UPPSALA UNIVERSITET
Matematiska institutionen
Rolf Larsson

TENTAMEN I MATEMATISK STATISTIK
Inferens 1 1MS035
2016–12–16
Lösningar

1. Slumpvariabeln X kan anta värdena $0, 1, 2, 3$ med sannolikheterna $1 - 6\theta$, θ, 2θ och 3θ. Man har observationerna $3, 0, 1, 2, 3, 2, 2, 3$. (5p)

   (a) Beräkna momentskattningen av θ. (1p)

   Lösning: Väntevärdet är

   $$m(\theta) = E(X) = 0 * (1 - 6\theta) + 1 * \theta + 2 * (2\theta) + 3 * (3\theta) = 14\theta.$$

   Det observerade medelvärdet $\bar{x} = 2$ ger att vi ska lösa $14\theta = 2$, så vi får momentskattningen $\theta^* = 1/7 \approx 0.143$.

   (b) Beräkna minsta kvadratskattningen av θ. (2p)

   Lösning: Svaret blir som i (a) eftersom θ är endimensionell, se boken s.286. Vi kan även få detta genom att minimera

   $$Q(\theta) = \sum_{i=1}^{n} (x_i - 14\theta)^2,$$

   där derivatorna är

   $$\begin{aligned} Q'(\theta) &= -28 \sum_{i=1}^{n} (x_i - 14\theta) = -28n(\bar{x} - 14\theta), \\ Q''(\theta) &= 28 * 14n > 0, \end{aligned}$$

   vilket ger ett minimum i $\theta = \bar{x}/14 = 1/7$.

   (c) Beräkna maximum likelihoodskattningen av θ. (2p)

   Lösning: Vi observerar en nolla, en etta tre tvåor och tre treor, så likelihooden är

   $$L(\theta) = (1 - 6\theta)^1 * \theta^1 * (2\theta)^3 * (3\theta)^3 = C_1(1 - 6\theta)\theta^7,$$

   där C_1 är en konstant. Detta ger log likelihooden (för en annan konstant C_2)

   $$l(\theta) = \ln L(\theta) = C_2 + \ln(1 - 6\theta) + 7 \ln \theta,$$

   med derivator

   $$\begin{aligned} l'(\theta) &= -\frac{6}{1 - 6\theta} + \frac{7}{\theta}, \\ l''(\theta) &= -\frac{36}{(1 - 6\theta)^2} - \frac{7}{\theta^2}. \end{aligned}$$

   Lösningen till $'(\theta) = 0$ är $\theta = 7/48 \approx 0.146$. Detta är ML-skattningen eftersom vi alltid har att $l''(\theta) < 0$, som ger maximum.

2. Man har ett slumpmässigt stickprov $x_1, x_2, ..., x_4$ från en slumpvariabel X med väntevärde μ och varians σ^2, samt ett slumpmässigt stickprov $y_1, y_2, ..., y_8$ från en slumpvariabel Y med väntevärde 2μ och varians $4\sigma^2$. Vi antar vidare att X och Y är oberoende. (5p)

Man vill skatta μ, och två skattningar är föreslagna:

$$\mu_1^* = \frac{2\bar{x} + \bar{y}}{4}, \quad \mu_2^* = \frac{\bar{x} + \bar{y}}{3}.$$

(a) Visa att båda skattningarna är väntevärdesriktiga. (1p)

Lösning: Motsvarande estimatorer är

$$\mu_1^* = \frac{1}{2}\overline{X} + \frac{1}{4}\overline{Y}, \quad \mu_2^* = \frac{1}{3}\overline{X} + \frac{1}{3}\overline{Y}.$$

Av räknelagarna för väntevärdet följer att

$$E(\mu_1^*) = \frac{1}{2}E(\overline{X}) + \frac{1}{4}E(\overline{Y}) = \frac{1}{2}\mu + \frac{1}{4}2\mu = \mu,$$
$$E(\mu_2^*) = \frac{1}{3}E(\overline{X}) + \frac{1}{3}E(\overline{Y}) = \frac{1}{3}\mu + \frac{1}{3}2\mu = \mu,$$

så båda är väntevärdesriktiga.

