STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
LÖSNINGAR
8 januari 2025

Lösningar

Tentamen i Statistisk analys, 8 januari 2025

Uppgift 1

a) Sant

b) Sant

c) Sant

d) Falskt

e) Falskt

Uppgift 2

a) Vi antar att vikten på de enskilda abborrarna är någorlunda normalfördelade. Vår punktskattning av μ blir $\hat{\mu} = 6.425$ hg. Standardavvikelsen för abborrarnas vikter blir $s = 1.95$ hg. Vi har 8 observationer vilket ger 7 frihetsgrader. Ett 99% konfidensintervall för μ ges således av

$$\bar{x} \pm t_{0.005}(7)\frac{s}{\sqrt{8}} = 6.425 \pm 3.50\frac{1.95}{\sqrt{8}} = 6.42 \pm 2.42 = [4.0,\ 8.84].$$

b) För att testa H_0 beräknar vi $t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{8}} = \frac{1.425}{1.95/\sqrt{8}} = 2.07$. För ett tvåsidigt test på 5%-nivån ska vi förkasta H_0 om $|t_{obs}| > t_{0.025}(7) = 2.36$ vilket inte gäller. Vi kan således inte förkasta hypotesen att sjöns abborrar i genomsnitt väger 5.0 hg.

Uppgift 3

a) En rimlig modell är att anta enkel linjär regression. Dvs att olika individers längder och vikter är oberoende och att vikten (Y) beror av längden linjärt plus lite variation. Mer

precist antar vi att $Y_i = \alpha + \beta x_i + \epsilon_i$, där $\epsilon_i \sim N(0, \sigma^2)$, oberoende for olika i. Vi får följande skattningar:

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n(\bar{x})^2} = \frac{143852 - 10 * 182.5 * 78.7}{333319 - 10 * 182.5^2} = \frac{224.5}{256.5} = 0.875$$
$$\alpha^* = \bar{y} - \beta^* \bar{x} = 78.7 - 0.875 * 182.5 = 81.0$$
$$s^2 = \frac{SSE}{n-2} = \frac{S_{yy} - S_{xy}^2/S_{xx}}{8} = \frac{62245 - 10 * 78.7^2 - 224.5^2/256.5}{8} = 111.6/8 = 13.95$$

Skattningen av standardavvikelsen blir således $\sigma^* = s = 3.73$.

b) Att testa om vikten är oberoende av längd svarar mott $H_0 : \beta = 0$. Ett 95% konfidensintervall för β ges av

$$\hat{\beta} \pm t_{0.025}(n-2)s/\sqrt{S_{xx}} = 0.875 \pm 2.31 * 3.73/\sqrt{256.5} = 0.875 \pm 0.538.$$

Eftersom intervallet inte inkluderar nollhypotesens värde ($\beta = 0$) så förkastas hypotesen. Längden har således en signifikant påverkan på mäns vikt (vilket ju inte är särskilt förvånande).

Uppgift 4

Detta är parvisa observationer (x_i, y_i) och fördelningarna verkar snälla utan några outliers (rimligt normalfördelade). Vi studerar därför i stället förändringen för respektive individ. Motsvarande differenser $d_i = y_i - x_i$ blir: 0.10, -0.05, 0.15, 0.15, -0.10, 0.0, 0.05, dvs ett stickprov av storlek 7. Vi är intresserade av om dessa differenser tenderar att vara positiva eller inte.

a) Vi beräknar $\sum_i d_i = 0.30$ och $\sum_i d_i^2 = 0.070$. Om vi låter $D := Y - X$ så skattas den förväntade förbättrade effekten mellan vår och höst, $\mu = E(D) = E(Y) - E(X)$, med $\hat{\mu} = \bar{d} = 0.043$. Skattningen av hur mycket bättre det går för gymnasister i årskurs 3 på våren jämfört med hösten är att de höjer sina resultat med i genomsnitt 0.043.

b) Standardavvikelsen σ för olika individers ändrade resultat skattas med

$$s_d = \sqrt{\frac{\sum_i d_i^2 - n\bar{d}^2}{n-1}} = \sqrt{(\frac{0.07 - 7 * 0.043^2}{6}} = 0.098.$$

För att testa hypotesen $H_0 : \mu = 0$ mot $H_1 : \mu > 0$ så förkastar vi H_0 om $t_{obs} > t_{0.05}(6) = 1.943$ där $t_{obs} = (\bar{d} - 0)/(s_d/\sqrt{n})$. I vårt fall får vi $t_{obs} = 0.043/(0.098/\sqrt{7}) = 1.16$. Vi är således långt ifrån att förkasta H_0. Slutsatsen är att vi med detta lilla datamaterial inte kan utesluta möjligheten att elever *inte* förbättrar sina resultat (i genomsnitt) mellan höst och vår i årskurs 3.

