**TENTAMEN:** Matematisk statistik för K (30 maj, 2007)

**Kortfattade lösningar:**

1) a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B|A)$ och vi får att $P(A) = (P(A \cup B) - P(B))/(1 - P(B|A)) = \frac{2}{7}$.

b) $P(A \cap B) = P(A)P(B|A) = \frac{2}{7} \cdot 0.3 = \frac{3}{35}$.

c) $P(A \cap B) = \frac{3}{35} \neq 0$, dvs. att $A$ och $B$ inte är disjunkta händelser.

d) $P(A \cap B) = \frac{3}{35} \neq 0.174 = \frac{2}{7} \cdot 0.6 = P(A)P(B)$ och därför är $A$ och $B$ inte oberoende.

2) a) $f(x) \geq 0$ och $\int_{-\infty}^{\infty} f(x)\,dx = 1$

b) $f(x) = \frac{3}{8} + cx$, $0 \leq x \leq 3$, är en täthetsfunktion om $\int_0^3 f(x)\,dx = 1$, dvs. om $c = -\frac{1}{36}$ (då är också $f(x) > 0$ på $[0,3]$).

3) a) Låt $\mu_1$ och $\mu_2$ vara de sanna väntevärdena av korrosionshastigheten av gjutgärn med 0.1% resp. 0.2% krominnehåll. $\bar{X}_1$ och $\bar{X}_2$, och $S_1$ och $S_2$ är de motsvaramde medelvärdena och stickprovsstandardavvikelserna. Man vill tetsa hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 \neq \mu_2$ och använder två stickprovs dubbelsidigt $T$-test. Teststatistikan

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}},$$

där $S_p^2 = ((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)/(n_1 + n_2 - 2)$, är $T_{n_1+n_2-2}$-fördelad om $H_0$ är sann. Nu är $s_p^2 = 0.0935$, $t = 1.356$ och $t_{0.005}^{(23)} = 2.807$. Nu är $-t_{0.005}^{(23)} = -2.807 < t = 1.356 < 2.807 = t_{0.005}^{(23)}$ och vi kan inte förkasta $H_0$. Dvs. att man kan inte säga att krominnehållet påverkar korrosionshastigheten

b) 99% konfidensintervall för $\mu_1 - \mu_2$ är $\bar{x}_1 - \bar{x}_2 \pm t_{0.005}^{(23)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.165 \pm 0.355$, dvs. $[-0.17, 0.5]$. Nu är $0 \in [-0.17, 0.5]$ och därför kan man inte säga att det finns någon skillnad mellan de olika krominnehållen. Man har räknat 99% konfidensintervall för att få motsvarande resultat med testet i a).

c) Både föreslår att det finns ingen skillnad mellan de två krominnehållen.

d) $X_1 \sim N(\mu_1, \sigma)$ och $X_2 \sim N(\mu_2, \sigma)$

4) $X_1, ..., X_n$ är ett stickprov från $N(\mu, 2)$. Styrkan av testet är
$P(H_0 \text{ förkastas} | H_1 \text{ sann}) = P(\frac{\bar{X}-10}{2/\sqrt{n}} \leq z_{0.01} | \mu = 9)$
$= P(\frac{\bar{X}-9}{2/\sqrt{n}} \leq z_{0.01} + \frac{1}{2/\sqrt{n}}) = \Phi(z_{0.01} + \frac{1}{2}\sqrt{n}) = 0.90 = \Phi(1.28)$,
där $\Phi$ är fördelningsfunktionen av den standardiserade normalfördelningen.
Nu är $z_{0.01} = -2.33$ (hitta $z_{0.01}$ sådant att $P(Z \leq z_{0.01}) = 0.01$, där $Z \sim N(0,1)$), och då blir $-2.33 + \frac{1}{2}\sqrt{n} = 1.28$ och $n = 52.1$. Man kan säga att stickprovsstorleken måste vara minst 53.