(b) Vilken skattning är effektivast? (2p)

Lösning: Räknelagarna för varians ger

$$V(\mu_1^*) = \left(\frac{1}{2}\right)^2 V(\overline{X}) + \left(\frac{1}{4}\right)^2 V(\overline{Y}) = \frac{1}{4}\frac{\sigma^2}{4} + \frac{1}{16}\frac{4\sigma^2}{8} = \frac{3}{32}\sigma^2,$$
$$V(\mu_2^*) = \left(\frac{1}{3}\right)^2 V(\overline{X}) + \left(\frac{1}{3}\right)^2 V(\overline{Y}) = \frac{1}{9}\frac{\sigma^2}{4} + \frac{1}{9}\frac{4\sigma^2}{8} = \frac{3}{36}\sigma^2,$$

så μ_2^* har minst varians, och är därför effektivast.

(c) Finns det någon effektivare skattning på formen $a\bar{x} + b\bar{y}$? Motivera ditt svar! (2p)

Lösning: Med räkningar som ovan fås

$$E(a\overline{X} + b\overline{Y}) = a\mu + b * 2\mu,$$

som är lika med μ om $a + 2b = 1$, d.v.s. om $a = 1 - 2b$. Vi har vidare att

$$V(a\overline{X} + b\overline{Y}) = a^2\frac{\sigma^2}{4} + b^2\frac{4\sigma^2}{8} = (1-2b)^2\frac{\sigma^2}{4} + b^2\frac{\sigma^2}{2} = g(b)\sigma^2,$$

där

$$g(b) = \frac{1}{4}\left\{(1-2b)^2 + 2b^2\right\} = \frac{1}{4}(1 - 4b + 6b^2).$$

Vi söker minimum för $g(b)$. Derivering ger

$$g'(b) = \frac{1}{4}(-4 + 12b)$$

och $g''(b) = 3$. Vi ser att $g'(b) = 0$ för $b = 1/3$ och att detta ger ett minimum eftersom $g''(b) > 0$. Detta ger oss att $a = 1/3$, och den optimala skattningen är alltså μ_2^*.

3. Som en följd av den s.k. ”Icings lag” kan tiden i dygn (måste ej vara ett heltal) efter den sista november fram till dess att isen lägger sig på sjön Fyrhörningen med god approximation betraktas som en slumpvariabel X med fördelningsfunktion

$$F(x) = 1 - \exp\left(-\frac{\theta x^2}{10000}\right),$$

där $x > 0$, och annars är $F(x) = 0$. Parametern θ är okänd. (5p)

(a) Långfärdsskridskoåkaren Låffe tror att $\theta = 9$. En vinter lägger sig isen 60 dagar efter den sista november. Hjälp Låffe att avgöra om det är rimligt att $\theta = 9$ genom att testa en lämplig hypotes. Använd signifikansnivån 5%. (2p)

Lösning: Frågan rymmer ingen misstanke om att θ skulle vara större eller mindre än 9. Därför är det naturligast med ett tvåsidigt test, alltså ett test av H_0: $\theta = 9$ mot alternativet H_1: $\theta \neq 9$. Under H_0 är halva P-värdet lika med

$$P(X \geq 60) = 1 - F(60) = \exp\left(-\frac{9 * 60^2}{10000}\right) \approx 0.0392.$$

Därmed är P-värdet $2 * 0.0392 \approx 0.078 > 0.05$, så vi kan inte förkasta H_0.

