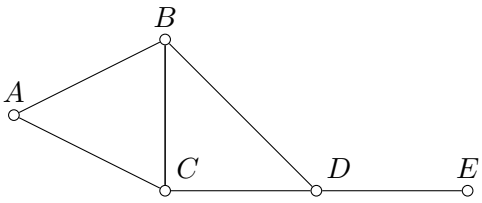
Matematiska Vetenskaper 23 maj 2023
Chalmers Tekniska Högskola Examinator: Tony Johansson

MVE302, MVE395, TMA321
Sammanfattande exempel

För repetitionsföreläsningarna delade jag upp kursen i 9 delar, och förberedde en uppgift för varje del. Då jag inte hann med alla under föreläsningen, kommer de istället här. Jag ger er uppgifterna här och lösningarna längre ner, om ni skulle vilja jobba på dem själv.

1. Betrakta följande nätverk av vägar:



Varje enskild väg är öppen med sannolikhet p, oberoende av de andra. Vad är sannolikheten att det finns en öppen väg från A till E som går genom B?

2. En stokastisk variabel X har täthetsfunktion

$$f_X(x) = c \sin x, \quad 0 < x < \pi.$$

(a) Vad är c?

(b) Vad är $\mathrm{Var}\,(X)$?

3. Låt $\{\tau_1, \tau_2, \dots\}$ vara en Poissonprocess med intensitet $\lambda > 0$. Vid varje impuls slås en sexsidig tärning (slaget tar ingen tid, det påbörjas och avslutas vid tidpunkt τ_k). Låt T vara tidpunkten för den första sexan som slås under processen. Vad har T för fördelning?

4. Låt X, Y vara diskreta slumpvariabler med $Y \sim \mathrm{Poi}(2)$ och (betingat) $X \sim \mathcal{U}\{0, 1, \dots, Y\}$, alltså

$$f_Y(y) = \frac{e^{-2}2^y}{y!}, \quad f_{X|Y}(x \mid y) = \frac{1}{y+1},$$

för heltal $0 \le x \le y$. Beräkna $\mathrm{Cov}\,(X, Y)$.

5. Låt $X_n \sim \Gamma(n, 1)$. Hitta konstanter a, b sådana att

$$\lim_{n\to\infty} \mathbb{P}(an - b\sqrt{n} \le X_n \le an + b\sqrt{n}) = 0.99.$$

6. Anta att en fördelning $\mathcal{F}(\theta)$ har parametriserad täthetsfunktion

$$f_\theta(x) = 2(x-\theta), \quad \theta < x < \theta + 1.$$

Hitta en ML-skattning för θ baserad på datapunkter $x_1 = 0.4, x_2 = 0.7, x_3 = 0.5$.

7. Hitta ett 99% konfidensintervall för μ, med normalfördelad data med okänt σ^2 och stickprovsdata

$$\overline{x} = 17.1, \quad s = 2.4, \quad n = 81.$$

8. Följande data kommer från normalfördelningar $X \sim N(\mu_x, \sigma^2)$, $Y \sim N(\mu_y, \sigma^2)$ där den gemensamma variansen σ^2 är okänd:

$$\begin{aligned} X:&\ 1.3,\ 1.7,\ 3.7,\\ Y:&\ 1.1,\ 0.4,\ 2.0,\ 1.0. \end{aligned}$$

Hitta ett 95% konfidensintervall för $\mu_x - \mu_y$, och ta ställning till $H_0 : \mu_x = \mu_y$ mot $H_A : \mu_x \neq \mu_y$ på signifikansnivå $\alpha = 0.05$.

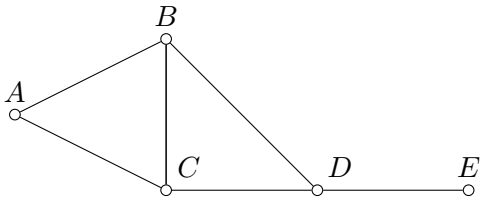
9. Med $n = 11$ datapunkter $(x_k, y_k), k = 1, \ldots, 11$ fås

$$\begin{aligned} \overline{x} &= 2.23, & \overline{y} &= 1.125,\\ S_{xx} &= 1.1, & S_{xy} &= 1.84, & S_{yy} &= 3.61. \end{aligned}$$

För en ny observation (x, Y), hitta ett övre begränsat 99% prediktionsintervall för Y.

