UPPSALA UNIVERSITET
Matematiska institutionen
Raimundas Gaigalas

Matematik och statistik NV1
Statistikdelen
2004-06-10

Lösningar till tenta 040610

1. Låt A vara händelsen "en rätt innehåller för mycket fett", $B =$ "en rätt innehåller för mycket snabba kolhydrater".

   (a) Vi beräknar först sannolikheten att en rätt har minst ett av dessa fel:
   $$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7 + 0.4 - 0.3 = 0.8.$$
   Sannolikheten att en rätt har varken för mycket fett eller snabba kolhydrater:
   $$P(A^* \cap B^*) = 1 - P(A \cup B) = 1 - 0.8 = 0.2.$$

   (b) $P(B|A \cup B) = \dfrac{P(B \cap (A \cup B))}{P(A \cup B)} = \dfrac{P(B)}{P(A \cup B)} = \dfrac{0.4}{0.8} = 0.5.$

   (c) Händelserna är beroende ty $0.3 = P(A \cap B) \neq P(A)P(B) = 0.28.$

   (d) Låt $X =$ antal dagar av de 150 när man äter mat med snabba kolhydrater $\sim \mathrm{Bin}(150, 0.4)$. Eftersom $npq = 150 \cdot 0.4 \cdot 0.6 = 36$, kan normalapproximation användas: $X \dot\sim \mathrm{N}(150 \cdot 0.4, 150 \cdot 0.4 \cdot 0.6)$=$\mathrm{N}(60, 36)$. Detta ger
   $$\begin{aligned} P(X \geq 55) &= 1 - P(X \leq 54) = 1 - P(\frac{X-60}{\sqrt{36}} \leq \frac{54-60}{\sqrt{36}}) \\ &\approx 1 - \Phi(\frac{54-60+0.5}{\sqrt{36}}) = 1 - \Phi(-0.92) = \Phi(0.92) = 0.8212. \end{aligned}$$

2. Inför $X =$ längden på en låt $\sim \mathrm{N}(6.4, 1.44)$ och $Y =$ längden på en paus $\sim \mathrm{N}(8.5, 3.24)$.

   (a) Sannolikheten att en låt framförs färdigt på 6.7 minuter:
   $$\begin{aligned} P(X \leq 6.7) &= P(\frac{X-6.4}{\sqrt{1.44}} \leq \frac{6.7-6.4}{\sqrt{1.44}}) = \Phi(\frac{6.7-6.4}{\sqrt{1.44}}) \\ &= \Phi(0.25) = 0.5987. \end{aligned}$$

   (b) Låt $Z =$ längden på 9 låtar med 8 pauser emellan. Då gäller $Z = \sum_{i=1}^{9} X_i + \sum_{i=1}^{8} Y_i \sim \mathrm{N}(9 \cdot 6.4 + 8 \cdot 8.5, 9 \cdot 1.4 + 8 \cdot 3.24) = \mathrm{N}(125.6, 38.88)$. Därmed
   $$\begin{aligned} P(Z \leq 131) &= P(\frac{X-125.6}{\sqrt{38.88}} \leq \frac{131-125.6}{\sqrt{38.88}}) = \Phi(\frac{131-125.6}{\sqrt{38.88}}) \\ &= \Phi(0.87) = 0.8078. \end{aligned}$$

   (c) $W =$ antal låtar av de 5 som framförs färdigt på 6.7 minuter $\sim \mathrm{Bin}(5, 0.5987)$. $U =$ antal låtar av de 5 som är längre än 6.7 minuter $\sim \mathrm{Bin}(5, 1-0.5987) = \mathrm{Bin}(5, 0.4013)$. Från Tabell 1:
   $$P(W = 2) = P(U = 3) = P(U \leq 3) - P(U \leq 2) = 0.9130 - 0.6826 = 0.2304.$$

(d) V = längden på 12 låtar med 11 pauser emellan. $V = \sum_{i=1}^{12} X_i + \sum_{i=1}^{11} Y_i \sim \mathrm{N}(12 \cdot 6.4 + 11 \cdot 8.5, 12 \cdot 1.4 + 11 \cdot 3.24) = \mathrm{N}(170.3, 52.92)$. Vi söker efter A sådant att $P(V \leq A) = 0.95$ dvs A är 0.05-kvantilen till fördelningen för V och kan beräknas med hjälp av kvantiler till standardnormalfördelningen:

$$P(\frac{V - 170.3}{\sqrt{52.92}} \leq \frac{A - 170.3}{\sqrt{52.92}}) = 0.95 \quad \Leftrightarrow \quad \frac{A - 170.3}{\sqrt{52.92}} = \lambda_{0.05} = 1.64$$
$$\Leftrightarrow \quad A = 1.64 \cdot \sqrt{52.92} + 170.3 = 182.2304.$$

3. (a) Vi approximerar värden i varje grupp med mittvärdena:
$\bar{x} = \frac{1}{24}(73 \cdot 4 + 75 \cdot 7 + 77 \cdot 9 + 79 \cdot 2 + 81 \cdot 1 + 83 \cdot 1) = \frac{1832}{24} = 76.3333$,
$s^2 = \frac{1}{23}(73^2 \cdot 4 + 75^2 \cdot 7 + 77^2 \cdot 9 + 79^2 \cdot 2 + 81^2 \cdot 1 + 83^2 \cdot 1 - \frac{1}{24}1832^2) = 6.1449$, $s = 2.4789$.

