STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
LÖSNINGAR
10 januari 2024

Lösningar

Tentamen i Statistisk analys, 10 januari 2024

Uppgift 1

a) Sant

b) Falskt

c) Sant

d) Sant

e) Sant

Uppgift 2

a) Data verkar hyfsat likna normalfördelningen. Vi får $\bar{x} = 69.82$ och $s = 1.15$. Ett 95% konfidensintervall ges således av $\bar{x} \pm t_{0.05}(7)s/\sqrt{8} = 69.82 \pm 2.36 * 1.15/\sqrt{8} = (68.86,\ 70.78)$.

b) Vår teststatistika blir

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{8}} = \frac{69.82 - 71}{1.15/\sqrt{8}} = -2.90.$$

Vi förkastar $H_0 : \mu = 71$, till förmån för $H_A : \mu \neq 71$ på 1%-nivån om $|t_{obs}| > t_{0.005}(7) = 3.50$. Eftersom detta inte gäller $(2.90 < 3.50)$ så förkastar vi ej H_0. Slutsatsen är alltså att vi inte kan utesluta möjligheten att $\mu = 71$.

Uppgift 3

a) Vi änvänder linjär regression vilket stöds ganska väl av en graf. Det betyder att $Y_i = \alpha + \beta x_i + \epsilon_i$, där ϵ_i är oberoende och $E(\epsilon_i) = 0$ och $V(\epsilon_i) = \sigma^2$.

b) Från de givna summorna och formelsamlingen får vi $S_{xx} = \sum_i x_i^2 - n^{-1}\left(\sum_i x_i\right)^2 = 3251 - 131^2/6 = 390,8$, $S_{yy} = \sum_i y_i^2 - n^{-1}\left(\sum_i y_i\right)^2 = 33344 - 440^2/6 = 1077.3$, och $S_{xy} = \sum_i x_i y_i - n^{-1}\left(\sum_i x_i\right)\left(\sum_i x_i\right) = 10234 - 131 * 440/6 = 627.3$.

σ^2 skattas med $s^2 = SSE/(n-2) = \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)/4 = 17.59$, så vi får $s = 4.19$.

Vi får därmed att skattningen för β ges av $\beta^* = S_{xy}/S_{xx} = 1.61$. Skattningens standardavvikelse ges av $s/\sqrt{S_{xx}} = 0.21$.

Skattningen för α blir $\alpha^* = \bar{y} - \beta^*\bar{x} = 440/6 - 1.61 * 131/6 = 38,2$. Standardavvikelsen är $s * \sqrt{\sum_i x_i^2}/\sqrt{nS_{xx}} = 4.19 * \sqrt{3251}/\sqrt{6 * 390.8} = 4.93$.

c) Utan reklam, dvs $x = 0$ blir $E(Y) = \alpha + \beta * 0 = \alpha$. Detta skattas således med $\alpha^* = 38,2$ tkr. Ett 95% konfidensintervall ges av $\alpha^* \pm t_{\alpha/2}(n-2) * s_{\alpha^*} = 38,2 \pm 2.776 * 4.93 = [24.5,\ 51.9]$.

d) Den föväntade ökningen av försäljning per tusen krona satsad reklam är just β. Detta skattas med $\beta^* = 1.61$ tkr. Ett 95% konfidensintervall ges av $\beta^* \pm t_{\alpha/2}(n-2) * s_{\beta^*} = 1.61 \pm 2.776 * 0.21 = [1.03,\ 1.19]$.

Uppgift 4

a) Nollhypotesten H_0 antar att säkerhetsöversynen inte har någon effekt vilket innebär skillnaden mellan skadebeloppen före och efter har en symmetrisk fördelning runt 0. Alternativhypotesen H_1 antar att säkerhetsöversyn har positiv effekt varför fördelningen för skillnaden mellan skadebelopp före och efter har en tyngdpunkt på negativa värden.

