STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Matematisk statistik, TB

MT4001
LÖSNINGAR
21 februari 2024

Lösningar

Tentamen i Statistisk analys, 21 februari 2024

Uppgift 1

a) Sant

b) Sant

c) Sant

d) Sant

e) Falskt

Uppgift 2

a) Vi ansätter linjär regression med ålder x som förklarande variabel och Y som responsvariabel. Modellen blir således

$$Y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, \ldots, n,$$

där α och β är okända parametrar och $\epsilon_1, \ldots, \epsilon_n$ är oberoende och lika fördelade med $E(\epsilon) = 0$ och $var(\epsilon) = \sigma^2$ där σ är okänd.

b) Från formelsamlingen har vi

$$\hat{\beta} = S_{xy}/S_{xx} = (3091 - 270 * 76.3/6) \,/\, \left(13900 - 270^2/6\right) = -342.5/1750 = -0.195$$

$$\hat{\alpha} = \bar{y} - \beta^* \bar{x} = 21,49$$

$$\hat{\sigma}^2 = s^2 = (S_{yy} - S_{xy}^2/S_{xx})/(n-2) = 4^{-1}\left(1045.23 - 76.3^2/6\right)$$
$$- 4^{-1}\left(3091 - 270 * 76.3/6\right)^2 / \left(13900 - 270^2/6\right) = 7.92/4 = 1.979.$$

Med andra ord skattas σ med $s = \sqrt{1.979} = 1.41$.

c) Ett 95% konfidensintervall ges av $\hat{\beta} \pm t_{0.025}(4) * s/\sqrt{S_{xx}} = -0.195 \pm 2.776 * 1.41/41.8 = -0.195 \pm 0.093 = [-0.288, -0.102]$. Eftersom nollhypotesens värde ($\beta = 0$) ligger utanför detta intervall så förkastas nollhypotesen. Slutsatsen är således att det finns en individeffekt: vissa har generellt fler kontakter och andra färre, och det är inte bara ren slump mellan olika dagar.

Uppgift 3

a) Observationerna är parvisa, så vi bildar skillnaden $d_i = y_i - x_i$ i resultat mellan är 2021 och år 2020 för respektive institution. Dessa blir: 1.6, 0.7, 2.2, -0.7, 0.3, 1.3, 1.5, 0.2. Dessa får ett medelvärde på $\bar{d} = 0.89$ och en standardavvikelse på $s = 0.94$. Vi bildar teststatistikan $t = (\bar{d} - 0)/(s/\sqrt{n}) = 2.68$. Vi gör ett 95% tvåsidigt test vilket betyder att vi ska förkasta H_0 : $\mu_d = 0$ om $|t| > t_{0.025}(7) = 2.36$. Eftersom detta är fallet så förkastas H_0. Dvs projektet ledde till att SUs institutioner fick ett bättre resultat år 2021.

b) Ett 95% konfidensintervall för effekten μ_d ges av $\bar{d} \pm t_{0.025}(7) * s/\sqrt{n} = 0.89 \pm 0.78 = [0.12, 1.67]$. I genomsnitt fick institutionerna en knapp miljon i bättre resultat.

c) Institutionerna har vissa liknande kostnader vilket gör att resultaten för olika institutiuoner knappast kan ses som oberoende, t ex har man samma samma procentuella löneökningar och hyreshöjningar. Storleken på institutionerna varierar ganska mycket. Resultatet på en institution som omsätter ca 200 MSek bör ha större spridning än en institution med 50 MSek i omsättning.

Uppgift 4

b) Pearsonkorrelationen ges av

$$r_{xy} = S_{xy}/\sqrt{S_{xx}S_{yy}} = (1244 - 91 * 121/10)/\sqrt{(943 - 91^2/10)(1917 - 121^2/10)} = 0.626.$$

c) Vi konstruerar ett test genom att beräkna $T = \sqrt{n-2}r_{xy}/\sqrt{1 - r_{xy}^2} = 2.27$. Denna ska jämföras med $t(n-2)$-fördelningen. I detta fall verkar ett ensidigt test vara rimligast eftersom det verkar otroligt att det skulle finnas ett negatvit samband mellan en individs kontakter olika dagar (2-sidigt

godkänns dock). Från tabellen ser vi att $t_{0.05}(8) = 1.86$. Eftersom vårt observerade t-värde är 2.27 vilket är större kan vi utan påstå att det finns en individeffekt. Ett tvåsidigt test hade emellertid inte förkastat H_0.

Uppgift 5

a) Det finns en observation (29 kontakter en enskild dag) som sticker ut rejält och därför påverkar både medelvärde och standardavvikelse rejält. Detta svarar mot att fördelningen över antal kontakter verkar ha en tjock högersvans, alternativt att det är något felaktigt med denna observation.

b) Vi rangordnar de olika x-värdena och de olika y-värdena (för att slippa problemet med en extrem observation. Data materialet blir då (två lika värden ges mittrangen): (7.5, 9), (1, 4.5), (3, 3), (10,6), (9,10), (4,1), (6, 4.5), (5,7), (7.5, 8), (2, 2). Spearmankorrelationen blir $r_s = 1 - [6/(n(n^2 - 1))] \sum_i (rang(x_i) - rang(y_i))^2 = 0.715$. Teststatistikan ges av $T = \sqrt{n-1} r_s = 2.15$. Denna ska jämföras med normalfördelningen och även här verkar ensidigt test mest rimligt. Vi bör således förkasta om $T > \lambda_{0.05} = 1.645$ så även detta test förkastar H_0 och talar således för att et finns en tydlig individeffekt: vissa individer har generellt fler (resp färre) kontakter. Värt att notera är att Spearmankorrelationen var större än Pearsonkorrelationen. Anledningen är att den extrema observation för att den linjära regression för mindre förklaring då en observation avviker kraftig från linjen.

Uppgift 6

a) Vi antar att medelvärdet kan approximeras med normalfördelningen och bildar således

$$T = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

för att avgöra om H_0 är rimlig eller inte. Eftersom vi har ett ensidigt test, $\mu_0 = 80$, och $\sigma = 10$ är känd så ska vi förkasta H_0 om $T > \lambda_{0.05} = 1.645$, dvs om $\bar{x} > 80 + 16.45/\sqrt{n}$.

b) Vi antar nu att sant $\mu = 82$, och vi ska beräkna vad sannolikheten att förkasta H_0 blir för olika värden på n. Vi får

$$\begin{aligned} P(\text{förkasta } H_0 | \mu = 82) &= P\left(\frac{\bar{X} - 80}{\sigma/\sqrt{n}} > 1.645 | \mu = 82 \right) \\ &= P\left(\frac{\bar{X} - 82}{\sigma/\sqrt{n}} > 1.645 - \frac{2}{\sigma/\sqrt{n}} \right) = 1 - \Phi(1.645 - 0.2\sqrt{n}) \end{aligned}$$

Vi vill att denna sannolikhet ska bli 0.8. Detta betyder att $1.645 - 0.2\sqrt{n}$ måste vara -0.84 (detta värde fås från normalfördelningstabellen: $P(Z > -0.84) = 0.80$. Ekvationen blir således

$$1.645 - 0.2\sqrt{n} = -0.84$$

Vi måste således ha minst $n = 155$ observationer för att kunna påvisa en ökning med 2 kg med 80% sannolikhet.