TENTAMEN: Matematisk statistik för K (20 augusti 2009)

Kortfattade lösningar:

1) Man kan definiera två händelser: A är händelsen att man får en femma i matte och B händelsen att man får en femma i fysik. Nu är $P(A) = 82/500$, $P(B) = 73/500$ och $P(A \cap B) = 42/500$.

   a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.226$

   b) $P(A \text{ men inte } B) = P(A) - P(A \cap B) = 0.08$

   c) $P(A^C \cup B^C) = 1 - P(A \cap B) = 1 - 42/500 = 0.916$

2) a) Väntevärdesriktighet och att variansen minskar då stickprovsstorleken ökar.

   b) Ja. $\mathbf{E}[\bar{X}] = \mu$ så att det är en väntevärdesriktig skattning för μ och $Var(\bar{X}) = \sigma^2/n$, vilket blir mindre då n blir större. (Här $\mathbf{E}[X] = \mu$ och $Var(X) = \sigma^2$.)

3) a) Testa hypotesen $H_0 : \mu = 100$ mot $H_1 : \mu > 100$. Då är $T = (\bar{X} - 100)/(s/\sqrt{n}) \sim T_{71}$ och får ett värde 1.697. Nu är $t = 1.697 > 1.294 = t_{0.10}$ och man förkastar H_0 på signifikansnivån 0.10. Cyanidnivån verkar vara högre än 100mg/kg.

   b) $\alpha = 0.05$: $t_{71,0.05} = 1.667$ och H_0 förkastas
   $\alpha = 0.01$: $t_{71,0.005} = 2.380$ och H_0 accepteras
   Desto större är α desto större är risken att man förkastar H_0 även om den är sann. Desto mindre risk man tar desto svårare det är att förkasta H_0.

4) 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 $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$.

5) 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.

6) Låt X vara densiteten. Då är $X \sim N(0.0046, 9.6 \cdot 10^{-8})$.

a) $P(0.004 < X < 0.005) = P(\frac{0.004-0.0046}{0.0003} < \frac{X-0.00046}{0.0003} < \frac{0.005-0.0046}{0.0003}) = P(-2 < Z < 1.33) = \Phi(1.33) - \Phi(-2) = 0.8854$, där $Z \sim N(0,1)$ och Φ är fördelningsfunktionen av den standardiserade normalfördelningen.

b) $P(X > x_0) = 0.05$ och man borde hitta x_0.
$P(X \leq x_0) = P(\frac{X-0.0046}{0.0003} \leq \frac{x_0-0.0046}{0.0003}) = \Phi(\frac{x_0-0.0046}{0.0003}) = 0.95 = \Phi(1.645)$ och $x_0 = 0.0051$.

7) a) Konfidensintervallet är $\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$, där $S_p^2 = ((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)/(n_1 + n_2 - 2)$. Nu är $\bar{x}_1 = 6150$, $\bar{x}_2 = 5250$, $\alpha = 0.01$, $t_{0.005}^{(28)} = 2.763$, $n_1 = 16$, $n_2 = 14$ och $s_p^2 = 6040.179$, och ett 99% konfidensintervall blir $900 \pm 79 = [821, 979]$ och det verkar som det är bra att använda kedjor (0 är inte på intervallet).

b) Man testar hypotesen $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 > \mu_2$. Teststatistikans värde är $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_p^2(1/n_1 + 1/n_2)} = 31.64$, vilket är större än $t_{0.01}^{28} = 2.467$. Därför förkastas H_0 på signifikansnivån 0.01 och man kan säga att det verkar bra att använda kedjor.

c) $X_{1i} \sim N(\mu_1, \sigma)$ och $X_{2i} \sim N(\mu_2, \sigma)$ och att de två stickproven är oberoende.