Lösningar till Matematisk statistik för Z & NP den 21 okt 2000

1. Oberoendet ger $P[A \cap B] = \frac{1}{7} \cdot \frac{1}{3} = \frac{1}{21}$. Vi får $P[A \cup B] = P[A] + P[B] - P[A \cap B] = \frac{1}{7} + \frac{1}{3} - \frac{1}{21} = \frac{3}{7}$, $P[A \setminus B] = P[A \cap B'] = \frac{1}{7} \cdot \frac{2}{3} = \frac{2}{21}$ och $P[A \,|\, B] = P[A] = \frac{1}{7}$.

2. Bayes formel säger

$$P[D+ \,|\, T_i] = \frac{P[T_i \,|\, D+]\, P[D+]}{P[T_i \,|\, D+]\, P[D+] + P[T_i \,|\, D-]\, P[D-]}$$

Insättning ger

$$P[D+ \,|\, T_0] = \frac{0.42 \cdot 0.92}{0.42 \cdot 0.92 + 0.96 \cdot 0.08} \approx 0.83$$

$$P[D+ \,|\, T_1] = \frac{0.24 \cdot 0.92}{0.24 \cdot 0.92 + 0.02 \cdot 0.08} \approx 0.99$$

Om test indikerar att alla artärerna är intakta, T_0, så minskar sannolikheten att patienten har sjukdomen, medan om testet indikerar att en artär är förkalkad, T_1, så är det genast mycket troligt att patienten har sjukdomen.

3. Vi ska beräkna $P[f = x]$ för $x = 0, 1, \ldots, n$. Tänk dig en etta då A inträffar och en nolla annars. Ett typiskt utfall sådant att $f = x$ kan då representeras med en sträng av x st ettor och $n - x$ st nollor. Varje sådant utfall har sannolikheten $p^x\,(1-p)^{n-x}$ och eftersom det finns $\binom{n}{x}$ olika sätt att välja de x st försök som A inträffar i, följer

$$P[f = x] = \binom{n}{x} p^x\,(1-p)^{n-x}$$

Detta är sannolikhetstätheten för en $\mathrm{Bin}(n, p)$-fördelad stokastisk variabel.

4. Observera först att $T > t$ om, och endast om, $T_1 > t$ och $T_2 > t$. Härur följer

$$P[T > t] = P[T_1 > t,\, T_2 > t] = P[T_1 > t] P[T_2 > t]$$

$$= e^{-t/\beta_1}\, e^{-t/\beta_1} = e^{-t(1/\beta_1 + 1/\beta_2)} = e^{-t(\beta_1+\beta_2)/(\beta_1\beta_2)}$$

Vi ser av detta uttryck att T är exponentialfördelad med väntevärde

$$E[T] = \frac{\beta_1\,\beta_2}{\beta_1 + \beta_2}$$

5. Medianen är det mittersta värdet, alltså 606. Första kvartilen är medianen i den första halvan av datamängden, alltså medianen i $186, 218, 473, 515, 606$. Vi ser att q_1 är 473. Analogt ses att q_3 är 869. Vissa anser att första halvan av datamängder är $186, 218, 473, 515$. Isåfall blir första kvartilen $q_1 = (218 + 473)/2 = 345.5$ och q_3 beräknas analogt till 873. Detta kan man naturligtvis inte invända emot.

6. Att $\bar{X}$ är väntevärdesriktig visas i beviset av sats 7.1.1 i Milton & Arnold (sid 227). Låt stickprovet vara $X_1, \ldots, X_n$. Oberoendet gör att

$$\mathrm{Var}\left[\sum_i X_i\right] = \sum_i \mathrm{Var}[X_i] = n\sigma^2$$

Vi får nu

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_i X_i\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

