Föreläsning 6. Flerdimensionella slumpvariabler

Inledande exempel: två-mynt experiment. Vi kastar två mynt n gånger. Tre slumpvariabler av intresse

$$\begin{aligned} X_1 &= \text{antal klave-klave}, \\ X_2 &= \text{antal krona-krona}, \\ X_3 &= \text{antal klave-krona eller krona-klave}, \end{aligned}$$

så att

$$X_1 + X_2 + X_3 = n.$$

Frågor

1. vad är fördelning av X_1?
2. vad är fördelning av X_2?
3. vad är fördelning av X_3?
4. är X_1 och X_2 oberoende slumpvariabler?
5. vad är fördelning av $Y = X_1 + X_2$?

Tvådimensionell simultan fördelning och marginella fördelningar

Tvådimensionell simultan fördelning av (X, Y) i diskreta fallet

$$p_{X,Y}(i,j) = \mathrm{P}(X = i, Y = j) = \mathrm{P}(\{X = i\} \cap \{Y = j\}).$$

I kontinuerliga fallet simultan täthetsfunktion $f_{X;Y}(x, y)$ definieras genom

$$\mathrm{P}(x < X \le x + \epsilon_1, y < Y \le y + \epsilon_2) \approx \epsilon_1 \epsilon_2 f_{X;Y}(x, y).$$

Två marginella fördelningar

$$\begin{aligned} p_X(i) &= \sum_j p_{X,Y}(i,j), \quad p_Y(j) = \sum_i p_{X,Y}(i,j), \\ f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx. \end{aligned}$$

Två oberoende slumpvariabler
X och Y är oberoende om

$$p_{X,Y}(i,j) = p_X(i)p_Y(j)$$

för alla par (i, j), eller om för alla par (x, y)

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

Väntevärdet av produkten av oberoende slumpvariabler
Om $X_1, X_2, \dots, X_n$ är oberoende slumpvariabler, då

$$\mathrm{E}(X_1 X_2 \cdots X_n) = \mathrm{E}(X_1)\mathrm{E}(X_2) \cdots \mathrm{E}(X_n).$$

Bevis för två slumpvariabler:

$$\mathrm{E}(XY) = \iint xy f_{X,Y}(x,y)dxdy = \iint xy f_X(x) f_Y(y)dxdy = \int x f_X(x)dx \int y f_Y(y)dy = \mathrm{E}(X)\mathrm{E}(Y).$$

Variansen av summan av oberoende slumpvariabler
Om $X_1, X_2, \ldots, X_n$ är oberoende slumpvariabler, då

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n).$$

Bevis. Låt

$$\hat{X}_i = X_i - \mathrm{E}(X_i).$$

Enligt föregående regel för samtliga par $i \neq j$

$$\mathrm{E}(\hat{X}_i \hat{X}_j) = \mathrm{E}(\hat{X}_i)\mathrm{E}(\hat{X}_j) = 0.$$

Därför

$$\begin{aligned}\mathrm{Var}(X_1 + \ldots + X_n) &= \mathrm{E}(\hat{X}_1 + \ldots + \hat{X}_n)^2 = \mathrm{E}(\hat{X}_1^2 + \ldots + \hat{X}_n^2 + 2\hat{X}_1\hat{X}_2 + 2\hat{X}_1\hat{X}_3 + \ldots) \\ &= \mathrm{E}(\hat{X}_1^2) + \ldots + \mathrm{E}(\hat{X}_n^2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n).\end{aligned}$$

Exempel: två-mynt experiment

Tvådimensionell simultan fördelning av (X_1, X_2). Om (i, j, k) är icke-negativa heltal sådana att

$$i + j + k = n,$$

då

$$p_{X_1,X_2}(i,j) = \mathrm{P}(X_1 = i, X_2 = j) = \binom{n}{i,j,k}(\tfrac{1}{4})^i(\tfrac{1}{4})^j(\tfrac{1}{2})^k,$$

där $\binom{n}{i,j,k}$ är multinomial koefficient

$$\binom{n}{i,j,k} = \frac{n!}{i!j!k!}$$

som ger antal möjliga placeringar för i klave-klave och j krona-krona utfall bland n försök.