(b) Vilken styrka har testet i (a) för $\theta = 2$? (3p)

Lösning: Det kritiska området är på formen $C = \{x < K_1\} \cup \{x > K_2\}$ där $P(X < K_1) = 0.025 = P(X > K_2)$ uträknat under H_0. För att bestämma K_2 ska vi lösa

$$0.025 = P(X > K_2) = \exp\left(-\frac{9K_2^2}{10000}\right),$$

vilket ger $K_2 = 100\sqrt{-\ln(0.025)/9} = 64.0$. På samma sätt är

$$0.025 = P(X < K_1) = 1 - \exp\left(-\frac{9K_1^2}{10000}\right),$$

så att $K_2 = 100\sqrt{-\ln(0.975)/9} = 5.3$.
Den sökta styrkan ges nu av sannolikheten för att $X \in C$ givet att $\theta = 2$. Detta blir

$$1 - \exp\left(-\frac{2 * 5.3^2}{10000}\right) + \exp\left(-\frac{2 * 64.0^2}{10000}\right) \approx 0.45.$$

4. Butikskedjan Interspurt beställer slalomhjälmar från två olika leverantörer, A och B. Från leverantör A beställer man 200 hjälmar, av vilka 8 visar sig vara defekta. Antalet beställda hjälmar från leverantör B var 400, och av dessa var 10 defekta.(5p)

   (a) Beräkna ett 95%-igt konfidensintervall för skillnaden mellan andelarna defekta hjälmar för de två leverantörerna. (4p)

   Lösning: Vi antar att antalet defekta hjälmar från leverantör A beskrivs av $X \sim \text{Bin}(200, p_1)$ och att antalet defekta hjälmar från leverantör B beskrivs av $Y \sim \text{Bin}(400, p_2)$. Vi vill beräkna ett konfidensintervall för $p_1 - p_2$.

   Vi har skattningarna $p_1^* = 8/200 = 0.04$ och $p_2^* = 10/400 = 0.025$. Tumregeln för normalapproximation är uppfylld, ty $200p_1^*(1-p_1^*) \approx 7.7 > 5$ och $400p_2^*(1-p_2^*) \approx 9.8 > 5$. Ett konfidensintervall med approximativ konfidensgrad 95% blir

$$
\begin{aligned}
I_{p_1-p_2} &= p_1^* - p_2^* \pm \lambda_{0.025}\sqrt{\frac{p_1^*(1-p_1^*)}{200} + \frac{p_2^*(1-p_2^*)}{400}} \\
&= 0.04 - 0.025 \pm 1.96 * \sqrt{\frac{0.04 * 0.96}{200} + \frac{0.025 * 0.975}{400}} \\
&= 0.015 \pm 0.031 = (-0.016,\ 0.046).
\end{aligned}
$$

   (b) Finns det belägg för att säga att andelarna defekta hjälmar är olika för de två leverantörerna? (1p)

   Lösning: Nej, inte på risknivån 5%, ty $p_1 - p_2 = 0$ ligger i det 95%-iga konfidensintervallet.

5. Jultomten vill välja ut renar till sin släde. Han har gett i uppdrag till tomtenissen Firpo och tomtenissan Firpelina att snabbhetsträna var sin kull med renar. Vid träningsperiodens slut väljer tomten slumpmässigt ut fem av Firpos och fem av Firpelinas renar och mäter upp deras toppfart i km/h. Resultaten ges i följande tabell.

Firpos kull	72.7	71.2	75.4	73.0	70.7
Firpelinas kull	77.2	73.4	75.3	78.0	79.1

Avgör om renar från olika kullar är lika eller olika snabba genom att genomföra ett lämpligt hypotestest. Var noga med att ange vilka förutsättningar du gör. (5p)

Lösning: Det finns ingen naturlig hopparning av renarna (de olika kullarna har ju olika renar), så vi använder metoden "stickprov i par". För Firpo observerar vi hastigheterna $x_1, ..., x_5$ som är ett slumpmässigt stickprov från $X \sim N(\mu_1, \sigma_1^2)$. För Firpelina observerar vi hastigheterna $y_1, ..., y_5$ som är ett slumpmässigt stickprov från $Y \sim N(\mu_2, \sigma_2^2)$. Vi vill testa H_0: $\mu_1 = \mu_2$ mot H_1: $\mu_1 \neq \mu_2$. (Det finns inget i uppgiftens formulering som antyder misstanken att någon kull skulle vara snabbare än den andra.)

