TENTAMEN: Matematisk statistik för K (24 augusti, 2007)

Kortfattade lösningar:

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

b) Två händelser är disjunkta om de inte kan inträffa samtidigt, dvs. att sannolikheten av snitthändelsen är noll. Nu är $P(A \cap B) = P(B|A)P(A) = 0.12 \neq 0$ och då är A och B inte disjunkta.

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

d) $P(A \cap B) = 0.12 \neq 0.4 \cdot 0.5 = P(A)P(B)$.

2) a) Väntevärdesriktighet och att variansen minskar då stickprovsstorleken ökar.

b) Ja. $\mathbf{E}[\bar{X}] = \mu$ så att det är en väntevärdesriktig skattning för μ och $Var(\bar{X}) = \sigma^2/n$, vilket blir mindre då n blir större. (Här $\mathbf{E}[X] = \mu$ och $Var(X) = \sigma^2$.)

3) a) Testa hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 \neq \mu_2$ på signifikansnivån 0.05. Man får att $S_p^2 = 153.64$ och $T = (\bar{X}_1 - \bar{X}_2)/S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sim T_{64}$ och får ett värde 0.164. Nu är $t = 0.164 < 1.998 = t_{.05}$ (och $0.164 > -1.998$) så att H_0 kan inte förkastas på signifikansnivån 0.05. Dvs. att poängtalen för pojkar och flickor är inte olika. (Du kan också räkna p-värdet om du vill.)

b) $\bar{x}_1 - \bar{x}_2 \pm t_{0.025}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.5 \pm 6.1$. Därför att $0 \in [-5.6, 6.6]$ kan man säga att det finns ingen skillnad mellan poängtal för pojkar och flickor enligt 95% konfidensintervallet. Man kan dra samma slutsats än i a).

c) Man måste anta att de två populationerna har samma varians. Man behöver inte anta normalfördelning därför att stickprovsstorlekerna är tillräckligt stora.

4) a) $P(X \leq 155) = P(\frac{X-150}{5} \leq \frac{155-150}{5}) = P(Z \leq 1) = \Phi(1) = 0.8413$, där $Z \sim N(0,1)$ och Φ är fördelningsfunktionen av den standardiserade normalfördelningen.

b) Låt Y vara antalet apelsiner (bland 5) som väger mer än 155g, $Y \sim Bin(5, 1 - 0.8413)$. Då är $P(Y \geq 1) = 1 - P(Y = 0) = 0.5785$.

5) $\sigma_A = \sigma_B = 1$

a) Man testar $H_0 : \mu_A = \mu_B$ mot $H_1 : \mu_A \neq \mu_B$ genom att använda teststatistikan $Z = (\bar{X}_A - \bar{X}_B)/\sigma$, där $\bar{X}_A \sim N(\mu_A, \sqrt{\frac{1}{10}})$, $\bar{X}_B \sim N(\mu_B, \sqrt{\frac{1}{12}})$, $\bar{X}_A - \bar{X}_B \sim N(\mu_A - \mu_B, \sqrt{\frac{1}{10} + \frac{1}{12}})$ och $\sigma = \sqrt{\frac{1}{10} + \frac{1}{12}}$. Då är teststatistikan $Z \sim N(0,1)$ och den får värdet 2.009. Då är $P(Z > 2.009 | H_0 \text{ sann}) = 1 - P(Z \leq 2.009 | \mu = 0) = 1 - \Phi(2.009) = 1 - 0.9778 = 0.0222$, där Φ är fördelningsfunktionen av den standardiserade normalfördelningen. Det exakta dubbelsidiga p-värdet är $2 \cdot 0.0222 = 0.044$.

b) $X_A \sim N(\mu_A, 1)$ och $X_B \sim N(\mu_B, 1)$.

6) a) Man testar $H_0 : \mu = 68$ mot $H_1 : \mu > 68$, där μ är väntevärdet av medelhöjden som antas vara normalfördelad (med okänd varians). Ett stickprovs T-test: Teststatistikan $T = (\bar{X} - \mu)/(S/\sqrt{n})$ är T_{19}-fördelad och får värdet 3.39. Värdet jämförs med $t_{0.01} = 2.539$ från T_{19}-fördelningen. Man förkastar H_0 på signifikansnivån 0.01 därför att teststatistikans värde 3.39 är större än 2.539. Man kan säga att datan stöder påståendet att medelhöjden är större än 68 tum.

b) Man använder teckentestet och testar $H_0 : M = 68$ mot $H_1 : M > 68$, där M är medianhöjden. Antalet positiva skillnader $(X_i - 68)$, Q_+ är 14. Vi har två nollor i datan och de har ändrats till minustecken. Då kan man räkna p-värdet $P(Q_+ \geq 14 | Q_+ \sim Bin(20, 0.5))$ är $(\frac{1}{2})^{20} \sum_{x=14}^{20} \binom{20}{x} = 0.058$. Nu är p-värdet större än signifikansnivån 0.01 och därför kan man inte förkasta H_0, dvs. att datan stöder inte påståendet att medelhöjden är större än 68 tum.

c) Höjden kan inte antas vara normalfördelad att därför ger testet baserat på normalfördelningsantagandet inte samma resultat som det ickeparametriska testet. Man borde lita på det ickeparametriska testet.

7) a) Man får att $\hat{\mu}_{Y|x_i} = 6.25 - 0.00231x_i$

b) Figuren på vänster är en "scatterplot" där man har plottat (x_i, y_i)-värdena. Figuren på höger är en residualplot, dvs. $(x_i, y_i - (6.25 - 0.00231x_i))$. Det verkar som värdena av söthetsindexet minskar då pektinmängden ökar (se figuren till vänster) men punkterna är väldigt splittrade. Från residualplotten ser man att väntevärdet av $E_i \approx 0$ men variansen verkar inte vara konstant. Man har också ett hål i x-värdena. En linjär regressionsmodell verkar inte vara så bra.