Total sannolikhet enligt multinomialsatsen

$$\sum_{i,j:i+j\le n} p_{X_1,X_2}(i,j) = \sum_{i,j,k:i+j+k=n} \binom{n}{i,j,k}(\tfrac{1}{4})^i(\tfrac{1}{4})^j(\tfrac{1}{2})^k = (\tfrac{1}{4} + \tfrac{1}{4} + \tfrac{1}{2})^n = 1.$$

Marginell fördelning av X_1

$$\begin{aligned}p_{X_1}(i) &= \sum_j p_{X_1,X_2}(i,j) = \sum_{j=0}^{n-i} \binom{n}{i,j,k}(\tfrac{1}{4})^i(\tfrac{1}{4})^j(\tfrac{1}{2})^{n-i-j} \\ &= (\tfrac{1}{4})^i \sum_{j=0}^{n-i} \frac{n!}{i!j!(n-i-j)!}(\tfrac{1}{4})^j(\tfrac{1}{2})^{n-i-j} = \frac{n!}{i!(n-i)!}(\tfrac{1}{4})^i \sum_{j=0}^{n-i} \frac{(n-i)!}{j!(n-i-j)!}(\tfrac{1}{4})^j(\tfrac{1}{2})^{n-i-j} \\ &= \frac{n!}{i!(n-i)!}(\tfrac{1}{4})^i(\tfrac{1}{4} + \tfrac{1}{2})^{n-i} = \binom{n}{i}(\tfrac{1}{4})^i(\tfrac{3}{4})^{n-i}, \quad i = 0, 1, \ldots, n;\end{aligned}$$

och marginell fördelning av X_2

$$p_{X_2}(j) = \sum_i p_{X_1,X_2}(i,j) = \binom{n}{j}(\tfrac{1}{4})^j(\tfrac{3}{4})^{n-j}, \quad j = 0, 1, \ldots, n.$$

Svaren på tidigare frågor:

1. Marginell fördelning av X_1 är $\mathrm{Bin}(n, \frac{1}{4})$.
2. Marginell fördelning av X_2 är $\mathrm{Bin}(n, \frac{1}{4})$.
3. Marginell fördelning av X_3 är $\mathrm{Bin}(n, \frac{1}{2})$.
4. Slumpvariablerna X_1 och X_2 är beroende på varandra eftersom $\mathrm{P}(X_1 = i)$ är inte lika med

$$\mathrm{P}(X_1 = i | X_2 = j) = \frac{\mathrm{P}(X_1 = i, X_2 = j)}{\mathrm{P}(X_2 = j)} = \frac{\frac{n!}{i!j!k!}(\frac{1}{4})^i(\frac{1}{4})^j(\frac{1}{2})^k}{\frac{n!}{j!(n-j)!}(\frac{1}{4})^j(\frac{3}{4})^{n-j}}$$
$$= \frac{(n-j)!}{i!k!}(\tfrac{1}{4})^i(\tfrac{1}{2})^k(\tfrac{4}{3})^{n-j} = \frac{(n-j)!}{i!k!}(\tfrac{1}{3})^i(\tfrac{2}{3})^k, \quad i = 0, 1, \ldots, n-j$$

   där $k = n - j - i$. Med andra ord, betingad fördelning av X_1 är $\mathrm{Bin}(n - j, \frac{1}{3})$.
5. Eftersom

$$Y = X_1 + X_2 = n - X_3$$

   och samtidigt $X_3 \sim \mathrm{Bin}(n, \frac{1}{2})$, vi får $Y \sim \mathrm{Bin}(n, \frac{1}{2})$.

Kovarians och korrelationskoefficient

Kovarians $\mathrm{Cov}(X, Y)$ är ett beroendemått för ett par slumpvariabler:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(X - \mu_X)(Y - \mu_Y) = \mathrm{E}(XY) - \mu_X \mu_Y.$$

Notera att

$$\mathrm{Cov}(X, X) = \mathrm{E}(X - \mu_X)^2 = \mathrm{Var}(X).$$

Tolkning av $\mathrm{Cov}(X, Y)$. Produkten

$$W = (X - \mu_X)(Y - \mu_Y)$$

visar hur två slumpvariablerna (X, Y) "ko-varierar" kring respektive väntevärden. Om $W > 0$ då (X, Y) samtidigt är antingen både högre än deras väntevärde eller både lägre än deras väntevärde. Om $W < 0$ då X och Y avviker från deras väntevärde åt olika håll. Därför positiv kovarians

$$\mathrm{Cov}(X, Y) = \mathrm{E}(W)$$

tolkas som positiv beroende mellan X och Y, medan $\mathrm{Cov}(X, Y) < 0$ antyder åt negativ beroende mellan X och Y.

