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,

$$1 \leq x - x_1 \leq 6 \iff -1 \geq x_1 - x \geq -6 \iff x - 1 \geq x_1 \geq x - 6 \iff x - 6 \leq x_1 \leq x - 1$$

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

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

Hence,

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

□

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,

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

Hence,

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

□

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:

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

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:

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

Lastly, if $\pi_h = 0.05$, then

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

□

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

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

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)

Let S_3 be the random variable defined by:

$$S_3 \;:=\; \begin{cases} 1, & \text{Strategy \#3 asserts that a given individual has IBD,} \\ 0, & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} P(S_3 = D|D=1) &= P(Z \geq 3)\,, \quad \text{where } Z \sim \text{Binomial}(n=4, p=\pi_1) \\ &= \binom{4}{3}\pi_1^3(1-\pi_1)^1 + \binom{4}{4}\pi_1^4(1-\pi_1)^0 \\ &= 4\pi_1^3(1-\pi_1) + \pi_1^4 \\ &= \pi_1^3\,(4-3\pi_1) \end{aligned}$$

Similarly,

$$P(S_3 = D|D=0) \;=\; \pi_0^3\,(4-3\pi_0)$$

Hence,

$$\begin{aligned} P(S_3 = D) &= P(S_3 = D, D=1) + P(S_3 = D, D=0) \\ &= P(S_3 = D|D=1)P(D=1) + P(S_3 = D|D=0)P(D=0) \\ &= \theta\pi_1^3\,(4-3\pi_1) + (1-\theta)\pi_0^3\,(4-3\pi_0) \end{aligned}$$

Now, observe that

$$\begin{aligned} P(S_1 = D) - P(S_3 = D) &= \left[\theta\pi_1^2\,(3-2\pi_1) + (1-\theta)\pi_0^2\,(3-2\pi_0)\right] - \left[\theta\pi_1^3\,(4-3\pi_1) + (1-\theta)\pi_0^3\,(4-3\pi_0)\right] \\ &= \theta\pi_1^2\left(3-2\pi_1-4\pi_1+3\pi_1^2\right) + (1-\theta)\pi_0^2\left(3-2\pi_0-4\pi_0+3\pi_0^2\right) \\ &= 3\theta\pi_1^2\left(\pi_1^2-2\pi_1+1\right) + 3(1-\theta)\pi_0^2\left(\pi_0^2-2\pi_0+1\right) \\ &= 3\theta\pi_1^2\,(\pi_1-1)^2 + 3(1-\theta)\pi_0^2\,(\pi_0-1)^2 \\ &\geq 0 \end{aligned}$$

Comment: This shows that Strategy #1 is always preferable over Strategy #3, regardless of the values of π_0 and π_1 (despite the latter involving more doctors). □

Exercise 1.22

Let

- A be the event that an individual has Alzheimer's Disease.
- D be the event that an individual has diabetes.
- M be the event that an individual is male.

Note that

$$\begin{aligned}\pi_1 &:= P(A|D) = \frac{P(A,D)}{P(D)} = \frac{P(A,D,M)+P(A,D,\overline{M})}{P(D)} \\ &= \frac{P(A,D,M)}{P(D,M)}\frac{P(D,M)}{P(D)} + \frac{P(A,D,\overline{M})}{P(D,\overline{M})}\frac{P(D,\overline{M})}{P(D)} \\ &= P(A|D,M)P(M|D) + P(A|D,\overline{M})P(\overline{M}|D) \\ &= \pi_{11}\cdot P(M|D) + \pi_{10}\cdot P(\overline{M}|D)\end{aligned}$$

Similarly,

$$\begin{aligned}\pi_0 &:= P(A|\overline{D}) = \frac{P(A,\overline{D})}{P(\overline{D})} = \frac{P(A,\overline{D},M)+P(A,\overline{D},\overline{M})}{P(\overline{D})} \\ &= \frac{P(A,\overline{D},M)}{P(\overline{D},M)}\frac{P(\overline{D},M)}{P(\overline{D})} + \frac{P(A,\overline{D},\overline{M})}{P(\overline{D},\overline{M})}\frac{P(\overline{D},\overline{M})}{P(\overline{D})} \\ &= P(A|\overline{D},M)P(M|\overline{D}) + P(A|\overline{D},\overline{M})P(\overline{M}|\overline{D}) \\ &= \pi_{01}\cdot P(M|\overline{D}) + \pi_{00}\cdot P(\overline{M}|\overline{D})\end{aligned}$$

We ASSUME

- $\pi_{00}\neq 0$, $\pi_{01}\neq 0$, and $\pi_0\neq 0$.
- *homogeneity of risk ratio across gender groups*, i.e.

