Petter Mostad
Matematisk Statistik
Chalmers

Förslag till lösningar för tentamen för
MSG830: Statistisk analys och experimentplanering
17 Januari 2014

1. -0.87: Korrelation är alltid mellan -1 och 1, och eftersom det är en trend i plottet där högre värden för x motsvarar lägre värden för y så är korrelationen negativ.

2. (a) Det är naturligt att använda en Poissonfördelning med rate 8. Sannolikheten för 5 personer blir
$$e^{-8}\frac{8^5}{k!} = 0.09160366$$
(b) Vi använder en approximation med en normalfördelning med väntevärde 8 och varians 8. Vi jämför då
$$\frac{14.5 - 8}{\sqrt{8}} = 2.298097$$
med en stanard normalfördelning och tabellen ger sannolikheten 0.011. (Det exakta värdet är 0.017).
(c) Använd Bayes formel. Under betyder X att Johan har X, och S att Johan har S.
$$\begin{aligned}
\pi(X) &= \frac{8}{500000} \\
\pi(S \mid X) &= 0.75 \\
\pi(S \mid \text{icke } X) &= 0.001 \\
\pi(X \mid S) &= \frac{\pi(S \mid X)\pi(X)}{\pi(S \mid X)\pi(X) + \pi(S \mid \text{icke } X)\pi(\text{icke } X)} = 0.012
\end{aligned}$$
(d) I utgångspunkten påväras inte sannolikheten alls. OM modellen är korrekt så är sannolikheten för om Johan är sjuk oberoende av antalet andra sjuka eller vem det är. MEN: OM modellen för sjukdomen är korrekt, att den förekommer i medelvärde hos 8 personer varje år, och oberoende i varje person, så är sannolikheten för 15 eller flera sjuka mycket liten (under 2 procent), och därmed är sannolikheten för 18 eller flera sjuka ännu mycket mindre. När man ändå har observerat 18 sjuka personer, så ger det grund till att misstänka att modellen kan vara fel: Sjukdomen kan vara vanligare än man trott, eller även kan det vara att fallen inte är oberoende. I så fall skulle sannolikheten att Johan har sjukdomen kunna öka något.

3. (a) Den “poolade” variansen blir
$$s_p^2 = \frac{0.2143 \cdot 39 + 0.4501 \cdot 39}{39 + 39} = 0.3322$$

Ett 95% kredibilitetsintervall blir

$$\begin{aligned} & \Big[3.62 - 3.39 - t_{0.025,40+40-2} \cdot \sqrt{0.3322 \cdot (1/40 + 1/40)}, \\ & \quad 3.62 - 3.39 + t_{0.025,40+40-2} \cdot \sqrt{0.3322 \cdot (1/40 + 1/40)}\Big] \\ = & \Big[0.23 - 1.99 \cdot 0.5763679 \cdot \sqrt{1/20}, \\ & \quad 0.23 + 1.99 \cdot 0.5763679 \cdot \sqrt{1/20}\Big] \\ = & \; [-0.026, 0.486] \end{aligned}$$

(b) Vi beräknade i (a) att 0 är inom ett 95% kredibilitetsintervall. Det betyder att en hypotestest där man testar om väntevärden är lika skulle ha ett p-värde över 0.05, och man skulle i ett sådant test inte förkasta nollhypotesen om att väntevärdena är lika. På detta sättet är det rimligt att anta att de är lika. Men om de värkligen är lika eller inte skulle bero även på annan information.

(c) I en hypotestest jämföras

$$\frac{0.4501}{0.2143} = 2.100327$$

med en F-fördelning med 39 och 39 frihetsgrader. Tabellen ger ett p-värde på mindre än 0.05. Enligt hypotestesten förkastas nollhypotesen om att fördelningsvarianserna är lika, och detta antagandet är alltså “inte rimligt”. Beräkningarna i (a) och (b) borde göras om.

4. Nollhypotesen i en Mann-Whitney U test är att alla data kommer från samma sannolikhetsfördelning. Idén med testen är att om H_0 är sann så är det osannolikt att de flesta x-värden är mindre än de flesta y-värden (eller att de flesta x-värden är större än de flesta y värden). Teststatistikan konstrueras genom att först ordna alla x och y värden i en gemensam lista. Låt R_1 betäckna summan av alla ordningstalen för x-värdena. Definera

$$U = nm + \frac{n(n+1)}{2} - R_1$$

Detta är (en version av) teststatistikan: Om den är mycket liten eller mycket stor förkastas nollhypotesen.

