STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
LÖSNINGAR
5 januari 2021

Lösningar

Tentamen i Statistisk analys, 5 januari 2021

Uppgift 1

a) Falskt

b) Sant

c) Sant

d) Sant

e) Falskt

Uppgift 2

a) Beräkningar ger att $\bar{x} = 3.019$ och $s^2 = \left(\sum_i x_i^2 - n\bar{x}^2\right)/7 = 0.136^2$. Vi får $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = 0.40$. Detta skall jämföras med t-fördelning med 7 frihetsgrader. Inte alls signifikant (oberoende av rimligt signifikansnivå)!

b) Ett 95% konfidensintervall för μ ges av $\bar{x} \pm t_{0.025}(7)s/\sqrt{8} = 3.019 \pm 0.113 = (2.906,\ 3.132)$.

Uppgift 3

a) Andel positiva i respektive stad skattas med $\hat{p}_A = 110/500 = 0.22$ respektive $\hat{p}_B = 91/500 = 0.182$. Under $H_0 : p_A = p_B =: p$ har de samma

sannolikhet p vilken skattas med $p^* = (110+91)/1000 = 0.201$. H_0 förkastas om

$$T = \frac{\hat{p}_A - \hat{p}_B}{p^*(1-p^*)(1/500+1/500)}$$

är stor eller liten i relation till standardiserad normalfördelning (se s 352). I vårt fall blir $t_{obs} = (0.22 - 0.182)/0.0253 = 1.50$. Detta är inte signifikant, så det går inte att påstå att stad A har större andel immuna and stad B, i alla fall inte att skillnaden är statistiskt säkerställd.

b) En förutsättning är att individernas tillstånd är oberoende. Eftersom det handlar om smittspridning gäller detta inte helt och hållet, men om urvalet dras slumpmässigt är detta inte något problem.

Uppgift 4

a) En lämplig modell är att anta linjär regression, så att $Y = \alpha + \beta x + \epsilon$, där Y är antal IVA-inläggningar, x är vecka och ϵ är slumpfelet. Eftersom Y är heltalsvärd är den inte exakt normalfördelad, ett rimligare antagande är Poisson, men Poission-fördelningen med stort väntevärde liknar mycket normalfördelningen.

Från de beräknade summorna får man enkelt: $\bar{x} = 49$, $\bar{y} = 167.8$, $S_{xx} = \sum_i x_i^2 - n\bar{x}^2 = 10$, $S_{yy} = \sum_i y_i^2 - n\bar{y}^2 = 456.8$ och $S_{xy} = \sum_i x_i y_i - n\bar{x}\bar{y} = 59$. Från formelsamlingen skattningar får vi $\beta^* = S_{xy}/S_{xx} = 5.9$, $\alpha^* = \bar{y} - \beta^*\bar{x} = -121.3$ och $s^2 = SSE/(n-2) = (S_{yy} - S_{xy}^2/S_{xx})/(n-2) = 6.02^2$.

Ett 95% konfidensintervall för β ges av $\beta^* \pm t_{0.025}(3)s\frac{1}{\sqrt{S_{xx}}} = 5.90 \pm 3.18 * 6.02 * 0.316 = 5.9 \pm 6.05$. Eftersom intervallet innehåller 0 så kan $H_0 : \beta = 0$ inte uteslutas. Ökningen är således inte statistiskt säkerställd (även om det är nära). Jag tycker det är mest rimligt med en två-sidig alternativhypotes varför jag gjorde tvåsidigt konfidensintervall. En som gör ensidig mothypotes gör ensidigt intervall, och då förkastas H_0 på denna signifikansnivå.

b) Ett 95% prediktionsintervall för antal inläggningar vecka 52 ges av $I_{Y_{52}}$: $\alpha^* + \beta^* * 52 \pm t_{0.025}(3)s\sqrt{1 + 1/n + (52-\bar{x})^2/S_{xx}} = 185.5 \pm 3.18 * 6.02 * 1.45 = 185.5 \pm 27.7 = (157,\ 214)$ (bäst att välja närmaste heltal utanför sifferintervallet).