Korrelationskoefficient

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \mathrm{E}(\tfrac{X - \mu_X}{\sigma_X})(\tfrac{Y - \mu_Y}{\sigma_Y})$$

är dimensionslös och alltid $-1 \leq \rho \leq 1$.

Exempel: två-mynt experiment

Låt oss beräkna $\mathrm{Cov}(X_1, X_2)$. Vi har

$$\mu_1 = \mathrm{E}(X_1) = \tfrac{n}{4}, \quad \mu_2 = \mathrm{E}(X_2) = \tfrac{n}{4}, \quad \sigma_1^2 = \mathrm{Var}(X_1) = \tfrac{3n}{16}, \quad \sigma_2^2 = \mathrm{Var}(X_2) = \tfrac{3n}{16},$$

samt

$$\mathrm{E}(X_1 X_2) = \sum_{j=1}^{n} \sum_{i=1}^{n-j} (ij) p_{X_1, X_2}(i, j) = \sum_{j=1}^{n} j \sum_{i=1}^{n-j} i p_{X_1, X_2}(i, j) = \sum_{j=1}^{n} j p_{X_2}(j) \sum_{i=1}^{n-j} i \mathrm{P}(X_1 = i | X_2 = j)$$
$$= \sum_{j=1}^{n} j p_{X_2}(j) \tfrac{n-j}{3} = \tfrac{n}{3} \sum_{j=1}^{n} j p_{X_2}(j) - \tfrac{1}{3} \sum_{j=1}^{n} j^2 p_{X_2}(j) = \tfrac{n}{3}\mathrm{E}(X_2) - \tfrac{1}{3}\mathrm{E}(X_2^2)$$
$$= \tfrac{n^2}{16} - \tfrac{n}{16}.$$

Det innebär att kovariansen är negativ

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1 X_2) - \mu_1\mu_2 = -\tfrac{n}{16}.$$

Korrelationskoefficient

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1\sigma_2} = -\frac{\frac{n}{16}}{\frac{3n}{16}} = -\frac{1}{3}.$$

Additionsregel för variansen

Om slumpvariabler X och Y är beroende då

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

Bevis. Låt

$$\hat{X} = X - \mathrm{E}X, \quad \hat{Y} = Y - \mathrm{E}Y.$$

Då

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{E}(\hat{X}+\hat{Y})^2 = \mathrm{E}(\hat{X}^2 + 2\hat{X}\hat{Y} + \hat{Y}^2)\\ &= \mathrm{E}(\hat{X}^2) + \mathrm{E}(2\hat{X}\hat{Y}) + \mathrm{E}(\hat{Y}^2) = \mathrm{Var}(X) + 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y).\end{aligned}$$

Exempel: två-mynt experiment
Vi vet att $X_1 \sim \mathrm{Bin}(n, \frac{1}{4})$, $X_2 \sim \mathrm{Bin}(n, \frac{1}{4})$ och $Y = X_1 + X_2 \sim \mathrm{Bin}(n, \frac{1}{2})$. Det innebär att

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \tfrac{3n}{16}, \quad \mathrm{Var}(X_1 + X_2) = \tfrac{n}{4}$$

Vi verifierar additionsregel för variansen

$$\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2) = \tfrac{3n}{16} + \tfrac{3n}{16} - \tfrac{n}{8} = \tfrac{n}{4} = \mathrm{Var}(X_1 + X_2).$$

Notera att om $X_1 \sim \mathrm{Bin}(n, \frac{1}{4})$ och $X_2 \sim \mathrm{Bin}(n, \frac{1}{4})$ vore oberoende, då vi skulle haft $Y \sim \mathrm{Bin}(2n, \frac{1}{4})$.