Uppgift 5

a) Eftersom $Y = X_1 + \cdots + X_k$ där X-variablerna är oberoende $Exp(\beta)$ så har vi $E(Y) = kE(X) = k/\beta$. Om vi ska skatta β från vårt stickprov $y_1, \ldots y_n$ med moment metoden sätter vi alltså $\bar{y} = E(Y) = k/\beta$ och löser ut β. Momentskattningen blir således $\beta^* = k/\bar{y} = nk/\sum_i y_i$.

b) Likelihooden ges av

$$L(\beta) = \prod_{i=1}^{n} f(y_i) = \frac{\beta^{nk}}{(k!)^n} e^{-\beta \sum_i y_i} \prod_{i=1}^{n} y_i^{k-1}.$$

Om vi hoppar över termer som inte beror på β (vi vill ju maximera med avseende på β) så blir log-likelihooden $\ell(\beta) = nk \log(\beta) - \beta \sum_i y_i + const.$ Vi maximerar genom att derivera och sätte derivatan till 0 vilket ger följande ekvation: $(nk/\beta) - \sum_i y_i = 0$. Skattningen blir således $\hat{\beta} = nk/\sum_i y_i$, dvs samma skattning som momentskattningen.

c) Vi har $k = 3$, $n = 4$ och $\sum_i y_i = 31.3$ så den gemensamma skattningen blir $\hat{\beta} = nk/\sum_i y_i = 0.38$.

d) Vi har att $\hat{\beta} = k/\bar{y} = nk/\sum_i y_i$. Men varje y_i kan skrivas som summan av k oberoende $Exp(\beta)$ variabler: $y_i = x_{i1} + \cdots + x_{ik}$. Så $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}$, en summa av nk oberoende $Exp(\beta)$. Vi får således att $\hat{\beta} = nk/\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} = 1/\bar{x}$.

Om vi hade observerat alla dessa x_{ij}-variabler (nu observerar vi ju bara $y_i = \sum_{j=1}^{k} x_{ij}$) så skulle vi skatta β med $1/\bar{x}$ (detta är både moment- och ML-skattning för $X \sim Exp(\beta)$). Slutsatsen är därför att ML-skattning av β för $\Gamma(k, \beta)$ om k är känd, är densamma som för fallet att vi hade observerat de enskilda x_{ij}-värdena med exponentialfördelningen.

Uppgift 6

Vi noterar att fördelningen tycks tjocksvansad med merparten värden mellan 3 och 15, samt en observation på 20.4 och en på 42.9. Det är således rimligt att fokusera på median som lämpligt lägesmått snarare än väntevärde.

a) För att skatta medianen sorterar vi först observationer så att vi får det ordnade stickprovet: $x_{(1)} = 3.1$, $x_{(2)} = 4.5$, $x_{(3)} = 7.9$, $x_{(4)} = 8.5$, $x_{(5)} = 9.1$, $x_{(6)} = 10.0$, $x_{(7)} = 12.1$, $x_{(8)} = 14.2$, $x_{(9)} = 20.4$, $x_{(10)} = 42.9$.

Eftersom vi har jämnt antal observationer ($n = 10$) blir stickprovets median $(x_{(5)} + x_{(6)})/2 = 9.55$ vilket också blir vår skattningen av medianen för den bakomliggande medianen: $x^*_{0.5} = 9.55$.

b) Möjliga icke-parametriska konfidensintervall för $x_{0.5}$ ges av $[x_{(1)},\ x_{(10)}]$, eller $[x_{(2)},\ x_{(9)}]$ eller $[x_{(3)},\ x_{(8)}]$, ...osv. Det största har högst konfidensgrad, det andra näst högst, osv. Konfidensgraden av det första intervallet blir $P(X_{(1)} \leq x_{0.5} \leq X_{(10)} = 1 - P(x_{0.5} < X_{(1)}) -$

$P(X_{(10)} < x_{0.5})$. De två negativa termerna är lika av symmetriskäl. Men $x_{0.5} < X_{(1)}$ betyder att stickprovets minsta värde ska vara större än den teoretiska medianen. Detta är samma sak som att *alla* $n = 10$ observationer är större än medianen. Varje enskild observation är större än medianen med sannolikhet 0.5 (per definition), så sannolikheten att alla observationer är större än medianen är således 0.5^{10}. Den efterfrågade sannolikheten, dvs $P(X_{(1)} \leq x_{0.5}\ X_{(10)}$, blir således $1 - 0.5^{10} - 0.5^{10} = 0.998$.

För att beräkna konfidensgraden för övriga intervall kan det vara klokt att definiera $Y :=$ antalet av de 10 observationerna som är större än medianen. Eftersom varje observation är större än medianen med sannolikhet 0.5 som blir $Y \sim Bin(n = 10, p = 0.5)$. Från detta får vi (analogt som för fallet ovan)

$$\begin{aligned} P(X_{(2)} \leq x_{0.5} \leq X_{(9)} &= 1 - P(x_{0.5} < X_{(2)}) - P(X_{(9)} > x_{0.5}) = 1 - P(Y \geq 9) - P(Y \leq 1) \\ &= 1 - 2\left(\binom{10}{9} + \binom{10}{10}\right) 0.5^{10} = 0.98. \end{aligned}$$

Den andra olikheten följer av följande resonemang: att $X_{(2)} > x_{0.5}$ betyder att den näst minst observationer är större än medianen. Detta betyder att 9 eller 10 av observationerna är större än medianen. Motsvarande resonemang förklarar den andra termen.

Konfidensgraden för det tredje intervallet blir således:

$$\begin{aligned} P(X_{(3)} \leq x_{0.5} \leq X_{(8)} &= 1 - P(x_{0.5} < X_{(3)}) - P(X_{(8)} > x_{0.5}) = 1 - P(Y \geq 8) - P(Y \leq 2) \\ &= 1 - 2\left(\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right) 0.5^{10} = 0.89. \end{aligned}$$

Ett 95% konfidensintervall efterfrågades, så vi bör välja det intervallet som har en konfidensgrader närmast över 95%.

Vårt erhållna intervall blir således $[x_{(2)},\ x_{(9)}] = [4.5,\ 20.4]$, och detta konfidensintervall har 98% konfidensgrad. Nästa smalare intervall ges av $[x_{(3)},\ x_{(8)}] = [7.9,\ 14.2]$, men detta intervall har bara 89% konfidensgrad.