Lösningar

1. Betrakta följande nätverk av vägar:



Varje enskild väg är öppen med sannolikhet p, oberoende av de andra. Vad är sannolikheten att det finns en öppen väg från A till E som går genom B?

Lösning. Notation: vi skriver $ABDE$ för händelsen att kanterna AB, BD, DE alla är öppna. Då kan vi få den sökta sannolikheten genom att summera över alla möjliga vägar.

$$\begin{aligned}\mathcal{A} &= \{\text{väg från } A \text{ till } E \text{ via } B \text{ öppen}\}\\ &= ABCDE \cup ACBDE \cup ABDE\\ &= (ABCD \cup ACBD \cup ABD) \cap DE.\end{aligned}$$

Låt $\mathcal{B}$ vara händelsen inom parentes; att det finns en väg från A till D via B. Då är $\mathcal{B}$ oberoende av DE, så

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{B})\mathbb{P}(DE) = p \times \mathbb{P}(\mathcal{B}).$$

För tre händelser A_1, A_2, A_3 har vi (rita Venndiagram!)

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3)\\ - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_2 \cap A_3) - \mathbb{P}(A_3 \cap A_1)\\ + \mathbb{P}(A_1 \cap A_2 \cap A_3).\end{aligned}$$

Vi får då

$$\begin{aligned}\mathbb{P}(\mathcal{B}) = \mathbb{P}(ABCD) + \mathbb{P}(ACBD) + \mathbb{P}(ABD)\\ - \mathbb{P}(ABCD \cap ACBD) - \mathbb{P}(ACBD \cap ABD) - \mathbb{P}(ABD \cap ABCD)\\ + \mathbb{P}(ABCD \cap ACBD \cap ABD).\end{aligned}$$

Varje sannolikhet här är p^k, där k är antalet kanter involverade. Till exempel är

$$\mathbb{P}(ABCD \cap ACBD) = \mathbb{P}(AB \cap BC \cap CD \cap AC \cap BD) = p^5.$$

Alltså är

$$\begin{aligned}\mathbb{P}(\mathcal{B}) &= p^3 + p^3 + p^2 - p^5 - p^4 - p^4 + p^5 \\ &= p^2 + 2p^3 - 2p^4,\end{aligned}$$

och den sökta sannolikheten är

$$\mathbb{P}(\mathcal{A}) = p \times \mathbb{P}(\mathcal{B}) = p^3 + 2p^4 - 2p^5.$$

2. En stokastisk variabel X har täthetsfunktion

$$f_X(x) = c \sin x, \quad 0 < x < \pi.$$

(a) Vad är c?

(b) Vad är $\mathrm{Var}\,(X)$?

Lösning.

(a) Täthetsfunktionen integral måste bli 1, så

$$1 = \int_0^\pi c \sin x dx = c\left[-\cos x\right]_{x=0}^{\pi} = 2c,$$

ger $c = 1/2$.

(b) Vi beräknar

$$\begin{aligned}\mathbb{E}\left[X\right] &= \int_0^\pi x \cdot \frac{1}{2} \sin x dx \\ &= \left[-\frac{x}{2}\cos x\right]_{x=0}^{\pi} + \frac{1}{2}\int_0^\pi \cos x dx \\ &= \frac{\pi}{2} + \frac{1}{2}\left[\sin x\right]_{x=0}^{\pi} = \frac{\pi}{2},\end{aligned}$$

och

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \int_0^\pi x^2 \cdot \frac{1}{2} \sin x dx \\ &= \left[-\frac{x^2}{2}\cos x\right]_{x=0}^{\pi} + \int_0^\pi x \cos x dx \\ &= \frac{\pi^2}{2} + \left[x \sin x\right]_{x=0}^{\pi} - \int_0^\pi \sin x dx \\ &= \frac{\pi^2}{2} - \left[-\cos x\right]_{x=0}^{\pi} \\ &= \frac{\pi^2}{2} - 2.\end{aligned}$$