Vi observerar medelvärdena $\bar{x} = 72.6$ och $\bar{y} = 76.6$, samt standardavvikelserna $s_x = 1.8426$ och $s_y = 2.2638$. Varianserna σ_1^2 och σ_2^2 är okända men antas vara lika. Vi får den poolade variansskattningen

$$
s_p^2 = \frac{4 * 1.8426^2 + 4 * 2.2638^2}{8} = 4.2600,
$$

och observerad teststorhet

$$T_{obs} = \frac{72.6 - 76.6}{\sqrt{4.26 * (1/5 + 1/5)}} = -3.06.$$

Eftersom $|T_{obs}| = 3.06 > 2.3060 = t_{0.025}(8)$, så förkastar vi nollhypotesen att de springer lika fort på nivån 5%.

6. Läkaren Elina vill undersöka den febernedsättande effekten av tabletten Cepofan för influensapatienter. Hon har en grupp av fyra patienter. På kvällen mäter hon deras kroppstemperatur (i grader Celcius), och ger dem sedan en tablett Cepofan. På morgonen mäter hon åter deras kroppstemperatur. Värdena anges i följande tabell. (5p)

Patient nr	1	2	3	4
Temp. på kvällen	40.5	37.8	39.2	38.4
Temp. på morgonen	38.2	37.4	38.4	36.7

(a) Beräkna ett 90%-igt konfidensintervall för den febernedsättande effekten hos en tablett Cepofan i grader. Var noga med att ange vilka förutsättningar du gör. (3p)

Lösning: Här mäter man på samma patient två gånger, så det är naturligt att se detta som stickprov i par. Vi antar att vi har observationer (x_i, y_i) från slumpvariabler (X_i, Y_i), $i = 1, 2, 3, 4$, där alla X_i och Y_i antas vara oberoende med $X_i \sim N(\mu_i, \sigma_1^2)$ och $Y_i \sim N(\mu_i + \Delta, \sigma_2^2)$.
Detta ger att $Z_i = Y_i - X_i \sim N(\Delta, \sigma^2$, där $\sigma^2 = \sigma_1^2 + \sigma_2^2$ är okänd. Vi söker ett 90%-igt konfidensintervall för Δ.
Vi observerar skillnaderna $-2.3, -0.4, -0.8, -1.7$, med medelvärde $\bar{z} = -1.3$ och standardavvikelse $s_z = 0.8602$. Det sökta konfidensintervallet blir

$$I_\Delta = \bar{z} \pm t_{0.05}(3)\frac{s_z}{\sqrt{4}} = 1.3 \pm 2.3534 * \frac{0.8602}{2} = 1.3 \pm 1.01 = (0.29,\ 2.31).$$

(b) Elinas försök får ses som en förstudie till ett mycket större experiment. Ungefär hur många patienter skulle man behöva i detta experiment för att få ett 90%-igt konfidensintervall av längd högst 0.2 grader? (2p)

Lösning: Vi antar att standardavvikelsen för det större experimentet är 0.8602 även den. Om antalet patienter i det större experimentet, n säg, är mycket stort bör vi kunna approximera $t_{0.05}(n-1)$ med $\lambda_{0.05} = 1.6449$. Att konfidensintervallets längd är högst 0.2 är då samma sak som att

$$2 * 1.6449 * \frac{0.8602}{\sqrt{n}} \leq 0.2.$$

Detta är ekvivalent med $n \geq (1.6449 * 0.8602/0.1)^2 = 200.2$. Minst cirka 201 patienter behövs alltså.

