PRELIMINÄRA LÖSNINGSFÖRSLAG Sannolikhetslära och inferensteori II, C1 (7.5 hp).

1. De kontinuerliga slumpvariablerna X och Y har täthetsfunktionen

$$f(x,y) = 1/3, \qquad x \leq y \leq x+1, \quad 0 \leq x \leq 3.$$

   (a) X och Y kan inte vara oberoende, eftersom de möjliga värdena på Y bestäms av X ($X \leq Y \leq X+1$).

   (b) Vi ska beräkna $E(Y|X=x)$. Vi börjar med att beräkna

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f(x)},$$

där

$$f(x) = \int_x^{x+1} \frac{1}{3} dy = \frac{1}{3}.$$

Vi får därmed

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f(x)} = \frac{1/3}{1/3} = 1, \qquad x \leq y \leq x+1,$$

dvs $Y|X=x$ är $U(x, x+1)$-fördelad. Alltså är

$$E(Y|X=x) = x + 1/2.$$

   (c) Beräkningen av kovariansen mellan X och Y, förenklas med räkneregeln

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Vi beräknar vart och ett av de tre väntevärdena:

$$E(X) = \int_0^3 \int_x^{x+1} \frac{x}{3} dx dy = \int_0^3 \frac{x}{3} dx = \frac{3}{2}.$$

Med hjälp av del (b) får vi nu

$$E(Y) = E(E(Y|X=x)) = E(X+1/2) = E(X) + 1/2 = 2.$$

Vidare är

$$E(XY) = \int_0^3 \int_x^{x+1} \frac{xy}{3} dy dx = \int_0^3 \frac{2x^2+x}{6} dx = \frac{15}{4}.$$

Vi får därmed

$$Cov(X,Y) = \frac{15}{4} - \frac{3}{2} \cdot 2 = \frac{3}{4} = 0.75.$$

2. Den kontinuerliga slumpvariabeln Y har täthetsfunktionen:

$$f(y) = \frac{1}{2}(1+\theta y), \qquad -1 < y < 1, \quad -1 < \theta < 1.$$

(a) För att försöka hitta en väntevärdesriktig skattare av θ kan vi använda exempelvis momentmetoden. Vi har att

$$E(Y) = \int_{-1}^{1} \frac{1}{2}(1+\theta y)y dy = \Big[\frac{y^2}{4} + \theta\frac{y^3}{6}\Big]_{-1}^{1} = \frac{\theta}{3}.$$

Alltså är $E(\bar{Y}) = E(Y) = \theta/3$ och $E(3\bar{Y}) = 3E(\bar{Y}) = \theta$. Så $\hat{\theta} = 3\bar{Y}$ är en väntevärdesriktig skattare av θ.

(b) Eftersom $\hat{\theta}$ är väntevärdesriktig så är den konsistent om $Var(\hat{\theta}) \to 0$ då $n \to \infty$ (Sats 9.1, s. 450 i kursboken). Vi har att

$$Var(3\bar{Y}) = 9Var(\bar{Y}) = \frac{9Var(Y)}{n}.$$

För ett fixt värde på θ så är $Var(Y)$ en konstant[1], som inte beror på n. Därmed gäller att

$$\frac{9Var(Y)}{n} \to 0$$

då $n \to \infty$, och $\hat{\theta} = 3\bar{Y}$ är därmed konsistent.

3. Låt $X_1, X_2, \dots, X_{n_1}, X_{n_1+1}, \dots, X_n$ vara oberoende $Bin(1,p)$-fördelade slumpvariabler, med $1 \le n_1 < n/2$ och $n_2 = n - n_1$. Vi studerar

$$\tilde{X} = \frac{n}{2}\Big(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i + \frac{1}{n_2}\sum_{i=n_1+1}^{n} X_i\Big).$$

(a) Vi skriver om $\tilde{X}$ som

$$\tilde{X} = \frac{n}{2n_1}\sum_{i=1}^{n_1} X_i + \frac{n}{2n_2}\sum_{i=n_1+1}^{n} X_i.$$

Vi har att den momentgenererande funktionen för X_i är $m_{X_i}(t) = q + pe^t$, så

$$m_{\frac{n}{2n_2}X_i}(t) = m_{X_i}\Big(\frac{n}{2n_2}t\Big) = q + pe^{\frac{n}{2n_2}t}.$$

[1] Beräkningar ger att $Var(Y) = \frac{1}{3} - (\frac{\theta}{3})^2 \le \frac{1}{3}$.

Räknereglerna för momentgenererande funktioner för linjärkombinationer av oberoende slumpvariabler ger därmed att

$$m_{\tilde{X}}(t) = \left(q + pe^{\frac{n}{2n_1}t}\right)^{n_1}\left(q + pe^{\frac{n}{2n_2}t}\right)^{n_2}.$$

(b) Låt $Z \sim HypGeo(n, X, n_1)$. Vi vill visa att $\tilde{X}$ har samma fördelning som $X + Y$ där

$$Y = \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n_1 - n_2}{2n_2}X.$$

Vi börjar med att studera delsumman $\sum_{i=1}^{n_1} X_i$. Där väljer vi ut n_1 observationer bland n möjliga, varav X stycken är "lyckade". Per definition är $Z = \sum_{i=1}^{n_1} X_i$ därmed fördelad som $HypGeo(n, X, n_1)$. Vidare gäller att $\sum_{i=n_1+1}^{n} X_i = X - Z$. Vi kan därmed skriva $\tilde{X}$ som

$$\tilde{X} = \frac{n}{2n_1}Z + \frac{n}{2n_2}(X - Z) = \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n}{2n_2}X.$$

