1 Chapter 1

Exercise 1.1(a)

Let X be the sum of the two number obtained.
Let X_1 be the number obtained on Die 1.
Let X_2 be the number obtained on Die 2.
Thus, $X = X_1 + X_2$, and

$$E_x = \{X = x\} = \{X_1 + X_2 = x\} = \{\ X_1 = x_1, X_2 = x - x_1 \mid 1 \leq x_1, x - x_1 \leq 6\ \}$$

Now,

$$1 \leq x - x_1 \leq 6 \quad \Longleftrightarrow \quad -1 \geq x_1 - x \geq -6 \quad \Longleftrightarrow \quad x - 1 \geq x_1 \geq x - 6 \quad \Longleftrightarrow \quad x - 6 \leq x_1 \leq x - 1$$

Hence,

$$E_x = \{X = x\} = \{X_1 + X_2 = x\} = \{\ X_1 = x_1, X_2 = x - x_1 \mid \max\{1, x-6\} \leq x_1 \leq \min\{6, x-1\}\ \}$$

$$\begin{aligned} P(E_x) &= \sum_{x_1=\max\{1,x-6\}}^{\min\{6,x-1\}} P(X_1 = x_1, X_2 = x - x_1) \;=\; \sum_{x_1=\max\{1,x-6\}}^{\min\{6,x-1\}} \frac{1}{6^2} \\ &= \frac{1}{6^2}\left(\min\{6, x-1\} - \max\{1, x-6\} + 1\right) \end{aligned}$$

Next, note that

$$\min\{6, x-1\} = \left\{ \begin{array}{cl} x-1, & \text{if } x = 2, 3, \ldots, 6 \\ 6, & \text{if } x = 7, 8, \ldots, 12 \end{array} \right. \quad \text{and} \quad \max\{1, x-6\} = \left\{ \begin{array}{cl} 1, & \text{if } x = 2, 3, \ldots, 6 \\ x-6, & \text{if } x = 7, 8, \ldots, 12 \end{array} \right.$$

Hence,

$$\begin{aligned} P(E_x) &= \frac{1}{6^2}\left(\min\{6, x-1\} - \max\{1, x-6\} + 1\right) \;=\; \frac{1}{36}\left\{ \begin{array}{cl} (x-1) - 1 + 1, & \text{if } x = 2, 3, \ldots, 6 \\ 6 - (x-6) + 1, & \text{if } x = 7, 8, \ldots, 12 \end{array} \right. \\ &= \frac{1}{36}\left\{ \begin{array}{cl} x-1, & \text{if } x = 2, 3, \ldots, 6 \\ 13-x, & \text{if } x = 7, 8, \ldots, 12 \end{array} \right. \end{aligned}$$

□

Exercise 1.18

Recapitulation of the rules of craps: *Let x be the number obtained on the first roll. If $x \in \{7, 11\}$, then the player wins. If $x \in \{2, 3, 12\}$, then the player loses. If $x \in \{4, 5, 6, 8, 9, 10\}$, then the player keeps rolling, until either 7 is rolled or x is rolled. If x is rolled first (before 7 is rolled), then the player wins. If 7 is rolled first (before x is rolled), then the player loses.*

Let W be the $\{0, 1\}$-valued random variable such that $W = 1$ if the player wins, and $W = 0$ if the player loses. We thus seek to compute $P(W = 1)$. Let X be (the random variable of) the sum of the two numbers obtained on the first roll. Note that $\text{Range}(X) = \{\, 2, 3, 4, \ldots, 12 \,\}$. Then,

$$\begin{aligned}
P(W = 1) &= \sum_{x=2}^{12} P(W = 1|X = x) \cdot P(X = x) \\
&= P(W = 1|X = 7)\, P(X = 7) + P(W = 1|X = 11)\, P(X = 11) + \sum_{x \in \{4,5,6,8,9,10\}} P(W = 1|X = x) \cdot P(X = x)
\end{aligned}$$

Now, note that $P(W = 1|X = 7) = P(W = 1|X = 11) = 1$, $P(X = 7) = \dfrac{6}{36} = \dfrac{1}{6}$, and $P(X = 11) = \dfrac{2}{36} = \dfrac{1}{18}$.

