TENTAMEN: Matematisk statistik för K (25 maj, 2012)

Kortfattade lösningar:

1) Låt A vara händelsen att den första aktiveringsapparaten fungerar och B händelsen att den andra fungerar. Man vet att $P(A) = 0.9$ och $P(B) = 0.8$ och att A och B är oberoende.

   a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.9 + 0.8 - 0.9 \cdot 0.8 = 0.98$

   b) $1 - P(A \cup B) = 0.02$

   c) $P(A \setminus B) = P(A) - P(A \cap B) = 0.9 - 0.72 = 0.18$

2) a) $P(0.5 \leq X \leq 1.5) = F(1.5) - F(0.5) = \frac{3}{4} \cdot \frac{3}{2} - \frac{1}{2} - \frac{1}{4} \cdot \frac{1}{8} = \frac{19}{32}$

   b) Sannolikheten $P(X > 1.5) = 1 - P(X \leq 1.5) = 1 - F(1.5) = 1 - \frac{3}{4} \cdot \frac{3}{2} + \frac{1}{2} = \frac{3}{8}$. Låt Y vara antalet variabler (som har samma fördelning som X och) som antar ett värde som är större än 1.5. Då är $Y \sim Bin(10, \frac{3}{8})$ och $P(Y = 2) = \begin{pmatrix} 10 \\ 2 \end{pmatrix} \left(\frac{3}{8}\right)^2 \left(1 - \frac{3}{8}\right)^{10-2} = 0.147$

3) a) Man har att $X_i \sim N(\mu, \sigma)$, $i = 1, ..., n$, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, och $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T_{n-1}$ där $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$. Man har att
   $P(-t_{0.05}^{(n-1)} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{0.05}^{(n-1)}) = P(-t_{0.05}^{(n-1)} S/\sqrt{n} \leq \bar{X} - \mu \leq t_{0.05}^{(n-1)} S/\sqrt{n})$
   $= P(\bar{X} - t_{0.05}^{(n-1)} S/\sqrt{n} \leq \mu \leq \bar{X} + t_{0.05}^{(n-1)} S/\sqrt{n}) = 0.90$
   Dvs att $\bar{X} \pm t_{0.05}^{(n-1)} S/\sqrt{n}$ är ett 90% konfidensintervall för μ.

   b) $\bar{x} = 7.3$, $s = 13.9$ och $n = 15$. Då är $\bar{x} \pm t_{0.05}^{(14)} s/\sqrt{15} = 37.3 \pm 6.3$ ett 90% konfidensintervall för μ. Hormonnivån antas vara normalfördelad.

   c) Hormonnivån under natten verkar skillja sig från hormonnivån under dagen. Värdet 25.5 ligger inte på det 90% konfidensintervallet i b).

4) a) Man minimerar kvadratsumman av residualerna $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$, där (x_i, y_i), $i = 1, ..., n$, är observationerna.

   b) Inga. Om man vill gå vidare och hitta konfidensintervall för skattningarna eller testa hypoteser angående dem, måste man anta att slumpfelen E_i är oberoende och $N(0, \sigma^2)$ fördelade.

   c) För att kolla om den linjära modellen är bra, om det finns hål i data eller outliers, och att $\mathbf{E}[E_i] = 0$ och $Var(E_i)$ är konstant

5) $\bar{x} = 54.11$, $s^2 = 207.19$ och $n = 20$

a) Man vill testa $H_0 : \mu = 50$ mot $H_1 : \mu < 50$, där μ är den genomsnittliga väntetiden. Man vet inte variansen men har uppskattad den av stickprovsvariansen S^2 och använder T testet. Teststatistikan blir

$$\frac{\bar{x} - 50}{s/\sqrt{20}} = \frac{54.11 - 50}{\sqrt{207.19/20}} = 1.277.$$

Man jämför detta värde med kritiska värdet från T_{19} fördelningen, nmligen $-t_{0.05}^{(19)} = -1.729$. H_0 kan inte förkastas på signifikansnivån 0.05 och man kan inte säga att den genomsnittliga väntetiden skulle vara högst 50s.

b) Man vill testa $H_0 : M = 50$ mot $H_1 : M < 50$, där M är medianväntetiden. Låt Q_+ vara antalet tider (bland de 20) som är längre än 50. Då är $Q_+ \sim Bin(20, 0.5)$ om H_0 är sann. Man räknar p-värdet $P(Q_+ \leq 6|H_0 \text{ sann}) = \sum_{x=0}^{6} \binom{20}{x} \left(\frac{1}{2}\right)^{20} = 0.058$. Man kan inte förkasta H_0 på signifikansnivån 0.05.

c) Både test säger att H_0 inte kan förkastas på signifikansnivån 0.05. I a) är det observerade medelvärdet större än 50 (några stora värden påverkan medelvärdet) och kan därför inte förkasta H_0 som säger att det genomsnittliga värdet skulle vara mindre än 50. I b) är man väldigt nära att förkasta H_0 på signifikansnivån 0.05 men har inte riktigt tillräckligt mycket bevis för att göra det.

d) I a) antar man att $X \sim N(\mu, \sigma)$ och i b) att X är en kontinuerlig stokastisk variabel.

e) Teckentestet i b) därför att enligt histogrammet är X inte normalfördelad.

6) a) Ett parvist test. Man kör två provkörningar med en bil och då blir de två körningarna beroende.

b) Man testar $H_0 : \mu_D = 0$ mot $H_1 : \mu_D > 0$, där μ_D är väntevärdet för $D = X_1 - X_2$. Man antar att $D_i \sim N(\mu, \sigma)$, $i = 1, ..., 10$ och använder T testet. Teststatistikan $\frac{\bar{D}}{S_D/\sqrt{n}}$, där S_D standardavvikelsen av D, är T_{n-1}-fördelad om H_0 är sann, och får värdet 9.15. Värdet är mycket större än det kritiska värdet $t_{0.01}^{(9)} = 2.821$ och man kan förkasta H_0 på signifikansnivån 0.01. Den nya däcktypen verkar minska bränslekonsumption.

c) Ja. Om man kan förkasta H_0 på signifikansnivån 0.01, kan man förkasta på alla signifikansnivåer α, där $\alpha > 0.01$.

7) $n_1 = n_2 = 100$, $\bar{x}_1 = 34.1$, $\bar{x}_2 = 36.0$, $s_1 = 5.9$, och $s_2 = 6.0$

a) Man testar hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 \neq \mu_2$ och använder teststatistikan $T = \frac{\bar{X}_1 - \bar{X}_2}{S_P\sqrt{1/n_1+1/n_2}}$, där $S_P^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1-n_2+2}$. (Teststatistikan får värdet -2.26.) Därför att stickprovsstorlekarna är så stora (100 vardera), är $T \approx N(0,1)$ och man kan räkna p-värdet genom att räkna $2 \cdot P(Z \leq -2.26 | H_0 \text{ sann}) = 2 \cdot 0.0119 = 0.0238$, där $Z \sim N(0,1)$. Därför att p-värdet 0.0238 är mindre än signifikansnivån 0.05, kan man förkasta H_0 på denna signifikansnivå. Blynivåerna verkar skillja sig från varandra.

b) $\bar{x}_1 + \bar{x}_2 \pm z_{0.025} \cdot s_P\sqrt{1/n_1 + 1/n_2} = -1.9 \pm 1.65$. Blynivåerna verkar skillja sig från varandra (0 är inte med på intervallet).

c) $\sigma_1^2 = \sigma_2^2 = \sigma^2$

d) Nej, den upptäckta skillnaden verkar mindre än 5 parts per million.