STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
LÖSNINGAR
23 augusti 2022

Lösningar

Tentamen i Statistisk analys, 23 augusti 2022

Uppgift 1

a) Sant

b) Sant

c) Falskt

d) Sant

e) Sant

Uppgift 2

a) Ett 95% konfidensintervall ges av $\bar{x} \pm t_{\alpha/2}(n-1)s/\sqrt{n} = 6.71 \pm 2.145 * \sqrt{0.074}/\sqrt{15} = 6.71 \pm 0.15 = [6.56,\ 6.86]$, eftersom $n = 15$ och $\alpha = 0.05$.

b) H_0 förkastas om $\bar{x}$ är tillräckligt liten. Gränsvärdet sätts så att det underskrids med max 1% sannolikhet om H_0 är sann. Detta gränsvärde ges således av $\mu_0 - t_\alpha(n-1)s/\sqrt{n} = 7 - 2.624 * \sqrt{0.074}/\sqrt{15} = 6.82$. Eftersom $\bar{x} = 6.71$ så förkastar vi H_0. Väntevärdet att kraftigt signifikant mindre än 7.

c) Även om X inte är normalfördelad som kommer $\bar{X}$ baserat på 15 observationer med god approximation vara det enligt centrala gränsvärdessatsen.

Uppgift 3

a) Vi använder linjär regression. Hur mycket medeltempersaturen förändras per lattitud är exakt detsamma som värdet på β eftersom $E(Y(x) - Y(x-1) = \alpha + \beta x - (\alpha + \beta(x-1)) = \beta$. Paramerern β skattas med

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i x_i y_i - n^{-1} \sum_i x_i \sum_i y_i}{\sum_i x_i^2 - n^{-1}(\sum_i x_i)^2} = -0.723$$

Stickprovsvariansen skattas med

$$s^2 = \frac{SSE}{n-2} = \frac{1}{9}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 9^{-1}\left(68.3054 - \frac{81.59^2}{112.78}\right) = 1.0312,$$

så $s = 1.016$.

Antalet frihetsgrader är $n - 2 = 9$ och $\alpha = 0.01$ så $t_{\alpha/2}(n-2) = 3.25$. Ett 99% konfidensintervall för β ges således av $\beta^* \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}} = -0.723 \pm 0.311 = [-1.034,\ -0.412]$.

b) Det gäller att $R^2 = 1 - \frac{SSE}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = 0.864$.

c) En skattning av den förväntade medeltemperaturen i Bollnäs (med lattitud $x_0 = 61.34$) ges av $\alpha^* + \beta^* x_0 = \bar{y} - \beta^* \bar{x} + \beta^* x_0 = \bar{y} + \beta^*(x_0 - \bar{x}) = 5.936 - 0.723(61.34 - 59.73) = 4.79$. Den förväntade medeltemperaturen i Bollnäs är således 4.79 grader.

Uppgift 4

a) Kontingenstabell. Vi beräknar först förväntat antal observationer i cell (i, j): $e_{ij} = n_i * O_{\cdot j}/n$. T ex blir $e_{11} = 50 * 24/150 = 8$ vilket även gäller för övriga i den kolumnen eftersom $n_1 = n_2 = n_3 = 50$. Man får: $e_{11} = e_{21} = e_{31} = 8$, $e_{12} = e_{22} = e_{32} = 13$, $e_{13} = e_{23} = e_{33} = 16$, $e_{14} = e_{24} = e_{34} = 13$. Man får då att $Q = \sum_{i,j}(O_{ij} - e_{ij})^2/e_{ij} = 12.16$. Antal frihetsgrader är $(3-1) * (4-1) = 6$. Från tabellen har vi att $\chi^2_{0.05}(6) = 12.59$. Således går det inte att hävda att färdsätten mellan städerna skiljer sig åt signifikant, men det är ju ganska nära, så en viss tendens åt att det är så föreligger likväl.

b) Andelen som tar bil i Stockholm är $\hat{p}_S = 13/50 = 0.26$ och i Göteborg $\hat{p}_G = 20/50 = 0.40$. Ett 95% konfidensintervall för skillnaden $p_S - p_G$ ges av

$$\hat{p}_S - \hat{p}_G \pm 1.96 * \sqrt{\frac{\hat{p}_S(1-\hat{p}_S)}{n_S} + \frac{\hat{p}_G(1-\hat{p}_G)}{n_G}} - 0.14 \pm 0.18.$$

Skattingen är således att Göteborg använder bil nominellt 14% mer än Stockholm, men skillnaden är inte statistiskt säkerställd.

Uppgift 5

a) Eftersom de möjliga utfallen är ganska få, men framför allt för att fördelningen verkar vara skev med några få väldigt höga värden så kan normalfördelningen absolut ifrågasättas. Det kan t.o.m. vara så att antal partners har ett väntevärde som är nästintill oändligt. I sådana fall är test av väntevärden olämpliga.

b) Alla observationer med samma värde ges samma rang, och den rangen bestäms så att summan av rangerna görs oförändrad. T ex är de fyra minsta observationerna i datamaterialet alla 0. Eftersom dessa ska ges rang 1 till 4 ges alla samma rang $(1 + 2 + 3 + 4)/4 = 2.5$.

c) Vi använder oss av Wilcoxons 2-stickprovs test (även kallat Mann-Whitney). Alla observationer rangordnas och summan av rangerna för det mindre stickprovet adderas. Detta ger för männen $r_{obs} = 76.5$. Antal obs för männen är $m = 8$ och för kvinnor $n = 10$. Man förkastar hypotesen om identiska fördelningar om $R \leq k^-$ eller om $R \geq k^+$, där k^- och k^+ är 2.5% resp 97.5% kvantiler som man finner i tabell. Från tabellen ser man att $k^- = 53$ och $k^+ = 99$ och eftersom 76.5 ligger emellan dessa värden så förkastar vi inte H_0. I själva verket ligger r_{obs} väldigt nära det förväntade värden $E(R) = 76$.

Uppgift 6

a) Parametern p ligger ju mellan 0 och 1. Om vi inte vet något på förhand om p så förefaller det rimligt att antag att den är likformigt fördelad på $[0, 1]$-intervallet. Eftersom observationerna kan antas oberoende och slumpvis utvalda så blir $X \sim Bin(50, p)$.

b) Det gäller ju att aposteriofördelningen är proportionell mot likelihooden multiplicerat med aprioriförde lningen. Det betyder således

$$f(p|X = 17) \propto L(p|X = 17)f(p) \propto p^{17}(1 - p)^{33} * 1 \text{ för } 0 \leq p \leq 1$$

c) Eftersom aposteriorifördelningen $f(p|X = 17)$ kan skrivas som $f(p|X = 17) \propto p^{18-1}(1 - p)^{34-1}$ så är detta således en beta-fördelning: $p|X = 17 \sim Beta(m = 18, n = 34)$. Denna fördelning har väntevärde $m/(m + n) = 34/52 = 0.654$. De flesta skulle nog ha gissat att p var den observerade

frekvensen, dvs 33/50=0.66, så nästan men inte riktigt detsamma som väntevärdet i aposteriorifördelningen.