From Exercise 1.1(a), we have:

$$\begin{aligned}
P(X = x) &= \frac{1}{6^2}\left(\min\{6, x-1\} - \max\{1, x-6\} + 1\right) = \frac{1}{36}\begin{cases} (x-1) - 1 + 1, & \text{if } x = 2,\, 3,\, \ldots,\, 6 \\ 6 - (x-6) + 1, & \text{if } x = 7,\, 8,\, \ldots,\, 12 \end{cases} \\
&= \frac{1}{36}\begin{cases} x - 1, & \text{if } x = 2,\, 3,\, \ldots,\, 6 \\ 13 - x, & \text{if } x = 7,\, 8,\, \ldots,\, 12 \end{cases}
\end{aligned}$$

Next, let Y_n be the random variable of the sum of the two numbers obtained on the $(n+1)$st roll. Then,

$$\begin{aligned}
P(W = 1|X = x) &= \sum_{n=1}^{\infty} \left[1 - P(Y_n = 7) - P(Y_n = x)\right]^{n-1} \cdot P(X = x) \\
&= P(X = x) \cdot \sum_{n=1}^{\infty} \left[1 - P(Y_n = 7) - P(Y_n = x)\right]^{n-1} \\
&= P(X = x) \cdot \frac{1}{1 - \left[1 - P(Y = 7) - P(Y = x)\right]} \\
&= \frac{P(X = x)}{P(Y = 7) + P(Y = x)} \\
&= \frac{P(X = x)}{\frac{1}{6} + P(Y = x)}
\end{aligned}$$

Hence,

$$
\begin{aligned}
P(W=1) &= \sum_{x=2}^{12} P(W=1|X=x)\cdot P(X=x) \\
&= P(W=1|X=7)\,P(X=7) + P(W=1|X=11)\,P(X=11) + \sum_{x\in\{4,5,6,8,9,10\}} P(W=1|X=x)\cdot P(X=x) \\
&= \frac{6}{36}+\frac{2}{36}+\sum_{x\in\{4,5,6,8,9,10\}}\frac{P(X=x)^2}{\frac{1}{6}+P(Y=x)} \\
&= \frac{6}{36}+\frac{2}{36}+\frac{(\frac{4-1}{36})^2}{\frac{1}{6}+\frac{4-1}{36}}+\frac{(\frac{5-1}{36})^2}{\frac{1}{6}+\frac{5-1}{36}}+\frac{(\frac{6-1}{36})^2}{\frac{1}{6}+\frac{6-1}{36}}+\frac{(\frac{13-8}{36})^2}{\frac{1}{6}+\frac{13-8}{36}}+\frac{(\frac{13-9}{36})^2}{\frac{1}{6}+\frac{13-9}{36}}+\frac{(\frac{13-10}{36})^2}{\frac{1}{6}+\frac{13-10}{36}} \\
&= \frac{6}{36}+\frac{2}{36}+\frac{(1/36)^2}{1/36}\left(\frac{3^2}{6+3}+\frac{4^2}{6+4}+\frac{5^2}{6+5}+\frac{5^2}{6+5}+\frac{4^2}{6+4}+\frac{3^2}{6+3}\right) \\
&= \frac{6}{36}+\frac{2}{36}+\frac{2}{36}\left(\frac{3^2}{6+3}+\frac{4^2}{6+4}+\frac{5^2}{6+5}\right) \;=\; \frac{1}{36}\left[6+2+2\left(\frac{9}{9}+\frac{16}{10}+\frac{25}{11}\right)\right] \\
&= \frac{1}{36}\left[8+2\left(\frac{536}{110}\right)\right] \;=\; \frac{1}{36}\left[\frac{1952}{110}\right] \;=\; \frac{1}{2^2\cdot 3^2}\left[\frac{2^5\cdot 61}{2\cdot 5\cdot 11}\right] \\
&= \frac{2^2\cdot 61}{3^2\cdot 5\cdot 11} \;\approx\; 0.4929293
\end{aligned}
$$