5) a) $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 \neq \mu_2$, där $\mu_1$ och $\mu_2$ är de sanna väntevärden av vattenabsorbansen för valstryck 10kg/cm$^2$ resp. 20kg/cm$^2$.

b) Man har gjort ett parvist $T$-test och använt teststatistikan $T = \bar{D}/(S_d/\sqrt{n})$, där $\bar{D}$ är medelvärdet av skillnaderna $D_i = X_{1i} - X_{2i}$ (där $X_{11}, ..., X_{1,14}$ är vattenabsorbanserna för valstryck 10kg/cm$^2$ och $X_{21}, ..., X_{2,14}$ de motsvarande värdena för valstryck 20kg/cm$^2$) och $S_d^2 = \sum_{i=1}^{14}(D_i - \bar{D})^2/13$. Värdet för teststatistikan är 4.1028.

c) $H_0$ förkastas. Det verkar som vattenabsorbansen är högre med valstryck 10kg/cm$^2$ än med 20kg/cm$^2$.

d) Man antar att skillnaden $D$ är normalfördelad.

6) a) När det finns ett linjärt samband mellan två variabler och man vill förklara en av dem m.h.a. den andra. Man kan använda residualplottar för att se om modellen är lämplig.

b) Man vet att

$$\frac{B_0 - \beta_0}{S\sqrt{\sum x_i^2/n\sum(x_i - \bar{x})^2}} \sim T_{n-2}$$

om $Y_i \sim N(\beta_0 - \beta_1 x_i, \sigma)$ (se formelbladet för beteckningarna). Ett 95% konfidensintervall för $\beta_0$: Man vet att
$P(-t_{0.025} \leq \frac{B_0-\beta_0}{S\sqrt{\sum x_i^2/n\sum(x_i-\bar{x})^2}} \leq t_{0.025}) = 0.95$. Då får man att
$P\left(-t_{0.025}S\sqrt{\sum x_i^2/n\sum(x_i - \bar{x})^2} \leq B_0 - \beta_0 \leq t_{0.025}S\sqrt{\sum x_i^2/n\sum(x_i - \bar{x})^2}\right)$
$= P\left(B_0 - t_{0.025}S\sqrt{\sum x_i^2/n\sum(x_i - \bar{x})^2} \leq \beta_0 \leq B_0 + t_{0.025}S\sqrt{\sum x_i^2/n\sum(x_i - \bar{x})^2}\right)$
$= 0.95$ och konfedensintervallet blir $B_0 \pm t_{0.025}^{(n-2)}S\sqrt{\sum x_i^2/n\sum(x_i - \bar{x})^2}$,
där $B_0$ (skattningen för $\beta_0$) och $S$ hittar man i formelbladet.

7) a) Man testar $H_0 : \mu = 1.5$ppm mot $H_0 : \mu > 1.5$ppm på signifikansnivån 0.05 och använder ett ensidigt ett stickprovs $T$-test. Teststatistikan $T = (\bar{X} - 1.5)/(S/\sqrt{n})$ är $T_7$-fördelad, $s = \sum_{i=1}^{8}(x_i - \bar{x})^2/7 = 0.064$ och $\bar{x} = 1.47$. Man får att $t = -1.27$ och jämför det med $t_{0.05}^{(7)} = 1.895$. Nu är $t = -1.27 < 1.895 = t_{0.05}^{(7)}$ och man kan inte förkasta $H_0$. Man kan inte säga att den genomsnittliga ammoniakkoncentrationen är större än 1.5ppm.

b) Teckentestet för att testa $H_0 : M = 1.5$ppm mot $H_0 : M > 1.5$ppm, där $M$ är medianen av ammoniakkoncentrationen. Teststatistikan $Q_+$ är antalet positiva differenser $X_i - M$, vilket i vårt fall får värdet 3 (differensen som är noll räknas som negativ). Om $H_0$ är sann är $Q_+ \sim Bin(8, 0.5)$. $p$-värdet blir $P(Q_+ \geq 3 | H_0$ sann, dvs. $M = 1.5) = 1 - P(Q_+ \leq 2) = 1 - (\frac{1}{2})^8(1 + 8 + 28) = 0.855$. $p$-värdet är större än signifikansnivån 0.05 0ch man kan inte förkasta $H_0$. Medianen är inte större än 1.5ppm.

c) a): $X \sim N(\mu, \sigma)$ och b): $X$ har en kontinuerlig fördelning

d) Ja. Det verkar som $X$ var approximativt normalfördelad. Om resultaten hade varit olika, skulle man lita på det icke-parametriska testet.