Då är

$$\mathrm{Var}\,(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = \frac{\pi^2}{2} - 2 - \left(\frac{\pi}{2}\right)^2 = \frac{\pi^2}{4} - 2.$$

3. Låt $\{\tau_1, \tau_2, \dots\}$ vara en Poissonprocess med intensitet $\lambda > 0$. Vid varje impuls slås en sexsidig tärning (slaget tar ingen tid, det påbörjas och avslutas vid tidpunkt τ_k). Låt T vara tidpunkten för den första sexan som slås under processen. Vad har T för fördelning?

Lösning. Vår ofelbara intuition säger oss att de tidpunkter då vi slår en sexa ska vara en Poissonprocess med intensitet $\lambda/6$, så svaret borde vara $\exp(\lambda/6)$.

Vi räknar på det. För ett givet $t > 0$ är antalet impulser i Poissonprocessen $X(t) \sim \text{Poi}(\lambda t)$. Vi har

$$\begin{aligned}\mathbb{P}(T > t) &= \sum_{k=0}^{\infty} \mathbb{P}(T > t \mid X(t) = k)\mathbb{P}(X(t) = k) \\ &= \sum_{k=0}^{\infty} \left(1 - \frac{1}{6}\right)^k \frac{(\lambda t)^k e^{-\lambda t}}{k!}.\end{aligned}$$

Här använder vi det faktum att om $T > t$, innebär det att alla $X(t)$ tärningskast under $[0, t]$ resulterat i något annan än en sexa. Denna summan kan vi beräkna ut, det är Taylorserien för exponentialfunktionen:

$$\begin{aligned}\mathbb{P}(T > t) &= e^{-\lambda t} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\lambda t \frac{5}{6}\right)^k \\ &= \exp\left\{-\lambda t + \lambda t \frac{5}{6}\right\} \\ &= e^{-\lambda t/6}.\end{aligned}$$

Detta stämmer överens med $T \sim \exp(\lambda/6)$, så det är svaret.

4. Låt X, Y vara diskreta slumpvariabler med $Y \sim \text{Poi}(2)$ och (betingat) $X \sim \mathcal{U}\{0, 1, \dots, Y\}$, alltså

$$f_Y(y) = \frac{e^{-2}2^y}{y!}, \quad f_{X|Y}(x \mid y) = \frac{1}{y+1},$$

för heltal $0 \le x \le y$. Beräkna $\text{Cov}\,(X, Y)$.

Lösning. Från tabell eller beräkning har vi $\mathbb{E}[Y] = 2$. Sen har vi

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{y\geq 0} \mathbb{E}[X \mid Y = y] f_Y(y) \\
&= \sum_{y\geq 0} \left(\sum_{x=0}^{y} x f_{X|Y}(x \mid y) \right) f_Y(y) \\
&= \sum_{y\geq 0} \left(\sum_{x=0}^{y} \frac{x}{y+1} \right) f_Y(y) \\
&= \sum_{y\geq 0} \frac{y}{2} f_Y(y) = \frac{\mathbb{E}[Y]}{2} = 1.
\end{aligned}$$

Slutligen är

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{y\geq 0} \mathbb{E}[XY \mid Y = y] f_Y(y) \\
&= \sum_{y\geq 0} \mathbb{E}[Xy \mid Y = y] f_Y(y) \\
&= \sum_{y\geq 0} y\mathbb{E}[X \mid Y = y] f_Y(y) \\
&= \sum_{y\geq 0} \frac{y^2}{2} f_Y(y).
\end{aligned}$$

Ett knep här är att se att

$$\frac{1}{2} \sum_{y\geq 0} y^2 f_Y(y) = \frac{1}{2}\mathbb{E}\left[Y^2\right] = \frac{1}{2}(\mathrm{Var}(Y) + \mathbb{E}[Y]^2) = 3.$$

Detta kan vi slå upp i tabell. (Jag visade ett annat knep på föreläsningen). Vi får