$\square$

Exercise 1.19(a)

Let n be the number of workers in the sample. Let X_i, $i = 1, 2, \ldots, n$, be $\{0, 1\}$-valued random variables defined by:

$$X_i \;=\; \begin{cases} 1, & \text{if the } i\text{th subject is highly exposed,} \\ 0, & \text{if the } i\text{th subject is NOT highly exposed} \end{cases}$$

Define

$$S_n \;:=\; \sum_{i=1}^{n} X_i\,, \quad \text{and} \quad S_{n-1} \;:=\; \sum_{i=1}^{n-1} X_i\,.$$

First, note that

$$\theta_n \;=\; P(S_n \text{ is even})\,, \quad \text{and} \quad \theta_{n-1} \;=\; P(S_{n-1} \text{ is even})\,.$$

Note also that

$$\begin{aligned} \theta_n \;&=\; P(S_n \text{ is even}) \;=\; P(X_n = 1)P(S_{n-1} \text{ is odd}) \;+\; P(X_n = 0)P(S_{n-1} \text{ is even}) \\ &=\; \pi_h\,(1-\theta_{n-1}) + (1-\pi_h)\,\theta_{n-1} \;=\; \pi_h + (1-2\pi_h)\,\theta_{n-1} \end{aligned}$$

Thus, the desired difference equation is:

$$\theta_n \;=\; \pi_h + (1-2\pi_h)\,\theta_{n-1} \tag{1.1}$$

Exercise 1.19(b)

To solve the difference equation (1.1) obtained in Exercise 1.19(a), we assume that θ_n has the following form:

$$\theta_n = \alpha + \beta\,\gamma^n \tag{1.2}$$

where α, β, and γ are unknown constants to be determined. We first make the following:

Observation: $\beta \neq 0$ and $\gamma \notin \{0,\, 1\}$.
Indeed, if $\beta = 0$ or $\gamma \in \{0,\, 1\}$, then θ_n would be constant in n. In that case, define $\theta := \theta_n = \theta_{n-1} = \cdots$. By the difference equation (1.1), we would then have

$$\theta = \pi_h + (1-2\pi_h)\,\theta \quad \Longrightarrow \quad 0 = \pi_h\,(1-2\theta) \quad \Longrightarrow \quad \theta = \frac{1}{2} \quad (\text{since } \pi_h > 0)$$

However, this contradicts the initial condition that $\theta_0 = 1$. Thus, this proves the assertion that $\beta \neq 0$ and $\gamma \notin \{0,\, 1\}$. (Note that if the sample size is 0, then the number of highly exposed subjects must be 0; hence $\theta_0 = P(S_0 \text{ is even}) = 1$, since we have here adopted the convention that 0 is "even.")

Now, substituting (1.2) into (1.1) yields:

$$\begin{aligned} \alpha + \beta\gamma^n \;&=\; \theta_n \;=\; \pi_h + (1-2\pi_h)\,\theta_{n-1} \\ &=\; \pi_h + (1-2\pi_h)\left(\alpha + \beta\gamma^{n-1}\right) \\ &=\; \alpha + \pi_h(1-2\alpha) + \beta\gamma^{n-1}(1-2\pi_h) \end{aligned}$$

Collecting terms involving γ on the right-hand side yields:

$$\pi_h(2\alpha - 1) \;=\; \beta\gamma^{n-1}\,(1 - 2\pi_h - \gamma)$$

Now, note that the left-hand side of the preceding equation is independent of γ, while the right-hand side is a scalar multiple of the $(n-1)$th power of γ; in other words, the right-hand side is a scalar multiple of a power of γ which is constant in n.

