TENTAMEN: Matematisk statistik för K (24 maj, 2006)

Kortfattade lösningar:

1) a) Händelse B givet att händelse A har inträffat.

   b) $P(A \cap B) = P(B|A)P(A) = 0.12$ är inte lika med 0 och därför är A och B inte disjunkta händelser.

   c) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.78$

   d) $P(A \cap B) = 0.12 \neq 0.4 \cdot 0.5 = 0.20 = P(A)P(B)$ och därför är A och B inte oberoende.

2) a) Sannolikheten att teststatistikan får det observerade värdet eller även mer extremt värde då H_0 är sann.

   b) Signifikansnivån α är sannolikheten att man förkastar H_0 då den är sann.

3) $IQ \sim N(115, 12)$. Då är $P(IQ < 95) = P((IQ - 115)/12 < (95 - 115)/12) = P(Z < -1.67) = \Phi(-1.67) = 0.0475$, där $Z \sim N(0,1)$ och Φ är fördelningsfunktionen av den standardiserade normalfördelningen. Då är det 28 (eller 29) elever som inte kommer in.

4) a) Nollhypotesen är $H_0 : \mu = 1.8$ och mothypotesen $H_1 : \mu \neq 18$. T-testet därför att variansen är okänd. Man måste anta att observationerna kommer från en normalfördelning. Teststatistikan är $(\bar{X} - 1.8)/(S/\sqrt{n})$ har T_{10}-fördelning och får värdet -1.64. $t_{.025} = 2.228$ och då är värdet $-1.64 (> -2.228)$ inte på kritiska området. H_0 accepteras på signifikansnivå 0.05 och man kan säga att den genomsnittliga tiden häcklipparen fungerar skiljer inte sig signifikant från 1.8 timmar.

   b) Nollhypotesen är $H_0 : M = 1.8$ och mothypotesen $H_1 : M \neq 18$, där M är medianen. Tiden är en kontinuerlig stokastisk variabel och man kan använda teckentestet. Inga antaganden behövs. Antalet positiva skillnader $(X_i - 1.8)$, Q_+ är 3. Vi har en nolla i datan och den tas bort så att man har 10 observationer kvar. Då kan man räkna p-värdet, som är $2P(Q_+ \leq 3 | Q_+ \sim Bin(10, 0.5)) = 2(\frac{1}{2})^{10} \sum_{x=0}^{3} \binom{10}{x} = .343$
   Nu är p-värdet större än signifikansnivån 0.05 och därför kan man inte förkasta H_0, dvs. att den genomsnittliga tiden häcklipparen fungerar skiljer inte sig signifikant från 1.8 timmar.

   c) Samma resultat i både fall, dvs att observationerna kan antas komma från en normalfördelning.

5) a) Konfidensintervallet är $\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$, där $S_p^2 = ((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)/(n_1 + n_2 - 2)$. Nu är $\bar{x}_1 = 6150$, $\bar{x}_2 = 5250$, $\alpha = 0.01$, $t_{0.005}^{(28)} = 2.763$, $n_1 = 16$, $n_2 = 14$ och $s_p^2 = 6040.179$, och ett 99% konfidensintervall blir $900 \pm 79 = [821, 979]$ och det verkar som det är bra att använda kedjor (0 är inte på intervallet).

b) Man testar hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 > \mu_2$. Teststatistikans värde är $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_p^2(1/n_1 + 1/n_2)} = 31.64$, vilket är större än $t_{0.01}^{28} = 2.467$. Därför förkastas H_0 på signifikansnivån 0.01 och man kan säga att det verkar bra att använda kedjor.

c) $X_{1i} \sim N(\mu_1, \sigma)$ och $X_{2i} \sim N(\mu_2, \sigma)$ och att de två stickproven är oberoende.

6) a) Nollhypotesen är $H_0 : \mu = 3$ och mothypotesen $H_1 : \mu < 3$.

b) Man skulle börja marknadsföra den nya produkten även om den inte var bättre än den gamla (typ I fel). Man skulle låta bli att börja marknadsföra den nya produkten även om den var bättre än den gamla (typ II fel).

c) $P(H_0 \text{ förkastas} | \mu = 2.5) = P((\bar{X} - 3)\sqrt{n}/0.5 < -1.645 | \mu = 2.5) = P((\bar{X} - 2.5)\sqrt{n}/0.5 < \sqrt{n} - 1.645) = P(Z < \sqrt{n} - 1.645) = \Phi(\sqrt{n} - 1.645) = 0.95$, där $Z \sim N(0, 1)$ och Φ är fördelningsfunktionen av den standardiserade normalfördelningen. Då är $\sqrt{n} - 1.645 = 1.645$ och $n = 10.8$. Det behövs 11 mätningar.

d) $\bar{x} = 2.34$ och $s = 0.71$. Teststatistikan $(\bar{X} - 3)/(S/\sqrt{(n)})$ är T_{15}-fördelad och får värdet -3.718, vilket är mindre än $-t_{.05} = -1.753$. H_0 förkastas på signifikansnivå 5% och man kan säga att den nya produkten verkar ha kortade torkningstid än den gamla.

e) $X_i \sim N(\mu, \sigma)$

7) a) Man får att $\hat{\mu}_{Y|x} = 0.8233 - 0.0589x$

b) Vänster: scatterplot där strömhastigheten är på x-axeln och återhämtningskvoten på y-axeln. Höger: residualplot där strömhastigheten är på x-axeln och skillnaden mellan den observerade återhämtningskvoten och dess uppskattade värde på y-axeln. Återhämtningskvoten verkar bero linjärt på strömhastigheten; den miskar då strömhastigheten ökar. Residualerna verkar vara genomsnitligt 0 och variansen konstant. Observation 9 kan tolkas som "outlier". En enkel linjär regressionsmodel verkar ok.

c) Om linjär model är lämplig, om väntevärdet av felet inte är 0, om variansen inte är konstant, om det finns hål i datamängden, och om det finns "outliers".