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 3 - 1 \cdot 2 = 1.$$

5. Låt $X_n \sim \Gamma(n, 1)$. Hitta konstanter a, b sådana att

$$\lim_{n\to\infty} \mathbb{P}(an - b\sqrt{n} \leq X_n \leq an + b\sqrt{n}) = 0.99.$$

Lösning. Först noterar vi att $X_n = E_1 + E_2 + \cdots + E_n$, där $E_k \sim \exp(1)$ är oberoende. Detta gör att vi kan tillämpa stora talens lag och centrala gränsvärdessatsen. Notera att

$$\mu = \mathbb{E}[E_k] = 1, \quad \sigma^2 = \mathrm{Var}(E_k) = 1.$$

För det första noterar vi att om $a < 1$, och n är stort, så kan vi välja $\varepsilon > 0$ så att

$$\begin{aligned}&\mathbb{P}\left(a - \frac{b}{\sqrt{n}} \leq \frac{X_n}{n} \leq a + \frac{b}{\sqrt{n}}\right)\\ &\leq \mathbb{P}\left(\frac{X_n}{n} < 1 - \varepsilon\right).\end{aligned}$$

Stora Talens Lag säger att denna sannolikhet går mot 0, och på liknande sätt går sannolikheten mot 0 om $a > 1$. Slutsatsen är att vi måste ha $a = 1$.

Nu väljer vi b med hjälp av centrala gränsvärdessatsen. Denna säger att

$$\lim_{n\to\infty} \mathbb{P}\left(-b \leq \frac{X_n - n}{\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(-b).$$

Vi ska då ha

$$0.99 = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1,$$

vilket ger $b = \Phi^{-1}(0.995) \approx 2.58$.

6. Anta att en fördelning $\mathcal{F}(\theta)$ har parametriserad täthetsfunktion

$$f_\theta(x) = 2(x - \theta), \quad \theta < x < \theta + 1.$$

Hitta en ML-skattning för θ baserad på datapunkter $x_1 = 0.4, x_2 = 0.7, x_3 = 0.5$.

Lösning. Här blir det viktigt att notera att $f_\theta(x) = 0$ utanför det definierade intervallet. Likelihoodfunktionen

$$L(\theta) = f_\theta(0.4) f_\theta(0.7) f_\theta(0.5)$$

är därför endast nollskild på intervallet $\theta \in [-0.3, 0.4]$. För dessa θ är

$$L(\theta) = 2^3(0.4 - \theta)(0.7 - \theta)(0.5 - \theta).$$

För att maximera detta är faktorn $2^3 = 8$ irrelevant, så den kan vi slänga bort. Derivatan av $L(\theta)/8$ blir ett andragradspolynom:

$$L'(\theta) = 8(-3\theta^2 + 3.2\theta - 0.83).$$

Detta är negativt för alla $\theta \in [-0.3, 0.4]$, alltså maximeras det för $\hat{\theta} = -0.3$, som blir vår ML-skattning.

7. Hitta ett 99% konfidensintervall för μ, med normalfördelad data med okänt σ^2 och stickprovsdata

$$\overline{x} = 17.1, \quad s = 2.4, \quad n = 81.$$

Lösning. Konfidensintervallet baseras på att

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

och formeln är

$$\mu \in \overline{X} \pm F^{-1}_{t_{n-1}}\left(1 - \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}} \quad (1-\alpha).$$

Ett 99% konfidensintervall motsvarar $\alpha = 0.01$.

Vi har

$$\mathbb{P}(-F^{-1}_{t_{n-1}}(0.995) \leq T \leq F^{-1}_{t_{n-1}}(0.995)) = 0.99.$$

Med $F^{-1}_{t_{80}}(0.995) = 2.64$ får vi då

$$\mu \in 17.1 \pm 2.64 \cdot \frac{2.4}{\sqrt{81}} = (16.4, 17.8).$$

Kommentar. Det kan vara intressant att se hur nära kvantilen är den för normalfördelningen:

$$F^{-1}_{t_{80}}(0.995) = 2.64 \quad \text{kontra} \quad \Phi^{-1}(0.995) = 2.57.$$

Detta är ganska nära, men för ett så stort värde som 0.995 tar det ganska lång tid innan vi konvergerar ordentligt. Inte ens med 1000 frihetsgrader är vi nere på 2.57, utan $F^{-1}_{t_{1000}}(0.995) = 2.58$.

