\
&= 0.0065 \pm 0.0101\,.
\end{aligned}$$

(c) Eftersom $0 \in I_{p_2 - p_1}$, så är det med konfidensgrad 95% inte statistiskt säkerställt att det finns en skillnad.

(5) (a) Enligt formlarna är skattningarna

$$\beta^*_{obs} = S_{xy}/S_{xx} = 133.476/40.558 = 3.29$$
$$\alpha^*_{obs} = \bar{y} - \beta^*_{obs} \cdot \bar{x} = 80.04 - 3.29 \cdot 16.65 = 25.26$$

$$y = 25.26 + 3.29 \cdot x\,.$$

(b) Utfallet av teststorheten är

$$t = \frac{\beta^*_{obs} - 0}{\sigma^*_{obs}/\sqrt{S_{xx}}} = \frac{3.29}{3.83/\sqrt{40.558}} = 5.47$$

där

$$(\sigma^*_{obs})^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 14.66$$

dvs.

$$\sigma^*_{obs} = 3.83\,.$$

Det kritiska området är

$$C = \left(-\infty, -t(n-2)_{\alpha/2}\right] \cup \left[t(n-2)_{\alpha/2}, +\infty\right) = (-\infty, -2.16] \cup [2.16, +\infty)\,.$$

där $n = 15$. Eftersom $t \in C$ så får vi förkasta H_0.

(c)

$$I_\beta = \beta^*_{obs} \pm t(n-2)_{\alpha/2} \cdot \frac{\sigma^*_{obs}}{\sqrt{S_{xx}}} = 3.29 \pm 1.3 = (1.99, 4.59)\,.$$

(6) (a) Väntevärdet beräknas

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 x\theta x^{\theta-1}\,dx = \left[\theta \frac{x^{\theta+1}}{\theta+1}\right]_0^1 = \frac{\theta}{\theta+1}\,.$$

Dessutom får vi

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^1 x^2\theta x^{\theta-1}\,dx = \left[\theta \frac{x^{\theta+2}}{\theta+2}\right]_0^1 = \frac{\theta}{\theta+2}$$

och

$$Var(X) = E[X^2] - (E[X])^2 = \frac{\theta}{\theta+2} - \left(\frac{\theta}{\theta+1}\right)^2\,.$$

(b) Vi får likelihoodfunktionen

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = \theta^n \cdot (x_1 \cdot \ldots \cdot x_n)^{\theta-1}$$

och logaritmerar vi så får vi

$$\ln L(\theta) = n\ln\theta + (\theta-1)\sum_{i=1}^{n} \ln x_i.$$

Deriverar vi detta,

$$\frac{d \ln L(\theta)}{d\theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i\,,$$

sätter derivatan $= 0$ och löser ut θ så fås

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln x_i}.$$

Kontroll ger även att andraderivatan är negativ,

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = -\frac{n}{\theta^2},$$

vilket garanterar att det verkligen är en max-punkt för likelihoodfunktion vi hittat.