(b) X = fördröjningstiden, $x_1, x_2, \ldots, x_{24}$ ett stickprov från $X \sim \mathrm{N}(\mu, \sigma^2)$, där μ och σ är okända. Punktskatting $\hat{\mu} = \bar{x}$, spridningsmått $D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. σ är okänd, skattas med s. Testvariabeln $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ är $t(n-1)$-fördelad. Ett 95 % konfidensintervall för μ blir

$$I_\mu = (\bar{x} \pm t_{0.025}(23) \frac{s}{\sqrt{24}}) = (76.3333 \pm 2.069 \cdot \frac{2.4789}{\sqrt{24}}) = (75.2864, 77.3803).$$

(c) Vi testar $H_0: \ \mu = 75$ mot $H_1: \ \mu > 75$. Testvariabeln $\frac{\bar{X} - 75}{\frac{s}{\sqrt{n}}}$ är $t(n-1)$-fördelad om H_0 är sann. Test: om $\frac{\bar{x} - 75}{\frac{s}{\sqrt{n}}} > t_{0.05}(23) = 1.714$, förkasta H_0. Vi har $\frac{\bar{x} - 75}{\frac{s}{\sqrt{24}}} = 2.6350 > 1.714$, H_0 förkastas. Med 95 % säkerhet är den genomsnittliga fördröjsningstiden längre än 75 μs.

(d) Eftersom antalet observationer är större än 20, blir $\bar{X}$ enligt centrala gränsvärdessatsen approximativt normalfördelad. Testvariabeln $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ blir i det här fallet approximativt $\mathrm{N}(0, 1)$-fördelad, fast en bättre approximation kan fås med $t(23)$-fördelningen. Därför är det möjligt att behålla samma konfidensintervall som ovan så eller kan $\lambda_{0.025}$-kvantilen användas i stället:

$$I_\mu = (\bar{x} \pm \lambda_{0.025} \frac{s}{\sqrt{24}}) = (76.3333 \pm 1.96 \cdot \frac{2.4789}{\sqrt{24}}) = (75.3416, 77.3251).$$

4. (a) χ^2-test för oberoende. Vi testar H_0 : faktorer "antal köpta skivor" och "en person delar med sig av filerna" är oberoende, dvs samma proportioner förekommer i båda serier. Testvariabeln

$$Q = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(x_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{x_{ij}^2}{\hat{e}_{ij}} - n$$

är $\chi^2((2-1)(3-1)) = \chi^2(2)$-fördelad om H_0 är sann. Test: om $Q > \chi^2_{0.05}(2) = 5.991$, förkasta H_0. I vårt fall

$$\begin{aligned} Q &= \frac{65^2}{\frac{171\cdot 369}{1000}} + \frac{207^2}{\frac{601\cdot 369}{1000}} + \frac{97^2}{\frac{228\cdot 369}{1000}} + \frac{106^2}{\frac{171\cdot 631}{1000}} + \frac{394^2}{\frac{601\cdot 631}{1000}} \\ &\quad + \frac{131^2}{\frac{228\cdot 631}{1000}} - 1000 = 4.7686 < 5.991, \end{aligned}$$

H_0 kan ej förkastas. Vi kan inte påstå att faktorerna är beroende.

(b) χ^2-test för proportioner. Vi testar $H_0: \ p_1 = \frac{1}{5}, p_2 = \frac{3}{5}, p_3 = \frac{1}{5}$ mot H_1 : andra sannolikheter förekommer. Testvariabeln

$$Q = \sum_{i=1}^{3} \frac{(x_i - np_i)^2}{np_i} = \sum_{i=1}^{3} \frac{x_i^2}{np_i} - n$$

är $\chi^2(3-1) = \chi^2(2)$-fördelad om H_0 är sann. Test: om $Q > \chi^2_{0.05}(2) = 5.991$, förkasta H_0. I vårt fall

$$Q = \frac{171^2}{200} + \frac{601^2}{600} + \frac{228^2}{200} - 1000 = 8.1267 > 5.991,$$

H_0 förkastas. Med 95 % säkerhet förekommer andelarna för antal köpta skivor enligt andra proportioner.

5. Ett linjärt samband beskrivs väl med linjär regression. Vi väljer $X =$ förvaringstiden, $Y =$ mängd vitamin C. $x_1, \ldots, x_6$ är givna, $y_1, \ldots, y_6$ är observationer av $Y_i = \alpha + \beta x_i + \epsilon_i$, där ϵ_i är $\mathrm{N}(0, \sigma^2)$-fördelade.

(a) Graden av linjärt samband mäts med korrelationskoefficienten $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$. Vi har $S_{xx} = 74.8333$, $S_{xy} = -572.8333$, $S_{yy} = 4964.833$ och $r = -0.9398$. r ligger nära $-1 \Rightarrow$ ett starkt (negativt) linjärt samband.

(b) Den skattade regressionslinjen: $y = \hat{\alpha} + \hat{\beta}x$, $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ och $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. $\bar{x} = 4.1667$, $\bar{y} = 71.8333$, $\hat{\beta} = -7.6548$, $\hat{\alpha} = 103.7283$. Därmed $y = 103.7283 - 7.6548x$.

(c) Insättning av $y = 0$ i den skattade linjens ekvation ger en skattning $x = \frac{0-103.7283}{-7.6548} = 13.5508$ dagar för tiden när ingen vitamin C finns kvar i frukterna. Man ska dock vara försiktig med en sådan prediktion ty det finns inget vetenskapligt underlag att det linjära sambandet håller även utanför det observerade intervallet. Det är överhuvudtaget tveksamt om en linjär modell är lämplig för att beskriva sådana fenomen även om de några få samlade observationer tyder på det...