STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MS 4001
LÖSNINGAR
13 januari 2020

Lösningar

Tentamen i Statistisk analys, 13 januari 2020

Obs: Fyll i kursvärderingen!!!

Uppgift 1

a) Falskt

b) Sant

c) Sant

d) Falskt

e) Falskt

Uppgift 2

a) Det gäller att $\mu^* = \bar{x} = 59.74$. Variansen skattas med stickprovsvariansen $s^2 = (n-1)^{-1}\sum_i (x_i - \bar{x})^2 = (n-1)^{-1}\left(\sum_i x_i^2 - n\bar{x}^2\right) = 11.28$. Eftersom fördelningen är snäll är medelvärdet $\bar{X}$ approximativt normalfördelad och ett 95% konfidensintervall ges således av ($\alpha = 0.05$ och $n = 11$)

$$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} = 59.74 \pm 2.23\frac{3.36}{3.32} = 59.74 \pm 2.26 = (57.48,\ 62.00).$$

b) En symmetrisk fördelning har alltid $\mu = \tilde{m}$, dvs väntevärdet är lika med medianen. Denna skattas lämpligen med medianen i datamaterialet som är $\mu^* = x_{(6)} = 59.2$ ($x_{(i)}$ är den i:te största observationen).

Uppgift 3

a) En rimlig modell är att anta att genomsnittlig vikt ökar linjärt med längden (skulle även kunna vara kvadratiskt). Vi antar därför modellen förkel linjär regression:

$$Y_i = \alpha + \beta x_i + \epsilon_i,\ \epsilon_1, \epsilon_2, \ldots, i.i.d.\ \sim N(0, \sigma^2).$$

Skattningarna blir

$$\beta^* = S_{xy}/S_{xx} = \frac{\sum_(x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{X})^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{103553 - 103413.1}{284769 - 284596.9}$$

$$= 0.813$$

$$\alpha^* = \bar{y} - \beta^* \bar{x} = -75.68.$$

(Att $\alpha^* < 0$ visar att linjär regression inte fungerar för alla längder, men i intervallet av intresse duger modellen gott.)

b) Det gäller att $E(Y(x+1) - Y(x)) = \alpha + \beta(x+1) - (\alpha + \beta x) = \beta$, så vi ska alltså skatta β som är den genomsnittliga viktökningen vid 1 cm längre längd. Variansen σ^2 skattas med $s^2 = SSE/(n-2) = (144.1 - 113.72)/8 = 3.80$. Ett konfidensintervall för β ges av

$$\beta^* \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} = 0.812 \pm 3.355 * \frac{\sqrt{3.81}}{13.12} = 0.29 =$$

c) Ett prediktionsintervall för en kvinna med längd 165cm ges av

$$\alpha^* + \beta^* * 165 \pm t_{\alpha/2}(n-2) * s * \sqrt{1 + n^{-1} + \frac{(165 - \bar{x})^2}{S_{xx}}},$$

vilket blir ??.

Uppgift 4

Vi antar att både stickproven består av sinsemellan oberoende observationer, att båda stickproven har varians σ^2 och att de vanliga studenterna har väntevärde μ_1 och distansstudenterna har väntevärde μ_2. Nollhypotesen är $H_0 : \mu_1 = \mu_2$ och mothypotesen att detta inte gäller, dvs $\mu_1 \neq \mu_2$. Vidare antar vi att fördelningarna är någorlunda snälla så att $\bar{X}$ och $\bar{Y}$ kan antas hyfsat normalfördelade enligt centrala gränsvärdessatsen. Ett 95% konfidensintervall för $\mu_1 - \mu_2$ ges då av

$$\bar{x} - \bar{y} \pm t_{\alpha/2}(n + m - 2) s_p \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Vi har $\bar{x} = 38.78$, $\bar{y} = 31.87$, $\alpha = 0.05$, $n = 9$ och $m = 8$. Den poolade stickprovsvarienansen ges av

$$s_p^2 = \frac{\sum(x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2}{n + m - 2} = \frac{\sum x_i^2 - n\bar{x}^2 + \sum_j y_j^2 - m\bar{y}^2}{n + m - 2} = 219.4,$$

så $s_p = \sqrt{219.4} = 14.8$. Till sist får vi från tabell att $t_{0.025}(15) = 2.13$. Vi får således följande konfidensintervall:

$$6.91 \pm 15.32 = (-8.41,\ 22.33).$$

Eftersom 0 ingår i detta intervall kan det inte uteslutas och vi förkastar således *inte* H_0 om att väntevärdena skulle vara lika.

Uppgift 5

Vi använder oss av Wilcoxons 2-stickprovs test (även kallat Mann-Whitney). Nollhypotesen H_0 är i detta fall att de två populationerna har samma fördelning för provresultat, och H_1 är att detta inte gäller. Alla observationer rangordnas gemensamt (se tabellen nedan).

Ranger-vanliga	4	13	7	11	17	12	3	15	8
Värden-vanliga	25	47	33	38	58	43	20	51	34
Värden-distans	10	31	48	32	5	36	56	37	
Ranger-distans	2	5	14	6	1	9	16	10	

Summan av rangerna för det mindre stickprovet (distanseleverna) adderas vilket ger $r_{obs} = 63$. Vi har $m = 8$ och $n = 9$. Från tabell ser vi att vi ska förkasta nollhypotesen om antingen $r_{obs} \leq 51$ eller om $r_{obs} \geq 93$. Eftersom detta inte föreligger så förkastas *inte* H_0. Inte heller icke-parametriska metoder kan påvisa att de två studentgrupperna skiljer sig åt signifikant. (Om vi använder normal approximation av R får vi att den under H_0 är normalfördelad med väntevärde 72 och standardavvikelse 10.4, och då är inte heller 63 signifikant.)

Uppgift 6

Vi har här att göra med en kontingenstabell. Nollhypotesen är att att de två behandlingsmetoderna är ekvivalenta. Detta testas genom att beräkna

$$X = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{i,j})^2}{\hat{n}_{i,j}},$$

Där $n_{i,j}$ är antalet observationer i respektive cell, och $\hat{n}_{i,j} = n_{i,\cdot}n_{\cdot,j}/n_{\cdot,\cdot}$ är det förväntade antalet observationer under H_0. Under H_0 är $X \sim \chi^2((n-1)(m-1))$, dvs att antaler frihetsgrader är antal rader minus ett multiplicerat med antal kolumner minus ett, i vårt fall $(2-1)*(3-1) = 2$.

För våra data blir $\hat{n}_{i,j} = 10$ för all i, j eftersom alla radsummer är 30 och kolumnsummor 20. Vi får att $X = 7.6$ vilket skall jämföras med $\chi^2_{0.05}(2) = 5.99$ vilket betyder att vi förkastar H_0. Vi drar alltså slutsatsen att de två behandlingesmotederna skiljer sig åt signifikant. Det tycks som att behandlingsmetod B har en snabbare verkan, alternativt ingen verkan och dödsfall. Vilken behandling som är att föredra är en annan fråga.