This happens if and only if either $\gamma \in \{0, 1\}$, or if the coefficient $\beta(1 - 2\pi_h - \gamma) = 0$. The preceding Observation (i.e. $\beta \neq 0$ and $\gamma \notin \{0, 1\}$) thus implies:

$$\gamma = 1 - 2\pi_h$$

Since $\pi_h > 0$, we furthermore conclude that

$$\alpha = \frac{1}{2}$$

We therefore have:

$$\theta_n \quad = \quad \frac{1}{2} + \beta\,(1 - 2\pi_h)^n$$

The initial condition $\theta_0 = 1$ now implies:

$$1 = \theta_0 = \frac{1}{2} + \beta\,(1 - 2\pi_h)^0 = \frac{1}{2} + \beta \quad \Longrightarrow \quad \beta = \frac{1}{2}$$

We may now conclude:

$$\theta_n \quad = \quad \frac{1}{2} + \frac{1}{2}(1 - 2\pi_h)^n$$

Lastly, if $\pi_h = 0.05$, then

$$\theta_{50} \quad = \quad \frac{1}{2} + \frac{1}{2}(1 - 2 \times 0.05)^{50} \quad \approx \quad 0.5025769$$

$\square$

Comment: For $0 < \pi_h < \frac{1}{2}$, the formula $\theta_n = \frac{1}{2} + \frac{1}{2}(1 - 2\pi_h)^n$ implies that $\theta_n > \frac{1}{2}$, for any $n = 1, 2, 3, \ldots$; in other words, there is a higher than $50:50$ chance that the number of highly exposed subjects in the sample is "even", whenever $0 < \pi_h < \frac{1}{2}$. This apparent asymmetry between odd and even is NOT surprising given the fact that 0 is regarded as "even" here, and that the probability that there are no highly exposed workers in the sample is high if π_h is "small" (e.g. $0 < \pi_h < \frac{1}{2}$).

Exercise 1.20(a)

$$p(D|S,x) = \frac{p(D,S,x)}{p(S,x)} = \frac{p(D,S,x)}{p(D,x)}\,\frac{p(D,x)}{p(S,x)} = p(S|D,x)\,\frac{p(D,x)/p(x)}{p(S,x)/p(x)} = p(S|D,x)\,\frac{p(D|x)}{p(S|x)}$$

Now, we are given that

$$p(S|D,x) = \pi_1\,, \quad \text{and} \quad p(D|x) = \frac{\exp\left(\beta_0 + \beta^T x\right)}{1+\exp\left(\beta_0+\beta^T x\right)}$$

So, we now proceed to compute $p(S|x)$. To this end,

$$\begin{aligned} p(S|x) &= \frac{p(S,x)}{p(x)} = \frac{1}{p(x)}\left(p(S,D,x)+p(S,\overline{D},x)\right) = \frac{1}{p(x)}\left(\frac{p(S,D,x)}{p(D,x)}p(D,x) + \frac{p(S,\overline{D},x)}{p(\overline{D},x)}p(\overline{D},x)\right) \\ &= p(S|D,x)p(D|x) + p(S|\overline{D},x)p(\overline{D}|x) \end{aligned}$$

Hence,

$$\begin{aligned} p(D|S,x) &= p(S|D,x)\frac{p(D|x)}{p(S|x)} = \frac{p(S|D,x)\,p(D|x)}{p(S|D,x)\,p(D|x)+p(S|\overline{D},x)\,p(\overline{D}|x)} = \frac{\pi_1\cdot p(D|x)}{\pi_1\cdot p(D|x)+\pi_0\cdot p(\overline{D}|x)} \\ &= \frac{\pi_1\cdot\dfrac{\exp\left(\beta_0+\beta^T x\right)}{1+\exp\left(\beta_0+\beta^T x\right)}}{\pi_1\cdot\dfrac{\exp\left(\beta_0+\beta^T x\right)}{1+\exp\left(\beta_0+\beta^T x\right)}+\pi_0\cdot\dfrac{1}{1+\exp\left(\beta_0+\beta^T x\right)}} = \frac{\pi_1\cdot\exp\left(\beta_0+\beta^T x\right)}{\pi_1\cdot\exp\left(\beta_0+\beta^T x\right)+\pi_0} \\ &= \frac{\dfrac{\pi_1}{\pi_0}\cdot\exp\left(\beta_0+\beta^T x\right)}{1+\dfrac{\pi_1}{\pi_0}\cdot\exp\left(\beta_0+\beta^T x\right)} = \frac{\exp\left[\,\log(\pi_1/\pi_0)+\beta_0+\beta^T x\,\right]}{1+\exp\left[\,\log(\pi_1/\pi_0)+\beta_0+\beta^T x\,\right]}\,, \end{aligned}$$

as required.