Skadebeloppen varierar i storlek och om man tittade på de absoluta förändringarna skulle ett par observationer styra medelförändringen helt. Av denna anledning är det rimligare att titta på ordningen (rangerna) på förändringarna.

b) Ett lämpligt test är teckenrangtestet. För varje observation bildar vi skillnaden mellan skadebopp året innan förändringen och året efter. Vi får att dessa skillnader d_i blir: -18.0, -1.7, -11.4, -3.3, -6.4, 2.4, -3.0 och 0.7. Om vi rangordnar absolutbeloppen av dessa skillnader får vi: 8, 2, 7, 5, 6, 3, 4, 1. Rangerna hörandes till positive förändringar blir således $T_+ 3 + 1 = 4$. Vi förkastar H_0 om T_+ är signifikant litet. Vilket värde K som utgör gränsen att förkaste hittar man i Tabell 10. För felrisken 0.05 ser man att med $n = 8$ observationer ska man förkasta H_0 om $T_+ \leq 5$. Eftersom vårt observerade värde är $T_+ = 4$ betyder det att vi ska förkasta H_0. Slutsatsen är således att säkerhetsöversynen har en gynnsam effekt och minskar skadeloppen på fastigheter.

Uppgift 5

a) Om utfallet, god eller dålig fångst, för varje fisketur sker oberoende av tidigare fångster och har sannolikheten $p = 0.5$ varje gång, så betyder det att antal gånger man behöver fiska till första goda fångsten blir $ffg(p = 0.5)$, dvs $p(k) = P(X = k) = p(1-p)^{k-1} = 0.5^k$.

Om fiskaren fiskade på *samma* ställe två dagar i rad är det rimligt att tro att god skörd en viss dag minskar chansen för god skörd andra dagen på samma ställe. I så fall skulle ffg inte vara en rimlig modell.

b) Vi får $p(1) = 1/2, \ p(2) = 1/4, \ p(3) = 1/8$ och $p_{4+} = 1/8$. Vid $n = 42$ mätningar blir motsvarande förväntade värden $e_i = np(i) = 40p(i)$. Enligt formeln först på tentan beräknar vi således

$$Q = \sum_i (n_1 - e_i)^2/e_i = (12-21)^2/21 + (14-10.5)^2/10.5 + (10-5.25)^2/5.25 + (6-5.25)^2/5.25 = 9.43.$$

Det finns fyra möjliga utfall, men eftersom $\sum_i n_i = n = 42$ är given så är det bara 3 fria utfall. Ingen parameter har skattats. Antalet frihetsgrader är således $f = 3$. Om H_0 är sann, dvs att data kommer från $ffg(p = 0.5)$ så bör Q approximativt följa $\chi^2(3)$, men om H_0 inte är sann bör Q vara större. Från tabellen ser vi att gränsvärdet för felrisk $\alpha = 0.05$ är $\chi^2_{0.05}(3) = 7.81$.

Eftersom $Q_{obs} = 9.43 > 7.81$ så förkastar vi H_0. Data tycks tydligen *inte* komma från $ffg(p = 0.5)$.

Uppgift 6

a) Likelihooden för $n_1, \ldots, n_5$ blir:

$L(p) \propto p_X(1)^{n_1} p_X(1)^{n_1} p_X(2)^{n_2} p_X(3)^{n_3} p_X(4)^{n_4} p_X(5)^{n_5} = p^{\sum_i n_i}(1-p)^{\sum_i (i-1)n_i} = p^{42}(1-p)^{54}.$

b)

Loglikelihooden blir således $\ell(p) \propto 42\log(p) + 54\log(1-p)$. Vi får

$$\ell'(p) = \frac{42}{p} - \frac{54}{1-p}.$$

Om vi sätter $\ell'(p) = 0$ och löser ut p så får man $\hat{p} = \frac{\sum_i n_i}{\sum_i n_i + \sum_i (i-1)n_i} = \frac{42}{42+54} = 42/96 = 0.4375$.

c) Totalt så är fiskaren ute och fiskar $\sum_i in_i = 96$ dagar och utav dessa får hen god fångst 42 gånger. $\hat{p} = 42/96$ blir således den relativa andelen fiskedagar med god fångst.