5. Vi får

$$\begin{aligned} SS_{\text{Gödsel}} &= 6 \cdot ((39.66667 - 41.89167)^2 + (44.48333 - 41.89167)^2 + \\ &\quad (44.93333 - 41.89167)^2 + (38.48333 - 41.89167)^2) = 195.2148 \\ SS_{\text{Pesticid}} &= 12 \cdot ((39.78333 - 41.89167)^2 + (44 - 41.89167)^2) = 106.6818 \\ SS_{\text{Total}} &= 23 \cdot 29.36775 = 675.4583 \end{aligned}$$

som ger ANOVA-tabellen

	Sum of squares	Deg. freedom	Mean squ.	F	p
Gödsel	195.2148	3	65.0716	3.30955	$0.025 < p < 0.05$
Pesticid	106.6818	1	106.6818	5.42602	$0.025 < p < 0.05$
Residualer	373.5617	19	19.66114		
Total	675.4583	23			

Båda p-värden är mindre än 0.05. Det betyder att båda val av gödsel och val av pesticid har en signifikant påvärkan på mängd potatis odlat, förutsatt att den additiva modellen som beräkningarna byggar på är korrekt.

6. Alexey säger sig vilja jämföra försurning av sjöar i Sverige och Finland generellt. Då borde sjöerna vara stickprov från alla sjöar i Sverige och Finland, vilket de inte är, de kommer från vissa områden. Detta kan helt klart ge felaktiga resultat (speciellt eftersom geografi har stor invärkan på försurning av sjöar). Ett annat allvarlig problem är att mätningarna är gjort på olika tider av året. Detta kan också påvärka, eftersom graden av försurning kan variera under året.

7. En möjlighet är

A	B	C
-	-	+
-	+	-
+	-	-
+	+	+

8. Ett typ 1 fel inträffer när nollhypotesen är sann men den blir förkastad i testet. Ett typ 2 fel inträffer när nollhypotesen är falsk, men den blir inte förkastad i testet.

9. Vi använder en “goodness-of-fit” test:

$$\frac{(14 - 40 \cdot 0.34)^2}{40 \cdot 0.34} + \frac{(12 - 40 \cdot 0.22)^2}{40 \cdot 0.22} + \frac{(3 - 40 \cdot 0.15)^2}{40 \cdot 0.15} + \frac{(11 - 40 \cdot 0.29)^2}{40 \cdot 0.29} = 2.706436$$

skall jämföras med en χ^2 fördelning med 3 frihetsgrader: Detta ger ett p-värde större än 0.05. Hypotesen om att biotopen *inte* skiljer sig från vanliga biotoper kan alltså inte förkastas.

10. (a) Posteriorifördelningen blir

$$p \mid \text{data} \sim \text{Beta}(4 + 27, 6 + 72 - 27) = \text{Beta}(31, 51)$$

(b) Ett 95% kredibilitetsintervall blir

$$[0.273, 0.483] \approx [0.27, 0.48]$$

(c) Väntevärdet:

$$\frac{31}{31+51} = 0.3780488$$

Fördelningsvariansen:

$$\frac{31 \cdot 51}{(31+51)^2(31+51+1)} = 0.002832866$$

Ett 95% kredibilitetsintervall kan approximeras med ett 95% kredibilitetsintervall för en normalfördelning med väntevärde 0.3780488 och fördelningsvarians 0.002832866:

$$[0.3780488 - 1.96 \cdot \sqrt{0.002832866}, 0.3780488 + 1.96 \cdot \sqrt{0.002832866}] \approx [0.27, 0.48]$$

11. (a) Figur 2 viser hur mycket varians som behållas om man tar med variansen bara långsmed de olika principalkomponenterna. Man ser att nästan all varians på detta sättet innehålls i komponent 1 (den längst till vänster) och komponent 2 (den nästa). Detta visar att data ligger nästan i ett plan spänt ut av dessa två vektorer, det är mycket liten varians utanför detta planet. En konsekvens är att Figur 2 ger en ganska korrekt bild av den relativa distansen mellan olika observationer.

(b) En viktig observation är att de 50 städerna värkar dela sig i två klart avgränsade grupper. Dessa två grupper visas till höger och till vänster i Figur 3. Figuren visar även villka av de ursprungliga observationsvariablerna som bidrar till de principalkomponenterna som data varierar med. Detta är variabel B och variabel D. Det är alltså dessa mätningar som klart skiljer städerna åt, och som bidrar med en klassificering i två grupper. Det värkar vara mycket mindre variabilitet i variablerna A, C, och E.