8. Följande data kommer från normalfördelningar $X \sim N(\mu_x, \sigma^2)$, $Y \sim N(\mu_y, \sigma^2)$ där den gemensamma variansen σ^2 är okänd:

$$\begin{aligned} X&: \ 1.3, \ 1.7, \ 3.7, \\ Y&: \ 1.1, \ 0.4, \ 2.0, \ 1.0. \end{aligned}$$

Hitta ett 95% konfidensintervall för $\mu_x - \mu_y$, och ta ställning till $H_0 : \mu_x = \mu_y$ mot $H_A : \mu_x \neq \mu_y$ på signifikansnivå $\alpha = 0.05$.

Lösning. Vi beräknar

$$\begin{aligned} \overline{x} &= 2.233, & \overline{y} &= 1.125, \\ s_x^2 &= 1.65, & s_y^2 &= 0.44. \end{aligned}$$

Stickprovsstorlekarna är $n = 3$ respektive $m = 4$. Den poolade stickprovsvaraiansen ges av

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2} = \frac{2 \cdot 1.65 + 3 \cdot 0.44}{3+4-2} = 0.92.$$

Rimlighetskoll. Vi har $\min\{s_x^2, s_y^2\} \leq s_p^2 \leq \max\{s_x^2, s_y^2\}$. Bra! Om man räknar fel kan man lätt få något utanför detta intervallet, och då vet man att man gjort fel.

Konfidensintervallet och hypotestestet baseras båda på att

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{m+n-2}.$$

Den aktuella kvantilen är $F_{t_5}^{-1}(0.975) = 2.57$. Ett konfidensintervall får vi som

$$\begin{aligned}
\mu_x - \mu_y &\in \overline{x} - \overline{y} \pm F_{t_5}^{-1}(0.975) s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \quad (95\%) \\
&= 2.233 - 1.125 \pm 2.57 \cdot \sqrt{0.92}\sqrt{\frac{1}{3} + \frac{1}{4}} \\
&= (-0.77,\ 2.99).
\end{aligned}$$

Vi kan nu direkt ta ställning till hypotestestet; eftersom konfidensintervallet innehåller 0, så tvingas vi behålla $H_0 : \mu_x - \mu_y = 0$ på signifikansnivå $\alpha = 0.05$.

9. Med $n = 11$ datapunkter $(x_k, y_k), k = 1, \ldots, 11$ fås

$$\begin{aligned}
\overline{x} &= 2.23, & \overline{y} &= 1.125, & \\
S_{xx} &= 1.1, & S_{xy} &= 1.84, & S_{yy} &= 3.61.
\end{aligned}$$

För en ny observation (x, Y), hitta ett övre begränsat 99% prediktionsintervall för Y.

Lösning. För vår linjära regressionsmodell $Y = a + bX + \varepsilon$, med $\varepsilon \sim N(0, \sigma^2)$, tar vi fram skattningar

$$\begin{aligned}
\hat{b} &= \frac{S_{xy}}{S_{xx}} = 1.67, \\
\hat{a} &= \overline{y} - \hat{b}\overline{x} = 1.48, \\
s^2 &= \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 0.06.
\end{aligned}$$

Det ensidiga prediktionsintervallet ges av (formel)

$$\begin{aligned} Y &\leq \hat{a} + \hat{b}x + F_{t_9}^{-1}(0.99)s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}} \\ &= 1.48 + 1.67x + 2.82\sqrt{0.06}\sqrt{1 + \frac{1}{11} + \frac{(x-2.23)^2}{1.1}}. \end{aligned}$$

Det är knappt värt att förenkla detta, men det kan ju vara trevligt:

$$Y \leq 1.48 + 1.67x + \sqrt{0.43x^2 - 0.87x + 2.68}.$$