$$R_1 = R_0 =: R, \quad \text{where} \quad R_1 := \frac{\pi_{11}}{\pi_{01}}, \quad R_0 := \frac{\pi_{10}}{\pi_{00}}. \tag{1.3}$$

We seek to derive sufficient conditions for

$$R_c = R, \quad \text{where} \quad R_c := \frac{\pi_1}{\pi_0}. \tag{1.4}$$

Now, it follows immediately from (1.3) and (1.4) that

$$\pi_{11} = R\cdot\pi_{01} \quad \text{and} \quad \pi_{10} = R\cdot\pi_{00}$$

Hence,

$$\pi_1 = R\cdot\big(\pi_{01}\cdot P(M|D) + \pi_{00}\cdot P(\overline{M}|D)\big)$$

which in turn implies:

$$\frac{\pi_1}{\pi_0} = R\cdot\left(\frac{\pi_{01}\,P(M|D)+\pi_{00}\,P(\overline{M}|D)}{\pi_{01}\,P(M|\overline{D})+\pi_{00}\,P(\overline{M}|\overline{D})}\right)$$

Thus, (1.4) will follow if the following holds:

$$\frac{\pi_{01}\,P(M|D)+\pi_{00}\,P(\overline{M}|D)}{\pi_{01}\,P(M|\overline{D})+\pi_{00}\,P(\overline{M}|\overline{D})} = 1 \tag{1.5}$$

Now, note:

$$\begin{aligned}
(1.5) \iff & \pi_{01}\,P(M|D) + \pi_{00}\,P(\overline{M}|D) - \pi_{01}\,P(M|\overline{D}) - \pi_{00}\,P(\overline{M}|\overline{D}) = 0 \\
\iff & \pi_{01}\,\big[P(M|D) - P(M|\overline{D})\big] + \pi_{00}\,\big[P(\overline{M}|D) - P(\overline{M}|\overline{D})\big] = 0 \\
\iff & \pi_{01}\,\big[P(M|D) - P(M|\overline{D})\big] + \pi_{00}\,\big[1 - P(M|D) - 1 + P(M|\overline{D})\big] = 0 \\
\iff & [\,\pi_{01} - \pi_{00}\,] \cdot \big[\,P(M|D) - P(M|\overline{D})\,\big] = 0
\end{aligned}$$

Thus, two separate sufficient conditions for (1.4) are:

$$\pi_{01} \;=\; \pi_{00} \quad \text{and} \quad P(M|D) \;=\; P(M|\overline{D})$$

Furthermore,

$$\begin{aligned}
& \text{independence of } M \text{ and } D, \text{ i.e. } P(M|D) = P(M) \\
\implies & \frac{P(M,D)}{P(D)} \;=\; P(M,D) + P(M,\overline{D}) \\
\implies & P(M,D) \;=\; P(M,D)P(D) + P(M,\overline{D})P(D) \\
\implies & P(M,D)\,[\,1 - P(D)\,] \;=\; P(M,\overline{D})P(D) \\
\implies & \frac{P(M,D)}{P(D)} \;=\; \frac{P(M,\overline{D})}{P(\overline{D})} \\
\implies & P(M|D) \;=\; P(M|\overline{D})
\end{aligned}$$

Therefore, we may now conclude that two separate sufficient conditions for (1.4) are:

- independence of M and D, i.e. $P(M|D) \;=\; P(M)$, and
- $\pi_{01} \;=\; \pi_{00}$, i.e. $P(A|\overline{D}, M) = P(A|\overline{D}, \overline{M})$.

□

2 Chapter 1

Exercise 2.1(a)

Note that our "stopping criterion" here is that the sequence of selected individuals contains at least one individual with the rare blood disorder and at least one individual without the disorder. Thus, the following must hold: Let n be the length of a stopping sequence; the first $n-1$ individuals of the stopping sequence must all be of one type, while the n^{th} individual is of the other type.

Let 1 represent an individual with the rare blood disorder, while 0 an individual without the disorder. Then, since there are only four inddividuals with the disorder and three who do not, the following sequences are the only possible stopping sequences:

$$01,\ 001,\ 0001,\ 10,\ 110,\ 1110,\ 11110$$

The probabilities of the above admissible stopping sequences are tabulated below:

sequence s	$P(s)$	length of s	sequence s	$P(s)$	length of s
01	$\frac{3}{7}\cdot\frac{4}{6}$	2	10	$\frac{4}{7}\cdot\frac{3}{6}$	2
001	$\frac{3}{7}\cdot\frac{2}{6}\cdot\frac{4}{5}$	3	110	$\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{3}{5}$	3
0001	$\frac{3}{7}\cdot\frac{2}{6}\cdot\frac{1}{5}\cdot\frac{4}{4}$	4	1110	$\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{3}{4}$	4
			11110	$\frac{4}{7}\cdot\frac{3}{6}\cdot\frac{2}{5}\cdot\frac{1}{4}\cdot\frac{3}{3}$	5