7. Vi har att $\bar{x} = 4.796$, $s = 2.305$ och $n = 10$. Man vill påvisa att $\mu > 4$ och har därför att testa nollhypotesen $H_0 : \mu \leq 4$ mot $H_1 : \mu > 4$. Det är stora värden på $\bar{x}$ som tyder på att alternativet kan vara sannt (ju större $\bar{x}$ är, desto mer vågar man tro att $\mu > 4$). Så förkastelseregeln blir $\bar{x} \geq c$ och det gäller att bestämma c. Testet ska vara på nivån $\alpha = 5\%$. Vi ska därför lösa kritiska värdet c ur $0.05 = P[\bar{X} \geq c]$. För att göra detta behöver vi en modell för våra mätningar. En bra sådan är att mätningarna är $\mathrm{N}(\mu, \sigma)$-fördelade, bl a därför att om detta ej är uppfyllt så ger cgs ändå att $\bar{X} \overset{\text{ap}}{\sim} \mathrm{N}(\mu, \sigma/\sqrt{n})$, vilket är vad vår analys egentligen förutsätter. Så vi räknar som om mätningarna är $\mathrm{N}(\mu, \sigma)$-fördelade. Vi vet då att $T = (\bar{X} - \mu)/(s/\sqrt{n}) \sim t(n-1) = t(9)$ och får därför

$$0.05 = P[\bar{X} \geq c] = P\left[T \geq \frac{c-4}{s/\sqrt{n}}\right]$$

ty det värsta som kan hända under H_0 är att $\mu = 4$. Ur tabell över $t(9)$-fördelningen fås att $P[T \geq 1.833] = 0.05$, vilket ger att

$$\frac{c-4}{s/\sqrt{n}} = 1.833 \quad \Rightarrow \quad c = 4 + 1.833 s/\sqrt{n} = 5.336$$

Vi ser att $\bar{x} = 4.796 \not\geq 5.336 = c$. Vi kan alltså inte förkasta H_0 på nivån 5%. Utslaget är tydligt, så den ev approximation vi gör saknar säkert betydelse.

8. Vi antar att stickproven är $\mathrm{N}(\mu_1, \sigma)$- resp $\mathrm{N}(\mu_2, \sigma)$-fördelade. Om standardavvikelserna ej skulle vara lika, fungerar inte den analys vi kommer att göra, så därför börjar vi med att kolla om antagandet verkar rimligt. Vi har att $X = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$ (σ_1, σ_2 är här standardavvikelserna i resp stickprov). Ett konfidensintervall med konfidensgraden 0.9 härleds ur $0.9 = P\left[c_1 \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq c_2\right]$. Ur $F(16,14)$-tabell fås att $P[X \leq 2.445] = 0.95$ och att $P[X < 0.421] = 0.05$. M a o, $c_1 = 0.421$ och $c_2 = 2.445$, vilket ger att $0.9 = P\left[0.421 \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq 2.445\right]$ och konfidensintervallet för kvoten σ_2^2/σ_1^2 blir $(0.88,\ 5.09)$. Konfidensgraden är 90%. Trots att $s_2^2/s_1^2 = 2.08$ har vi alltså ingen egentlig orsak att tro något annat än att $\sigma_1 = \sigma_2$ (ty $1 \in (0.88,\ 5.09)$). Vi skattar den gemensamma variansen σ^2 med sammanvägningen

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 9.282$$

och σ skattas av $s_p = 3.047$. Ett 95% konfidensintervall för differensen $\mu_1 - \mu_2$ härleds ur

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2) = t(30)$$

Ur $t(30)$-tabellen fås $P[T > 2.750] = 0.005$ och genom att skriva om $-2.75 \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq 2.75$ och sätta in uppmätta värden erhålles

$$\mu_1 - \mu_2 = 2.30 \pm 2.97 \quad (99\%)$$

Ett annat sätt att skriva detta är $-0.67 \leq \mu_1 - \mu_2 \leq 5.27$ (99%). Om normalfördelningsantagandet ej är korrekt, så gör det inte så mycket eftersom analysen utgår ifrån medelvärdena och dessa är iallafall approximativt normalfördelade p g a cgs.