Comment: The above derivations show that, in a case-control study, if one has knowledge (or good estimate) of the ratio π_1/π_0, one can obtain an estimate for $p(D|x)$, the disease risk associated to covariate value x, from the quantity $p(D|S,x)$, which can be estimated from case-control study data as follows:

$$p(D|S,x) \;\approx\; \frac{\#(\text{subjects in sample with disease and covariate value } x)}{\#(\text{subjects in sample with covariate value } x)}$$

However, in practice, the ratio π_1/π_0 is rarely, if ever, known. And, without knowledge or estimate of π_1/π_0, the disease risk $p(D|x)$ associated to covariate value x can NOT be estimated based on data from a case-control study.

Exercise 1.20(b)

First, note that

$$\frac{p(D|x^*)}{p(\overline{D}|x^*)} = \frac{\exp\left(\beta_0+\beta^T x^*\right)/(1+\exp\left(\beta_0+\beta^T x^*\right))}{1/(1+\exp\left(\beta_0+\beta^T x^*\right))} = \exp\left(\beta_0+\beta^T x^*\right)$$

Similarly,

$$\frac{p(D|x)}{p(\overline{D}|x)} = \exp\left(\beta_0+\beta^T x\right)$$

Hence,

$$\theta_r = \theta_r(x^*,x) = \frac{p(D|x^*)/p(\overline{D}|x^*)}{p(D|x)/p(\overline{D}|x)} = \frac{\exp\left(\beta_0+\beta^T x^*\right)}{\exp\left(\beta_0+\beta^T x\right)} = \exp\left[\,\beta^T(x^*-x)\,\right]\,,$$

as required. Next,

$$\theta_c \;=\; \theta_c(x^*, x) \;=\; \frac{p(D|S,x^*)/p(\overline{D}|S,x^*)}{p(D|S,x)/p(\overline{D}|S,x)} \;=\; \frac{\exp\left[\,\log(\pi_1/\pi_0) + \beta_0 + \beta^T x^*\,\right]}{\exp\left[\,\log(\pi_1/\pi_0) + \beta_0 + \beta^T x\,\right]} \;=\; \exp\left[\,\beta^T(x^* - x)\,\right]\,,$$

as required.

Comment: Exercise 1.20(a) showed that, without knowledge or estimate of the ratio π_1/π_0, case-control study data can NOT be used to estimate the disease $p(D|x)$ associated to covariate value x. On the other hand, case-control study data can be readily used to estimate the odds ratio

$$\theta_c \;=\; \theta_c(x^*, x) \;:=\; \frac{p(D|S,x^*)/p(\overline{D}|S,x^*)}{p(D|S,x)/p(\overline{D}|S,x)}$$

Exercise 1.20(b) shows that θ_c is equal to

$$\theta_r \;=\; \theta_r(x^*, x) \;:=\; \frac{p(D|x^*)/p(\overline{D}|x^*)}{p(D|x)/p(\overline{D}|x)}$$

Thus, Exercise 1.20(a) and Exercise 1.20(b) together show that, while case-control study data can NOT be used to estimate disease risk $p(D|x)$ associated to covariate value x, they can be used to estimate the disease odds ratio

$$\theta_r \;=\; \theta_r(x^*, x) \;:=\; \frac{p(D|x^*)/p(\overline{D}|x^*)}{p(D|x)/p(\overline{D}|x)}$$

associated to the covariate value x^* against x.

