TENTAMEN: Matematisk statistik för K (25 maj, 2005)

Kortfattade lösningar:

1) a) $P(E \cap F) = P(F|E)P(E) = 0.20 \cdot 0.55 = 0.11$

   b) $P(E \cup F) = P(E) + P(F) - P(E \cap F) = 0.55 + 0.40 - 0.11 = 0.84$

   c) $P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{0.11}{0.40} = 0.275$

2) a) Ett 99% konfidensintervall är $\bar{x} \pm t_{\alpha/2} s/\sqrt{n}$, där $\bar{x} = 2.89$, $s = 3.92$, $\alpha = 0.01$, $n = 10$ och $t_{0.005} = 3.25$ (antalet frihetsgrader är $10-1=9$). Då blir konfidensintervallet $2.89 \pm 3.25 \cdot 3.92/\sqrt{10} = 2.89 \pm 4.03$, dvs. $[-1.14, 6.92]$.

   b) Enligt a) är den sanna medelnivån av bly i dricksvattnet mellan -1.14 och 6.92 μg/l. Blynnivån kan inte vara negativ så att man borde säga att medelnivån är lägre än 6.92 μg/l men större än 0.

   c) 99% av stickproven av storlek 10 skulle ge ett konfidensintervall som täcker det sanna medelblynivån (man är 99% konfident att medelnivån av bly ligger på det beräknade intervallet). Bara en gång i 100 skulle man få ett intervall som inte innehåller det sanna värdet.

   d) Stickprovet kommer från en normalfördelning.

3) a) Vikten $X \sim N(1.03, 0.014)$. Då är $P(X < 1) = P(\frac{X-1.03}{0.014} < \frac{1-1.03}{0.014}) = P(Z < -2.14) = 0.0162$, där $Z \sim N(0,1)$.

   b) Nu är vikten $X \sim N(1.05, 0.016)$ och $P(X < 1) = P(\frac{X-1.05}{0.016} < \frac{1-1.05}{0.016} = P(Z < -3.13) = 0.0009$. Detta betyder att andelen underviktiga paket minskar.

4) Låt X vara vithetsnivån med låg nivå och Y med hög nivå av väteperoxid. Väntevärdena och varianserna av X och Y är μ_X resp. μ_Y och σ_X^2 resp. σ_Y^2

   a) För att testa hypotesen $H_0 : \mu_X = \mu_Y$ mot $H_1 : \mu_X \neq \mu_Y$ kan man använda två stickprovs T-testet (σ okänd). Teststatistikan är $T = (\bar{X} - \bar{Y})/(S\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}) \sim T_{n_X+n_Y-2}$, där $S^2 = ((n_X - 1)S_X^2 + (n_Y - 1)S_Y^2)/(n_X + n_Y - 2)$. Nu är $s^2 = 14.44$ och $t = -1.50$. $\alpha = 0.05$ och därför jämförs värdet av teststatistikan med $t_{0.025} = 2.035$ (eller $-t_{0.025} = -2.035$). Antalet frihetsgrader är $n_X + n_Y - 2 = 33$. Värdet av teststatistikan är inte på kritiska området ($-2.035 < -1.50 < 2.035$) och därför förkastar man inte H_0. Man kan inte säga att det finns någon skillnad mellan de olika vithetsnivåerna.

   b) $X \sim N(\mu_X, \sigma)$ och $Y \sim N(\mu_Y, \sigma)$. Varianserna kan antas vara lika men Y verkar inte vara normalfördelad.

5) a) Man har två stokastiska variabler X och Y och man är intresserad av sambandet mellan dem. Y är stokastisk och X är mätt utan fel. Man plottar först grafen Y mot x och om sambandet ser linjärt ut, kan en linjär regressionsmodel vara en lämplig model.

b) Man antar att felet $E_i \sim N(0, \sigma)$. (Då är också Y normalfödelad). Då vet man att $(\hat{\beta}_1 - \beta_1)/(S/\sqrt{\sum(x_i - \bar{x})^2}) \sim T_{n-2}$ och att $P(-t_{0.05} \leq (\hat{\beta}_1 - \beta_1)/(S/\sqrt{\sum(x_i - \bar{x})^2}) \leq t_{0.05}) = 0.90$, där $P(T > t_{0.05}) = 0.05$, $T \sim T_{n-2}$ och n är stickprovsstorleken. Genom att lösa olikheten ovan m.a.p. β_1 får man att ett 90% konfidensintervall för β_1 är $\hat{\beta}_1 \pm t_{0.05}S/\sqrt{\sum(x_i - \bar{x})^2}$, där $\hat{\beta}_1 = (n\sum x_iY_i - \sum x_i \sum Y_i)/(n\sum x_i^2 - (\sum x_i)^2)$ $S^2 = \sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/(n-2)$ och $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$.

6) De två stickproven är inte oberoende och därför utför man ett parvist test (som är samma som ett stickprovs T-test) för att testa $H_0 : \mu_D = \mu_A - \mu_B = 0$ mot $H_1 : \mu_D \neq 0$. Man definierar $D_i = A_i - B_i$ och får att $\bar{d} = -2.906$ och $s_d = 8.895$. Teststatistikan (som är T_{31}-fördelad) är $\bar{D}/(S_d/\sqrt{n})$ och dess värde är $-2.906/(8.895/\sqrt{32}) = -1.848$. Värdet jämförs med $t_{0.025} = 2.040$ (och $-t_{0.025} = -2.040$), antal frihetsgrader är $n - 1 = 31$. Teststatistikans värde är inte på kritiska området och vi kan inte förkasta H_0. Man kan inte säga att det finns någon skillnad mellan IQ-poängtal av tvillingspar beroende på vem har uppfostrat dem. (Antalet observationer är tillräckligt stort och vi kan anta att $\bar{D}$ är approximativt normalfördelad.)

7) 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 icke-parametriska testet. Man borde lita på det ickeparametriska testet.