TENTAMEN: Matematisk statistik för K (28 maj, 2008)

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) Man testar $H_0 : \mu_1 = \mu_2$ mot $H_1 : \mu_1 < \mu_2$ (där μ_1 är det förväntade poängtalet för barnen utsatta för kokain och μ_2 motsvarande för den andra gruppen) på signifikansnivån 0.01 genom att använda 2 stickprovs T-test. Teststatistikan är $T = (\bar{X}_1 - \bar{X}_2)/S_p\sqrt{1/n_1 + 1/n_2}$, där $S_p^2 = ((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)/(n_1 + n_2 - 2)$, och den är T_{374}-fördelad då H_0 är sann. $\bar{X}_1$ och $\bar{X}_2$ är medelvärden i kokaingruppen resp. icke-kokaingruppen och S_1^2 och S_2^2 är motsvarande stickprovsvarianserna. Nu är de observerade värdena $s_p = 3.0$ och $t = -2.908$. Därför att $t = -2.908 < -2.326 = -t_{0.01}^{374}$ är på kritiska området, förkastas H_0 på signifikansnivån 0.01. Dvs. att kokainutsätta barn verkar ha lägre poängtal i mattetestet än de kokainfria barn. Man måste anta att varianserna i de två grupperna är lika.

   b) Ett 99% konfidensintervall för $\mu_1 - \mu_2$ är $\bar{X}_1 - \bar{X}_2 \pm t_{0.005}^{374} S_p\sqrt{1/n_1 + 1/n_2}$ och detta blir $[-1.70, -0.10]$. Här måste man också anta att variansen är samma i de två grupperna. Därför att 0 inte är med på intervallet och alla värden på intervallet är negativa, verkar det som de kokainutsätta barn verkar ha lägre poängtal i mattetestet än de kokainfria barn.

   c) Desto större konfidensgrad, desto längre intervall; desto större varians, desto längre intervall; och desto större stickprovsstorlekar, desto kortare intervall.

   d) Inte nödvändigtvist därför att konfidensintervallet i b) motsvarar ett dubbelsidigt test men i a) gör man ett enkelsidigt test.

3) I samtliga fall (faktorer) har man med 95% sannolikhet hittat ett intervall, som täcker det "normala" värdet. Dvs. att sannolikheten att intervallet inte täcker det "normala" värdet är 0.05. Låt X vara antalet intervall som inte täcker det "normala" värdet. Då är $X \sim Bin(50, 0.05)$ och $P(X \geq 2) = 1 - P(X \leq 1) = 1 - P(X = 1) - P(X = 0) = 0.72$.

4) a) Ett parvist T-test därför att man jämför den faktiska och förutsagda temperaturen samma dag. Man testar $H_0 : \mu_D = 0$ mot $H_1 : \mu_D \neq 0$, där μ_D är den förväntade skillanden av de två temperaturen. Teststatitikan är $T = \bar{D}/(S_D/\sqrt{n})$, vilken är T_{11}-fördelad då H_0 är sann. Nu är $D_i = X_i - Y_i$, och $\bar{D}$ och S_D är medelvärdet resp. stickprovsstandardavvikelsen av D. Nu är $t = -2.569$, vilket är på kritiska området därför att $-2.569 < -2.201 = -t_0.025$. H_0 förkastas på signifikansnivån 0.05. Det verkar som att den faktiska och förutsagda temperaturen inte är samma.

b) Teststyrkan $= P(H_0$ förkastas$|H_1$ sann$) =$
$P(T < -t_{0.025}$ eller $T > t_{0.025}|\mu_D = \mu_1 - \mu_2 = -1) =$
$P(T < -t_{0.025}|\mu_D = -1) + P(T > t_{0.025}|\mu_D = -1) =$
$P(\bar{D}/(S_D/\sqrt{n}) < -t_{0.025}|\mu_D = -1)$
$+P(\bar{D}/(S_D/\sqrt{n}) > t_{0.025}|\mu_D = -1) =$
$P((\bar{D}+1)/(S_D/\sqrt{n}) < -t_{0.025} + 1/(S_D/\sqrt{n}))$
$+P((\bar{D}+1)/(S_D/\sqrt{n}) > t_{0.025} + 1/(S_D/\sqrt{n})) =$
$P(T' < -t_{0.025} + 1.713)) + P(T' > t_{0.025} + 1.713)) =$
$P(T' < -0.488) + 1 - P(T' \leq 3.914) =$
$1 - P(T' \leq 0.488) + 1 - P(T' \leq 3.914),$
där $T' = (\bar{D}+1)/(S_D/\sqrt{n})$ är T_{11}-fördelat. Nu är $P(T' \leq 0.488)$ mellan 0.6 och 0.75 och $P(T' \leq 3.914)$ mellan 0.995 0ch 0.999, och då är teststyrkan mellan 0.251 och 0.405.