Exercise 1.21(a)

Let D be the random variable defined by:

$$D \;:=\; \begin{cases} 1, & \text{if a given individual has IBD,} \\ 0, & \text{otherwise.} \end{cases}$$

Let S_1 be the random variable defined by:

$$S_1 \;:=\; \begin{cases} 1, & \text{Strategy \#1 asserts that a given individual has IBD,} \\ 0, & \text{otherwise.} \end{cases}$$

Let S_2 be the random variable defined by:

$$S_2 \;:=\; \begin{cases} 1, & \text{Strategy \#2 asserts that a given individual has IBD,} \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$\begin{aligned} P(S_1 = D) &= P(S_1 = D, D = 1) + P(S_1 = D, D = 0) \;=\; P(S_1 = D | D = 1)P(D = 1) + P(S_1 = D | D = 0)P(D = 0) \\ &= P(S_1 = D | D = 1)\theta + P(S_1 = D | D = 0)(1 - \theta) \end{aligned}$$

Next, note that

$$\begin{aligned} P(S_1 = D | D = 1) &= P(X \geq 2), \quad \text{where } \; X \sim \text{Binomial}(n = 3, p = \pi_1) \\ &= \binom{3}{2} \pi_1^2 (1 - \pi_1)^1 + \binom{3}{3} \pi_1^3 (1 - \pi_1)^0 \\ &= 3\pi_1^2 (1 - \pi_1) + \pi_1^3 \;=\; \pi_1^2 \,(3 - 2\pi_1) \end{aligned}$$

Similarly,

$$P(S_1 = D | D = 0) \;=\; \pi_0^2 \,(3 - 2\pi_0)$$

Therefore,

$$P(S_1 = D) \;=\; P(S_1 = D | D = 1)\theta + P(S_1 = D | D = 0)(1 - \theta) \;=\; \theta\pi_1^2 \,(3 - 2\pi_1) + (1 - \theta)\pi_0^2 \,(3 - 2\pi_0)$$

On the other hand, note that

$$P(S_2 = D | D = 1) \;=\; \pi_1 \quad \text{and} \quad P(S_2 = D | D = 0) \;=\; \pi_0$$

Hence,

$$\begin{aligned} P(S_2 = D) &= P(S_2 = D | D = 1)P(D = 1) + P(S_2 = D | D = 0)P(D = 0) \\ &= P(S_2 = D | D = 1)\theta + P(S_2 = D | D = 0)(1 - \theta) \\ &= \theta\pi_1 + (1 - \theta)\pi_0 \end{aligned}$$

Thus, a sufficent condition for $P(S_1 = D) \geq P(S_2 = D)$ is the following:

$$\pi_1^2 \,(3 - 2\pi_1) \geq \pi_1 \quad \text{and} \quad \pi_0^2 \,(3 - 2\pi_0) \geq \pi_0$$

Now,

$$\begin{aligned} \pi_1^2 \,(3 - 2\pi_1) \geq \pi_1 &\iff \pi_1 \,(3 - 2\pi_1) \geq 1 \\ &\iff 2\pi_1^2 - 3\pi_1 + 1 \leq 0 \\ &\iff (2\pi_1 - 1)(\pi_1 - 1) \leq 0 \\ &\iff \frac{1}{2} \leq \pi_1 \leq 1 \end{aligned}$$

Similarly,

$$\pi_0^2\,(3-2\pi_0) \geq \pi_1 \quad\Longleftrightarrow\quad \frac{1}{2} \leq \pi_0 \leq 1$$

We may now conclude that a sufficient condition for $P(S_1 = D) \geq P(S_2 = 0)$ is

$$\frac{1}{2} \leq \pi_0\,,\,\pi_1 \leq 1$$

Comment: The above sufficient condition shows that as long as the probability of each doctor giving a correct diagnosis is at least $\frac{1}{2}$ (i.e. $\frac{1}{2} \leq \pi_0\,,\,\pi_1 \leq 1$), Strategy #1 will outperform Strategy #2, in the sense that the probability that Strategy #1 giving a correct diagnosis will exceed that of Strategy #2.

Exercise 1.21(b)

References