STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
LÖSNINGAR
7 februari 2025

Lösningar

Tentamen i Statistisk analys, 7 februari 2025

Uppgift 1

a) Sant

b) Sant

c) Falskt

d) Sant

e) Falskt

Uppgift 2

a) Modellen är att $Y_i = \alpha + \beta x_i + \epsilon_i$ för $i = 1, \dots k$, där $\epsilon_1, \dots, \epsilon_k$ är oberoende och likafördelade slumpvariabler med väntevärde 0 och samma varians σ^2 (ibland antas även normalfördelning). Att variansen är konstant verkar någorlunda rimligt, liksom antagandet att bnp växer ungefär linjärt. Året 2020 var såklart speciellt (pandemin) vilket är förklaringen till att bnp inte ökade det året så denna observation är lite speciell, men det antagande som framförallt inte gäller är den om obeorende mellan olika år: vilket bnp det blir i år är väldigt beroende av vilket bnp det blev förra året.

b) Den förväntade årliga tillväxttakten är β som skattas med $\beta^* = S_{xy}/S_{xx} = (92060.2 - 18171 * 45.59/9)/(36687309 - 18171^2/9) = 14.32/60 = 0.239$. Den genomsnittliga linjära tillväxten av bnp är således 0.239 tusen miljarder kr, eller 239 miljarder kronor,

c) Ett 95% prediktionsintervall för bnp år 2024 (under antagandet om oberoende observationer), ges av

$$\alpha^* + \beta^* * 2024 \pm t_{0.025}(9-2) * s * \sqrt{1 + \frac{1}{n} + \frac{(2024 - \bar{x})^2}{S_{xx}}}.$$

Vi har $n = 9$, $\alpha^* = \bar{y} - \beta^*\bar{x} = -477.4754$, $t_{0.025}(7) = 2.365$, samt $s^2 = (S_{yy} - S_{xy}^2/S_{xx})/7 = (3.569 - 14.32^2/60)/7 = 0.0219$, så $s = 0.148$. Detta ger intervallet

$$6.26 \pm 2.365 * 0.148\sqrt{1 + 1/7 + 5^2/60} = 6.26 \pm 0.44 = [5.82, 6.67].$$

Notera att detta är ett väldigt brett intervall vilket beror på vårt (inkorrekta) antagande att bnp 2024 inte beror på bnp året innan.

Uppgift 3

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 4

Under H_0 har ungdomarna från de olika länderna samma (okända) sannolikhet att prioritera skola, vård/omsorg eller försvar. Dessa sannolikheter skattas med den totala andelen som prioriterar respektive område: $p_S = 44/150$, $p_V = 48/150$ och $p_F = 58/150$. De förvänta antalen e_{ij}i respektive land blir dessa sannolikheter multiplicerat med 50 (eftersom det är 50 ungdomar i respektive land. De förväntade andelarna blir således:

Land	Skola	Vård/omsorg	Försvar	Summa
Estland	14.67	16	19.33	50
Belgien	14.67	16	19.33	50
Portugal	14.67	16	19.33	50
Antal	44	48	58	150

b) Vi kan nu beräkna $Q = \sum_{ij} \frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 4.91$. Detta ska jämföras med en χ^2-fördelning med $(3-1)*(3-1) = 4$ frihetsgrader. Då är 4.91 inte alls särskilt extremt. Gränsen för att förkasta H_0 ligger vid $\chi^2_{0.05}(4) = 9.49$. Så trots att det verkar som att en större andel i Estland värderar försvaret högre så kan en sådan avvikelse absolut ske av ren slump. Det finns ingen anledning att påstå att ländernas ungdomar har signifikant skilda prioriteringar.

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) Variansen σ^2 skattas med $s^2 = (n-1)^{-1} \sum_i (x_i - \bar{x})^2 = 0.300$, så σ skattas med $s = 0.548$.

b) En referensvariabel för σ^2 ges enligt formelsamlingen av $R = (n-1)s^2/\sigma^2 \sim \chi^2(n-1)$. Det gäller således att $P(\chi^2_{0.95}(n-1) \leq (n-1)s^2/\sigma^2 \leq \chi^2_{0.05}(n-1)) = 0.9$. Om vi stuvar om i olikheterna får vi $(n-1)s^2/\chi^2_{0.05}(n-1) \leq \sigma^2 \leq (n-1)s^2/\chi^2_{0.05}(n-1)$. Motsvarande numeriska nedre och övre gränser blir $[7 * 0.3003/14.067, 7 * 0.3003/2.167] = [0.149, 0.970]$.

Ett 90% konfidensintervall för σ erhålls genom att ta roten ur respektive gräns: $[0.39,\ 0.98]$. Punktskattningen $s = 0.548$ ligger således klart närmre den nedre gränsen av konfidensintervallet, men det är inte konstigt eftersom χ^2-fördelningen inte är symmetrisk.