TENTAMEN: Matematisk statistik för K (28 maj, 2009)

Kortfattade lösningar:

1) Låt A vara händelsen att komponent A fungerar och B händelsen att komponent B fungerar. Nu är $P(A) = 0.95$, $P(B) = 0.93$ och $P(a \cap B) = 0.90$.

   a) $P(A)P(B) = 0.8835 \neq 0.90 = P(A \cap B)$, dvs. att A och B är inte oberoende.

   b) $P(A|B) = P(A \cap B)/P(B) = 0.968$

   c) $P(B^C|A^C) = P(A^C \cap B^C)/P(A^C) = P((A \cup B)^C)/P(A^C)$
   $= (1 - P(A \cup B))/(1 - P(A)) = 0.4$

2) a) T-test för att testa $H_0 : \mu = 83$ mot $H_1 : \mu \neq 83$, där μ är den sanna genomsnittliga värdet av högsta temperaturen i juli. Teststatistikan är $T = (\bar{X} - \mu)/(S/\sqrt{n}) \sim T_{13}$ då H_0 är sann. Den får värdet -2.145. Därför att värdet inte är på kritiska området (större än $t^{(13)}_{0.005} = -3.012$), kan man inte förkasta H_0 på signifikansnivån 0.01. Det finns inget skäl att tro att meteorologen inte hade rätt.

   b) Nu testar man $H_0 : M = 83$ mot $H_1 : M \neq 83$, där M är den sanna medianen för högsta temperaturen i juli. Teststatistikan Q_+ är $Bin(14, 0.5)$-fördelad om H_0 är sann. Nu får den värdet 4 ("nollan" har sätts till "+" därför att mothypotesen är tvåsidig och det finns färre "-" än "+"). p-värdet blir $2 \cdot P(Q_+ \leq 4|H_0 \text{ sann}) = 2 \sum_{x=0}^{4} \binom{14}{x} 0.5^x 0.95^{14-x} = 0.179$. Därför att p-värdet är större än 0.01 kan man inte förkasta H_0 på signifikansnivån 0.01. Man kan inte säga att den median högsta temperaturen i juli skilljer sig från 83° Fahrenheit.

   c) I a) antar man att observationerna kommer från $N(\mu, \sigma)$, där σ är okänt och i b) antar man att observationerna kommer från en kontinuerlig fördelning.

   d) Teckentestet därför att enligt histogrammet kan man inte anta normalfördelning.

3) $\int_{-1}^{1} f(x)\,dx = \int_{-1}^{1} (a + bx)\,dx = 2a = 1$ vilket ger att $a = 0.5$. Vi vet också att $\mathbf{E}[X] = \int_{-1}^{1} xf(x)\,dx = \int_{-1}^{1} (0.5x + bx^2)\,dx = 2b/3 = 1/5$, vilket ger att $b = 0.3$.

4) a) Längden av intervallet är 2.6 och den måste vara lika med $\bar{x} + z_{\alpha/2} \cdot 5/\sqrt{n} - (\bar{x} - z_{\alpha/2} \cdot 5/\sqrt{n}) = 2z_{\alpha/2} \cdot 5/\sqrt{100} = z_{\alpha/2}$. Man har att $P(Z \leq z_{\alpha/2}) = P(Z \leq 2.6) = 0.9953 = 1 - \alpha/2$. Då blir $\alpha = 0.0094$ och konfidensgraden $1 - \alpha = 0.9906$ eller 99.06%.

b) Längden $2.6 = 2 * z_{\alpha/2} * 5/\sqrt{n}$ och då är $\sqrt{n} = \frac{2 \cdot 1.96 \cdot 5}{2.6}$ och $n \approx 57$.

5) a) $\mu_x = [X]$ och $\mu_Y = [Y]$. Vi testar $H_0 : \mu_X = \mu_Y$ mot $H_1 : \mu_X < \mu_Y$ därfär att sjuksköterskorna tror på att mördar med fler mödravårdsbesök föder tyngre barn än mödrar med färre besök.

b) Ett två stickprovs T-test: Teststatistikan $T = (\bar{X} - \bar{Y})/S_p\sqrt{1/n_X + 1/n_Y} \sim T_{26}$ om H_0 är sann. $S_p = ((n_X - 1)S_X^2 + (N_Y - 1)S_Y^2)/(n_X + n_Y - 2)$ och får värdet 467.0453. Teststatistikan får värdet -1.714. Detta jämförs med $-t_{0.05}^{(26)} = -1.706$. Nu är teststatistikans värde på kritiska området (< -1.706) och man kan förkasta H_0 på signifikansnivån 0.05. Det verkar som antalet mödravårdsbesök påverkar födelsevikten av barnet positivt.

c) Nej därför att $-t_{0.025}^{(26)} = -2.056 < -1.714 = t$, dvs. att t skulle inte ligga på kritiska området och man skulle inte kunna förkasta H_0.

d) $X \sim N(\mu_X, \sigma)$ och $Y \sim N(\mu_Y, \sigma)$. Stickproven är oberoende.

6) a) $Chol = 80.813 + 2.472Age$. Kolesterolnivån verkar öka med ökande ålder av patienten.

b) I 40-års ålder är kolesterolnivån ungefär 180 enligt regressionslinjen. Man kan inte säga någonting om kolesterolnivån i 80-års ålder därför att 80 ingår inte på intervallet av de observerade ålder-värdena.

c) Residualplot. Värdena är omkring 0, variansen ungefär konstant, *Age*-värdena kunde vara mer jämt fördelade på observationsintervallet, men man kan säga att den linjära regressionsmodellen verkar rimlig.

d) Cholesterol är en stokastisk variabel som förklaras av Age (fixerade värden). Man gissar att sambandet mellan Cholesterol och Age är linjärt. Om man bara skattar regressionslinjen, behövs det inga antaganden av fördelningen av feltermen osv.

7) a) $H_0 : \mu = 0$ mot $H_1 : \mu \neq 0$

b) $\alpha = 0.05$ och man förkastar H_0 på denna signifikansnivå därför att p-värdet ($7.6804e^{-21}$) är mycket mindre än signifikansnivån 0.05.Man ser också att det 95% konfidensintervallet inte täcker H_0-värdet 0.

c) $\bar{x} \pm t_{0.005}^{(99)} s/\sqrt{n} = 1.0980 \pm 2.626 \cdot 0.9213/\sqrt{100} = 1.0980 \pm 0.2419$, dvs. att det 99% konfidensintervallet är $(0.8561, 1.3399)$.