Vi söker Y sådant att $X + Y = \tilde{X}$, så att

$$\begin{aligned}
Y = \tilde{X} - X &= \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n}{2n_2}X - X \\
&= \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n - 2n_2}{2n_2}X \\
&= \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n_1 - n_2}{2n_2}X,
\end{aligned}$$

där vi i sista steget utnyttjade att $n - 2n_2 = n_1 + n_2 - 2n_2 = n_1 - n_2$. Vi kan därmed konstatera att $\tilde{X}$ har samma fördelning som $X + Y$, vilket skulle visas.

4. Låt $Y_1, Y_2, \ldots, Y_n$ vara ett slumpmässigt stickprov från en fördelning med täthetsfunktion

$$f(y) = (1 + \theta)\, y^{\theta},\ \ \theta > -1,\ \ 0 < y \leq 1$$

(a) För att finna det bästa kritiska området finner vi först ett uttryck för likelihoodfunktionen

$$L(\theta) = \prod_{i=1}^{n}\left[(1 + \theta)\, y_i^{\theta}\right] = (1 + \theta)^n \left(\prod_{i=1}^{n} y_i\right)^{\theta}$$

vilket innebär att likelihoodkvoten blir

$$\frac{L(\theta_0)}{L(\theta_a)} = \frac{(1+\theta_0)^n \left(\prod_{i=1}^n y_i\right)^{\theta_0}}{(1+\theta_a)^n \left(\prod_{i=1}^n y_i\right)^{\theta_a}} = \left(\frac{1+\theta_0}{1+\theta_a}\right)^n \left(\prod_{i=1}^n y_i\right)^{\theta_0-\theta_a} < k$$

vilket leder till att

$$W = \prod_{i=1}^n y_i > \left[k\left(\frac{1+\theta_0}{1+\theta_a}\right)^{-n}\right]^{1/(\theta_0-\theta_a)} = k'$$

där olikheten har bytt riktning eftersom $\theta_a > \theta_0$. Eftersom det kritiska området inte beror på θ_a kommer det test där detta kritiska område används och mothypotesen är på formen $H_a : \theta > \theta_0$ att vara ett UMP-test.

(b) Antag att man vill testa

$$\begin{aligned} H_0 &: \theta = 0 \\ H_a &: \theta > 0 \end{aligned}$$

med en signifikansnivå på $\alpha = 0.05$ (dvs på 5%-nivån). Till vår hjälp har vi endast en observation på Y. Ange det kritiska området (förkastelseområdet) och testets styrkefunktion. Vi ska alltså förkasta nollhypotesen då

$$Y > k'$$

där k' bestäms av

$$\Pr(Y > k' \mid \theta = 0) = 0.05$$

Då $\theta = 0$ följer att $f(y) = 1,\ 0 < y \leq 1$, dvs Y är då $U(0,1)$. Således följer att $k' = 0.95$, dvs

$$C = \{y : y > 0.95\}$$

Fördelningsfunktionen för Y ges av

$$F(y) = \int_0^y (1+\theta)\, t^\theta dt = \left[t^{\theta+1}\right]_0^y = y^{\theta+1},\ \theta > -1,\ 0 < y \leq 1$$

Således följer att

$$\text{power}(\theta) = \Pr(Y > 0.95 \mid \theta) = 1 - 0.95^{\theta+1},\ \theta > -1$$

(ab*) Om man i a-uppgiften istället på traditionellt sätt logaritmerar så blir lösningsförfarandet mer komplext. Vi får

$$\ln\left[\frac{L(\theta_0)}{L(\theta_a)}\right] = n\ln\left(\frac{1+\theta_0}{1+\theta_a}\right) + (\theta_0-\theta_a)\sum_{i=1}^{n}\ln y_i < \ln k$$

vilket leder till att

$$\sum_{i=1}^{n}\ln y_i > \frac{\ln k - n\ln\left(\frac{1+\theta_0}{1+\theta_a}\right)}{(\theta_0-\theta_a)} = k'$$

där olikheten har bytt riktning eftersom $\theta_a > \theta_0$. Eftersom det kritiska området inte beror på θ_a kommer det test där mothypotesen är på formen $H_a : \theta > \theta_0$ att vara ett UMP-test. Med en enda observation ska vi förkasta nollhypotesen då

$$X = \ln Y > k'$$

där k' bestäms av

$$\Pr(X > k' \mid \theta = 0) = 0.05$$

Då $\theta = 0$ följer att Y är $U(0,1)$ som medför att

$$\begin{aligned} F_X(x) &= \Pr(X \le x) = \Pr(\ln Y \le x) = \Pr(Y \le e^x) = \\ &= F_Y(e^x) = e^x, \quad -\infty < x \le 0 \end{aligned}$$

Vi finner nu den kritiska punkten k' via

$$0.05 = \Pr(X > k' \mid \theta = 0) = 1 - F_X(k') = 1 - e^{k'}$$

vilket innebär att $k' = \ln 0.95$, dvs

$$C = \{\ln y : \ln y > \ln 0.95\}$$

eller som tidigare

$$C = \{y : y > 0.95\}$$