5) a) Om min egen hypotes (gissning) är att det genomsnittliga sanna värdet är större än nollhypotesvärdet, då testar jag H_0 mot $H_1 : \mu > \mu_0$; om min hypotes är att det genomsnittliga sanna värdet är mindre än nollhypotesvärdet, då testar jag H_0 mot $H_1 : \mu < \mu_0$; och om jag bara tycker att värdet borde skilja sig från nollhypotesvärdet testar jag H_0 mot $H_1 : \mu \neq \mu_0$. Jag borde inte använda datan för att bestämma H_1.

b) Ha ett tillräckligt stort stickprov.

6) a) Man kan se en linjär trend men inte så tydligt. Det verkar som X och Y ökar samtidigt, vilket skulle ge ett positivt värde för korrelationskoefficienten. Man skulle gissa att korrelationskoefficienten blir närmare till 0 än 1.

b) Genom att använda formeln för korrelationskoefficienten (sista formeln i formelbladet), får man $r = 0.309$. Det verkar inte vara så stark linjärt samband mellan de två variablerna men de är svagt positivt korrelerade. Det är svårt att säga mer utan att veta någonting om noggranheten av skattningen.

c) I både är man intresserad av linjärt samband mellan två saker. Korrelation används då både två variabler är stokastiska; regression då man vill förklara en stokastisk variabel m.h.a. en fixt variabel.

7) a) T-test för att testa $H_0 : \mu = 5.67$ mot $H_1 : \mu \neq 5.67$, där μ är det sanna väntevärdet av vikten. Teststatistikan är $T = (\bar{X} - \mu)/(S/\sqrt{n})$ (S är stickprovsstandardavvikelsen av vikten) och den är T_9-fördelad då H_0 är sann. Nu är $\mu = 5.67$, $s = 0.0669$, $\bar{x} = 5.6766$ och $n = 10$ och t blir 0.312. Teststatistikans värde är inte på kritiska området (mellan $-t_{0.05} = -1.833$ och $t_{0.05} = 1.833$) och därför förkastar man inte H_0 på signifikansnivån 0.1. Den genomsnittliga vikten verkar inte skilja sig från 5.67. För att utföra testet måste man anta att vikten är normalfördelad.

b) Teckentestet för att testa $H_0 : M = 5.67$ mot $H_1 : M \neq 5.67$, där M är den sanna medianvikten. Teststatistikan är Q_+ =antalet positiva $X_i - M$. Det observerade värdet av Q_+ är 6 och p-värdet blir
$2P(Q_+ \geq 6|H_0 \text{ sann}) = 2P(Q_+ \geq 6|Q_+ \sim Bin(10, 0.5))$
$= 2\sum_{x=6}^{10} \binom{10}{x} 0.5^{10} = 0.75$. p-värdet är mycket större än 0.1 och man förkastar inte H_0 på signifikansnivån 0.1. Medianvikten verkar inte skilja sig från 5.67g. För att utföra testet måste man anta att fördelningen av vikten är kontinuerlig.

c) Vikten verkar inte vara normalfördelad så att man borde inte använda T-testet i a), hellre teckentestet i b).

d) Teckentestet därför att när man har en kontinuerlig fördelning, är sannolikheten att man observerar ett värde som är precis samma som medianen noll. Om man får sådana värden, kan man antingen kasta bort dem (om de är få jämfört med stickprovsstorleken) eller sätta skillnaden $X_i - M$ antingen till "-" eller till "+" så att detta inte hjälper att förkasta H_0.