Uppgift 5

a) Data kan skrivas som en kontingenstabell med O_{i1} = observerat antal kunder utan skada i stock i och $O_{i2} = 100 - O_{i1}$ observerat antal kunder *med* skada i stock i. Under H_0 om att alla kundstockar är lika blir förväntat antal irespektive cell $e_{i1} = n_i \frac{O.1}{n..} = 100\frac{162}{400} = 40.5$ och $e_{i2} = n_i \frac{O.2}{n..} = 100\frac{238}{400} = 59.5$.

Nollhypotesen testas genom att jämföra $X \sum_{ij}(O_{ij}-e_{ij})^2/e_{ij}$ med χ^2-fördelningen med (4-1)(2-1)=3 frihetsgrader.

Vi får

$$X = \frac{(32-40.5)^2 + (47-40.5)^2 + (39-40.5)^2 + (44-40.5)^2}{40.5} + \frac{(32-40.5)^2 + (47-40.5)^2 + (39-40.5)^2 + (44-40.5)^2}{59.5} = 5.35.$$

Vi har att $\chi^2_{0.95}(3) = 7.81$ så eftersom $5.35 < 7.81$ så kan vi inte förkasta H_0 om att alla kundstockar är jämbördiga.

b) Om man tittar i χ^2-tabellan för 3 frihetsgrader ser man att $\chi^2_{0.75}(3) = 4.11$ och $\chi^2_{0.9}(3) = 6.25$ och 5.31 ligger emellan dessa värden, lite närmre det senare. En rimlig skattning är därför att det ungefär svarar mot $\chi^2_{0.85}(3)$ vilket betyder att p-värdet är ungefär 0.15.

Uppgift 6

a) Testet ifrån uppgift 3 gick ut på att förkasta H_0 om T är stor/liten relativt $N(0,1)$ där

$$T = \frac{\hat{p}_A - \hat{p}_B}{p^*(1-p^*)(1/500 + 1/500)}.$$

Vi väljer här att ersätta p^* med det teoretiska mittvärdet 0.225, men det går även att använda skattningen från uppgift 3. Detta betyder att vi förkastar H_0 om $|\hat{p}_A - \hat{p}_B| > 1.96\sqrt{0.225 * 0.775 * (1/500 + 1/500)} = 0.0518$.

Om $p_A = 0.25$ och $p_B = 0.20$ så är $Z := \hat{p}_A - \hat{p}_B = (N_A - N_B)/500$ approximativt normalfördelad med väntevärde 0.05 och varians $(0.25*0.75 + 0.2*0.8)/500 = 0.0264^2$.

Således gäller (under förutsättning att $p_A = 0.25$ och $p_B = 0.20$)

$$\begin{aligned} P(|Z| > 0.0518) &= P((Z - 0.05)/0.0264 > (0.0518 - 0.05)/0.0264) \\ &\quad + P((Z - 0.05)/0.0264 < (-0.0518 - 0.05)/0.0264) \\ &= (1 - \Phi(0.068)) + \Phi(-3.86) = 0.472 + 0 = 0.472 \end{aligned}$$

Styrkan är således endast 24% om skillnaden mellan p_A och p_B är så liten som 5%.

b) För godtyckligt n så gäller att vi förkastar H_0 om $\hat{p}_A - \hat{p}_B > 1.96\sqrt{0.225 * 0.775 * (1/n + 1/n)} = 1.157/\sqrt{n}$ (vi bortser från den osannolika händelsen att det är mindre än motsvarande negativa belopp). Dessutom så är $Z := \hat{p}_A - \hat{p}_B = (N_A - N_B)/n$ approximativt normalfördelad med väntevärde 0.05 och varians $(0.25 * 0.75 + 0.2 * 0.8)/n = 0.3475/n$.

Därmed förkastar vi H_0 med sannolikhet

$$P(\frac{Z - 0.05}{\sqrt{0.3475/n}} > \frac{1.157/\sqrt{n} - 0.05}{\sqrt{0.3475/n}}.$$

Vi vill att denna sannolikhet ska vara lika med 0.8. Vi finner i Tabellen för normalfördelningen att $1 - \Phi(z) = 0.8$ då $z = -0.84$. Således måste vi ha

$$\frac{1.157/\sqrt{n} - 0.05}{\sqrt{0.3475/n}} = -0.84,$$

vilket ger att $n = 1090$. Så för satt ha styrka 80% i testet av $H_0 : p_A = p_B$ när $p_A = 0.25$ och $p_B = 0.20$ behövs minst 1090 individer testas i respektive stad.