TENTAMEN: Matematisk statistik för K (27 maj, 2011)

Kortfattade lösningar:

1) Låt A vara händelsen att den första man tar är felaktig och B händelsen att den andra man tar är felaktig

   a) $P(A) = \frac{3}{10}$
   b) $P(B|A) = \frac{2}{9}$
   c) $P(B) = P(B \cap A) + P(B \cap A') = P(B|A)P(A) + P(B|A')P(A') = \frac{2}{9} \cdot \frac{3}{10} + \frac{3}{9} \cdot \frac{7}{10} = \frac{3}{10}$
   d) $P(B) \neq P(B|A)$ dvs. att A och B är inte oberoende.

2) a) $\int_{-\infty}^{\infty} f(x)\,dx = c\int_0^1 (2x - x^2)dx = 1$ vilket ger att $c = \frac{3}{2}$.
   b) $F(x) = 0$, då $x < 0$, $F(x) = \frac{3}{2}\int_0^x (2t - t^2)dt = \frac{3}{2}x^2 - \frac{1}{2}x^3$, då $0 \leq x \leq 1$, och $F(x) = 1$, då $x > 1$.
   c) $\mathbf{E}[X] = \int_{-\infty}^{\infty} xf(x)\,dx = \frac{3}{2}\int_0^1 x(2x - x^2)dx = \frac{5}{8}$.
   d) $P(0.5 < X \leq 1.2) = P(0.5 < X \leq 1) = \int_{0.5}^1 f(x)\,dx = \frac{11}{16}$

3) $n = 100$, $\bar{x} = 50$, $\sigma = 2$

   a) $\bar{x} \pm z_{0.025}\sigma/\sqrt{n} = 50 \pm 0.39$.
   b) Längden av intervallet är $2z_{\alpha/2} \cdot 2/10 = 50.3 - 49.7 = 0.6$, vilket ger att $z_{\alpha/2} = 1.5$. $1 - \Phi(1.5) = 1 - 0.9332 = 0.0668$ och $\alpha = 2 \cdot 0.0668 = 0.1336$. Konfidensgraden är $100(1 - 0.1336)\% = 86.64\%$.
   c) Inga.

4) Utför ett teckentest för medianen M genom att testa $H_0 : M = 30$ mot $H_1 : M > 30$.

   a) $q_+ = 16$, $\alpha = 0.01$ och p-värdet blir $P(Q_+ \geq 16 | Q_+ \sim Bin(20, 0.5)) = \sum_{x=16}^{20} \binom{20}{x} (\frac{1}{2})^x (1 - \frac{1}{2})^{20-x} = 0.0059$. Nu är p-värdet mindre än $\alpha = 0.01$ och H_0 kan förkastas på signifikansnivå 0.01. Det verkar som medianåldern är högre än 30 år.
   b) Åldern är en kontinuerligt fördelad stokastisk variabel och det inte finns personer som har disputerad exakt vid 30.
   c) 0.0059

5) a) $H_0 : \mu = 4.0$ mot $H_1 : \mu \neq 4.0$.

b) Typ I: man skulle säga att den genomsnittliga väggtjockleken skilljer sig från 4.0 även om den inte gör. Typ II: man skulle säga att den genomsnittliga väggtjockleken är 4.0 även om den inte är.

c) Teststatistikan $T = \frac{\bar{X}-4.0}{S/\sqrt{n}}$, som är T_6-fördelad om H_0 är sann. Det observerade värdet för T är 1.81 och $t_{0.025}^{(6)} = 2.447$. Teststatistikans värde är inte på kritiska området och man kan inte förkasta H_0 på signifikansnivån 0.05. Man kan inte säga att den genomsnittliga tjockleken skilljer sig från 4.0.

d) $X_1, ..., X_7$ är ett stickprov från $N(\mu, \sigma)$. Histogrammet visar att fördelningen ser inte riktigt normal ut.

e) Typ II

6) a) $H_0 : \mu_X = \mu_Y$ mot $H_1 : \mu_X < \mu_Y$

b) 2 stickprovs T-test, två oberoende stickprov.

c) Teststatistikan $T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{1/n_X+1/n_Y}}$ som är T_{10}-fördelad om H_0 är sann, och $S_p^2 = \frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}$. Nu är $s_p^2 = 8.56$ och teststatistikan får värdet -0.89. Kritiska värdet är $-t_{0.10} = -1.372$ vilket är mindre än teststatistikans värde. H_0 kan inte förkastas på signifikansnivån 0.10. Man kan inte säga att den nya metoden är billigare än den gamla.

d) $X_1, ..., X_6$ ett stickprov från $N(\mu_X, \sigma)$ och $Y_1, ..., Y_6$ ett stickprov från $N(\mu_Y, \sigma)$ (och att stickproven är oberoende).

e) $P(H_0 \text{ förkastas} | H_1 \text{ sann}) = P(\frac{\bar{X}-\bar{Y}}{S_p\sqrt{1/6+1/6}} \leq -1.372 | \mu_X - \mu_Y = 2)$

7) a) $y = -2090.9 + 737.1x$

b) $e_i = y_i - \hat{y}_i = y_i + 2090.9 - 737.1x_i$

c) Sambandet mellan pH och skörd ser linjärt ut och på det sättet skulle en linjär modell vara bra. Men det finns ett värde som har väldigt stor residual jämfört med de andra, och det påverkar skattningen av regressionslinjen.

d) Skörden skulle öka med $737.1 \cdot 0.1 = 73.71$ pounds per acre.