7. Antalet julklappar som ett barn i Norbyskolan får kan antas följa en Poissonfördelning med väntevärde μ, lika från år till år. Johannes går i en klass i Norbyskolan med 30 barn. Förra året fick dessa barn sammanlagt 282 julklappar. (5p)

   (a) Beräkna ett 99%-igt konfidensintervall för μ. (2p)

   Lösning: Vi har ett slumpmässigt stickprov $x_1, ..., x_{30}$ från $X \sim \text{Po}(\mu)$, där X beskriver hur många julklappar ett Norbybarn får en viss jul. Vi skattar μ med $\mu^* = \bar{x} = 282/30 = 9.4$, och motsvarande estimator är $\overline{X}$. Enligt tumregel kan vi approximera med normalfördelning om $n\mu^* > 15$, och här har vi $n\mu^* = 30 * 9.4 = 282 > 15$. Detta ger oss referensvariabeln

$$\frac{\overline{X} - \mu}{d} \approx N(0,1),$$

   där d är medelfelet. Eftersom $V(\overline{X}) = V(X)/30 = \mu/30$ har vi i vårt fall $d = \sqrt{\mu^*/30} = \sqrt{9.4/30} \approx 0.5598$. Detta ger oss ett konfidensintervall med approximativ konfidensgrad 99% enligt

$$I_\mu = \mu^* \pm \lambda_{0.005} d = 9.4 \pm 2.5758\sqrt{\frac{9.4}{30}} = 9.4 \pm 1.44 = (7.96,\ 10.84).$$

   (b) Det finns bara ett sätt för ett Norbybarn att bli på dåligt julhumör, och det är att det får färre än tre julklappar. Beräkna ett 99%-igt konfidensintervall för sannolikheten att Johannes blir på dåligt julhumör i år. (3p)

   Lösning: Vi söker ett konfidensintervall för

$$g(\mu) = P(X < 3) = e^{-\mu}\left(1 + \mu + \frac{\mu^2}{2}\right).$$

   Detta är en monotont avtagande funktion i μ eftersom

$$g'(\mu) = e^{-\mu}\left\{-\left(1 + \mu + \frac{\mu^2}{2}\right) + 1 + \mu\right\} = -\frac{\mu^2}{2}e^{-\mu}$$

   som är < 0 för alla μ.

   Intervallet i (a) ger att $7.96 \leq \mu \leq 10.84$, och det kan nu omformas som $g(10.84) \leq g(\mu) \leq g(7.96)$, vilket ger vårt sökta konfidensintervall som

$$\{g(10.84),\ g(7.96)\} = (0.0014,\ 0.0142).$$

8. Betrakta regressionsmodellen

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

där $i = 1, 2, ..., n$ och alla ε_i är oberoende $N(0, \sigma^2)$.

Data har simulerats fyra gånger från denna modell, varje gång med olika värden på parametrarna α, β och σ^2. Antalet observationer har valts till $n = 50$ och $x_i = i$ för $i = 1, 2, ..., 50$. Efter att simuleringarna har gjorts plottades data, parametrarna skattades och förklaringsgraden räknades ut. Plottar av data visas i figur 1-4 nedan.

Para ihop figur 1-4 med de skattade modellerna och motsvarande förklaringsgrader i (a)-(d) nedan. Motivera din lösning. (5p)

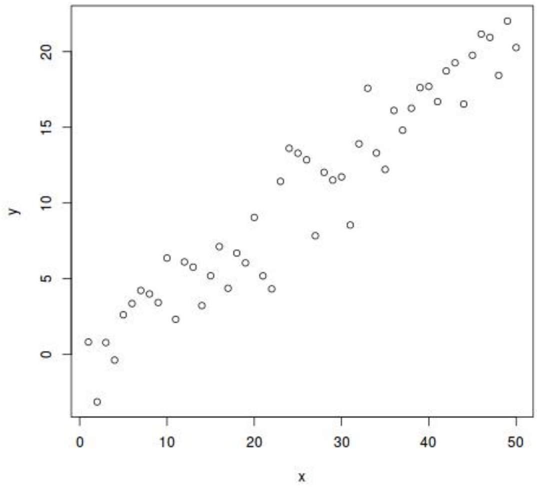
$$\begin{array}{lll}
y_i^* = 20.75 - 0.48x_i, & R^2 = 0.3\% & (a) \\
y_i^* = -7.61 + 0.61x_i, & R^2 = 17\% & (b) \\
y_i^* = -0.38 + 0.52x_i, & R^2 = 98\% & (c) \\
y_i^* = 4.15 + 1.74x_i, & R^2 = 53\% & (d)
\end{array}$$

Lösning: Vi tittar först på modell (c). Detta är den modell som har högst förklaringsgrad, och alltså minst spridning av observationspunkterna kring en tänkt anpassad regressionslinje. Detta ser ut att motsvara figur 1, där man också ser att lutningen 0.52 verkar rimlig.

