UPPSALA UNIVERSITET
Matematiska institutionen
Raimundas Gaigalas

STATISTIK IP
2004-03-16

Lösningar till tenta 040316

1. Inför händelser C, D och E att respektive bro är farbar. Då är C, D och E oberoende och $P(C) = 0.7$, $P(D) = 0.4$, $P(E) = 0.8$.

   (a) Sannolikheten att någon av broarna D och E är farbar:

   $$\begin{aligned} P(D \cup E) &= P(D) + P(E) - P(D \cap E) = P(D) + P(E) - P(D) \cdot P(E) \\ &= 0.4 + 0.8 - 0.4 \cdot 0.8 = 0.88. \end{aligned}$$

   (b) På grund av oberoendet:

   $$P(C \cap (D \cup E)) = P(C) \cdot P(D \cup E) = 0.7 \cdot 0.88 = 0.616.$$

2. (a) $P(X \leq 3) = p(-1) + p(1) + p(3) = 0 + \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$

   (b) $$\begin{aligned} E(X) &= (-1) \cdot p(-1) + 1 \cdot p(1) + 3 \cdot p(3) + 5 \cdot p(5) \\ &= (-1) \cdot 0 + 1 \cdot \frac{1}{6} + 3 \cdot \frac{1}{3} + 5 \cdot \frac{1}{2} = \frac{11}{3}, \\ E(X^2) &= (-1)^2 \cdot p(-1) + 1^2 \cdot p(1) + 3^2 \cdot p(3) + 5^2 \cdot p(5) \\ &= (-1)^2 \cdot 0 + 1^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{3} + 5^2 \cdot \frac{1}{2} = \frac{47}{3}, \\ \mathrm{Var}(X) &= E(X^2) - (E(X))^2 = \frac{47}{3} - \left(\frac{11}{3}\right)^2 = \frac{20}{9}, \\ D(X) &= \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{20}{9}} = 1.4907. \end{aligned}$$

   (c) $$P(X = -1 \mid X \leq 3) = \frac{P(X = -1, X \leq 3)}{P(X \leq 3)} = \frac{P(X = -1)}{P(X \leq 3)} = 0.$$

3. Inför X = antal bilar i ett hushåll. Då gäller att $E(X) = 0.7$, $D(X) = 0.3$. Vi vill studera Y = totala antalet bilar i området. Men $Y = \sum_{i=1}^{300} X_i$ och enligt centrala gränsvärdessatsen $Y \overset{\cdot}{\sim} \mathrm{N}(300 \cdot 0.7, 300 \cdot 0.09) = \mathrm{N}(210, 27)$. Sannolikheten att alla bilar i området får plats på 215 parkeringsplatser:

   $$\begin{aligned} P(Y \leq 215) &= P\left(\frac{Y - 210}{\sqrt{27}} \leq \frac{215 - 210}{\sqrt{27}}\right) \approx \Phi\left(\frac{215 - 210 + \frac{1}{2}}{\sqrt{27}}\right) \\ &= \Phi(1.06) = 0.8554. \end{aligned}$$

   Anm: halvkorrektion behövs för vi approximerar en diskret fördelning med en kontinuerlig, men det är ok att räkna utan halvkorrektion. I så fall fås $\Phi(0.96) = 0.8315$.

4. (a) X = antal detaljer som kasseras under en timme $\sim \mathrm{Po}(3.5)$. Y = antal detaljer som kasseras under två timmar $\sim \mathrm{Po}(2 \cdot 3.5) = \mathrm{Po}(7)$. Den sökta sannolikheten:

   $$P(Y > 9) = 1 - P(Y \leq 9) = 1 - 0.8305 = 0.1695.$$

   (b) Z = antal detaljer som kasseras under åtta timmar $\sim \mathrm{Po}(8 \cdot 3.5) = \mathrm{Po}(28)$. Eftersom $\lambda = 28 > 15$, kan normalapproximation tillämpas: $X \overset{\cdot}{\sim} \mathrm{N}(28, 28)$. Därmed

   $$\begin{aligned} P(Z \leq 29) &= P\left(\frac{Y - 28}{\sqrt{28}} \leq \frac{29 - 28}{\sqrt{28}}\right) \approx \Phi\left(\frac{29 - 28 + \frac{1}{2}}{\sqrt{28}}\right) \\ &= \Phi(0.19) = 0.5753. \end{aligned}$$

5. Homogenitetstest. Märkligt nog, procentsatserna i tidningen adderar inte till 100 %. Antalet personer som deltog i undersökningen i april är inte angivet heller så vi antar att det var samma antal som i maj dvs 1213. Vi ska undersöka de 99 motsvarande 98 % av svaren som finns tillgängliga:

	Ja	Nej	Tveksamma	Summor
april	412	607	182	1201
maj	449	582	158	1189
	861	1189	340	2390

Vi testar H_0 : de två serierna är homogena, 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

$$\begin{aligned} Q &= \frac{412^2}{\frac{861\cdot1201}{2390}} + \frac{607^2}{\frac{1189\cdot1201}{2390}} + \frac{182^2}{\frac{340\cdot1201}{2390}} + \frac{449^2}{\frac{861\cdot1189}{2390}} + \frac{582^2}{\frac{1189\cdot1189}{2390}} \\ &\quad + \frac{158^2}{\frac{340\cdot1189}{2390}} - 2390 = 3.7496 < 5.991, \end{aligned}$$

H_0 kan ej förkastas. Rubriken "Ökat stöd till EMU" är inte statistiskt fastställd.

6. X = längd på låtar på *Diamond dogs* $\sim \mathrm{N}(\mu_1, \sigma^2)$, Y = längd på låtar på *Station to station* $\sim \mathrm{N}(\mu_2, \sigma^2)$. $x_1, \ldots, x_{11}$ observationer av X, $y_1, \ldots, y_6$ observationer av Y. Vi testar $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 < \mu_2$.

   Punktskattning $\widehat{\mu_1 - \mu_2} = \bar{x} - \bar{y}$. Spridningsmått $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$. σ^2 är okänd, skattas med den sammanvägda skattningen $s^2_{(2)} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$. Testvariabeln $\dfrac{\bar{X} - \bar{Y}}{s_{(2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ är $t(n_1 + n_2 - 2)$-fördelad om H_0 är sann.

   Test: om $\frac{\bar{x}-\bar{y}}{s_{(2)}\sqrt{\frac{1}{11}+\frac{1}{6}}} < -t_{0.05}(15) = -1.753$, förkasta H_0. I vårt fall $\bar{x} = 207.909$, $\bar{y} = 377.6667$, $s_1^2 = 6878.891$, $s_2^2 = 14869.87$, $s_{(2)} = 97.68597$, $\frac{\bar{x}-\bar{y}}{s_{(2)}\sqrt{\frac{1}{11}+\frac{1}{6}}} = -3.4241 < -1.753$, H_0 förkastas. Med 95 % säkerhet är låtarna på *Diamond dogs* kortare.

7. Linjär regressionsmodell. Vi väljer X = markdjup, Y = temperatur. $x_1, \ldots, x_7$ är givna, $y_1, \ldots, y_7$ är observationer av $Y_i = \alpha + \beta x_i + \epsilon_i$, där ϵ_i är $\mathrm{N}(0, \sigma^2)$-fördelade.

   (a) Den skattade regressionslinjen: $y = \hat{\alpha} + \hat{\beta}x$, $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ och $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. $S_{xy} = 9723.029$, $S_{xx} = 132661.4$, $\bar{x} = 213.2857$, $\bar{y} = 38.28571$, $\hat{\beta} = 0.0733$, $\hat{\alpha} = 22.6536$. Därmed $y = 22.6536 + 0.0733x$.

   (b) Korrelationskoefficienten $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$. $S_{yy} = 770.5486$ medför $r = 0.9617$. Ett starkt linjärt samband.

   (c) Insättning av $x = 500$ i den skattade linjens ekvation ger en skattning $y = 22.6536 + 0.0733 \cdot 500 = 59.2996$ för temperaturen.

8. (a) X = pH-värdet, $x_1, \ldots, x_{30}$ observationer av X. Punktskatting $\hat{\mu} = \bar{x}$. Eftersom antalet observationer är ganska stort (större än 20), blir $\bar{X}$ enligt centrala gränsvärdessatsen approximativt normalfördelad. Spridningsmått $D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. σ är okänd, skattas med s. Testvariabeln $\dfrac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$ är N$(0, 1)$-fördelad.

Ett 95 % konfidensintervall för μ blir

$$I_\mu = (\bar{x} \pm z_{0.025} \frac{s}{\sqrt{30}}) = (8.0733 \pm 1.96 \cdot \frac{0.23479}{\sqrt{30}}) = (7.9893, 8.1573).$$

Det är också möjligt att använda $t_{0.025}(29)$-kvantilen i stället. I det fallet

$$I_\mu = (\bar{x} \pm t_{0.025}(29) \frac{s}{\sqrt{30}}) = (8.0733 \pm 2.045 \cdot \frac{0.23479}{\sqrt{30}}) = (7.9856, 8.1610).$$

(b) Konfidensgraden 95 % är sannolikheten att ett konfindensintervall täcker det riktiga värdet. Därmed blir antalet konfidensintervall som missar det riktiga pH-värdet bland de 5 erhållna Bin$(5, 0.05)$-fördelat.