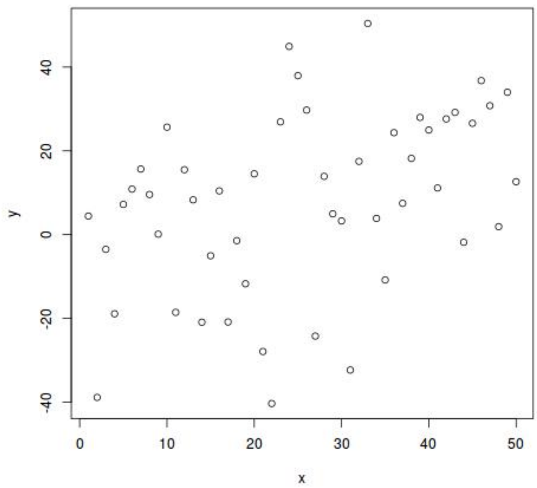
Av de övriga figurerna ser vi att vi har både högst lutning och förklaringsgrad i figur 3 (observera skalan på y-axeln), vilken då bör svara mot modell (d).

Återstår så modellerna (a) och (b) samt figurerna 2 och 4. Genom att observera skalan på y-axeln kan man inse att modellen med lägst förklaringsgrad, alltså (a), bör svara mot figur 4, där man ju har störst spridning på y-värdena. Dessutom kan möjligen ett svagt negativt samband (svagt negativ lutning på regressionslinjen) skönjas där, till skillnad mot i figur 2, där sambandet ser ut att vara svagt positivt.

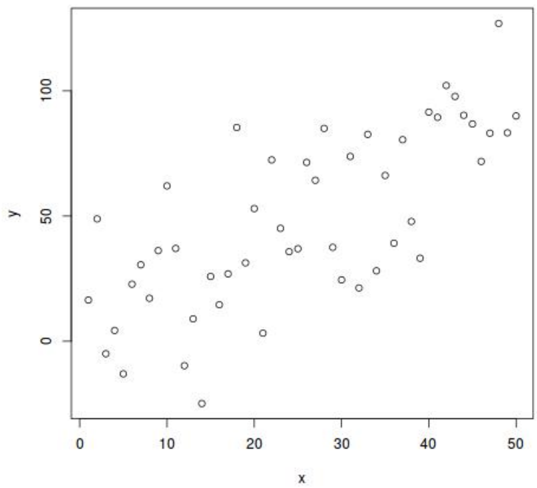
Svaret är alltså: 1-c, 2-b, 3-d, 4-a.



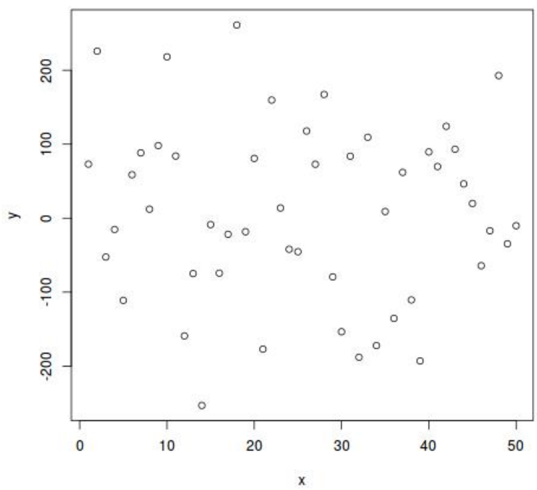
Figur 1: Plot för uppgift 8.



Figur 2: Plot för uppgift 8.



Figur 3: Plot för uppgift 8.



Figur 4: Plot för uppgift 8.