We thus see that, letting N denote (the random variable of) the stopping sequence:

$$\begin{aligned}
P(N=2) &= P(01)+P(10) &&= \frac{3\cdot 4+4\cdot 3}{7\cdot 6} &&= \frac{4}{7}\\
P(N=3) &= P(001)+P(110) &&= \frac{3\cdot 2\cdot 4+4\cdot 3\cdot 3}{7\cdot 6\cdot 5} &&= \frac{2}{7}\\
P(N=4) &= P(0001)+P(1110) &&= \frac{3\cdot 2\cdot 1\cdot 4+4\cdot 3\cdot 2\cdot 3}{7\cdot 6\cdot 5\cdot 4} &&= \frac{4}{35}\\
P(N=5) &= P(11110) &&= \frac{4\cdot 3\cdot 2\cdot 1\cdot 3}{7\cdot 6\cdot 5\cdot 4\cdot 3} &&= \frac{1}{35}
\end{aligned}$$

Hence,

$$E[N] = \sum_{n=2}^{5} n\cdot P(N=n) = 2\cdot\left(\frac{4}{7}\right)+3\cdot\left(\frac{2}{7}\right)+4\cdot\left(\frac{4}{35}\right)+5\cdot\left(\frac{1}{35}\right) = \frac{13}{5} = 2.6$$

Comment (the underlying probability space used in Exercise 2.1(a)):

Let

$$\Omega := \{\,(x_i)_{i=1}^{\infty}\,|\,x_i\in\{0,1\}\,\} = \left\{\begin{array}{c}\text{all infinite sequences}\\ \text{of 0's and 1's}\end{array}\right\}$$

Note that each finite sequence $y = (y_1, y_2, \ldots, y_n)$ of zeros and ones can be regarded as a subset of Ω as follows:

$$y = (y_1, y_2, \ldots, y_n) \quad \longleftrightarrow \quad \{\, (x_i)_{i=1}^{\infty} \in \Omega \mid x_i = y_i,\ i = 1, 2, \ldots, n \,\}$$

Let Θ be the set of all finite sequences of zeros and ones. Then, by the preceding convention, we have that $\Theta \subset \text{PowerSet}\,(\Omega)$. Note the the underlying set of the probability space used in Exercise 2.1(a) is Ω, the probability measure on Ω is first defined subsets of Ω belonging to Θ, and then extend to the σ-algebra generated by Θ. Note also that, given any two members of Θ (as subsets of Ω), either they are disjoint or one is a subset of the other.

Exercise 2.1(b)

First, consider the concrete example that $M = 10$, $N = 100$, $k = 3$, and $X = 5$. Then,

$$\begin{aligned} P(X = 5\,;\, 100, 10, 3) &= \binom{5-1}{3-1} \cdot \frac{10}{100} \cdot \frac{9}{99} \times \frac{90}{98} \cdot \frac{89}{97} \times \frac{8}{96} = \binom{5-1}{3-1} \cdot \frac{10}{100} \cdot \frac{10-1}{100-1} \cdot \frac{10-2}{100-2} \cdot \frac{90}{100-3} \cdot \frac{90-1}{100-4} \\ &= \binom{5-1}{3-1} \frac{10\,!}{(10-3)!} \cdot \frac{90\,!}{(90-(5-3))!} \cdot \frac{(100-5)!}{100\,!} \\ &= \binom{5-1}{3-1} \binom{100-5}{10-3} \Big/ \binom{100}{10} \end{aligned}$$

In general, we therefore have:

$$P(X = x\,;\, N, M, k) = \frac{\binom{x-1}{k-1}\binom{N-x}{M-k}}{\binom{N}{M}}$$

□

Exercise 2.26(a)

Recall that:

$$h(t) \;:=\; \lim_{h\to 0} \frac{P(t \le T \le t+h \,|\, t \le T)}{h}, \qquad F(t) \;=\; P(T < t), \qquad f(t) \;=\; F'(t), \qquad S(t) = 1 - F(t)$$

We want to prove:

$$h(t) \;=\; \frac{f(t)}{S(t)}.$$

PROOF

$$\begin{aligned} P(t \le T \le t+h \,|\, t \le T) \;&=\; \frac{P(t \le T \le t+h \cap t \le T)}{P(t \le T)} \;=\; \frac{P(t \le T \le t+h)}{P(t \le T)} \;=\; \frac{P(T \le t+h) - P(T \le t)}{P(t \le T)} \\ &=\; \frac{F(t+h) - F(t)}{1 - F(t)} \end{aligned}$$

Hence,

$$\frac{P(t \le T \le t+h \,|\, t \le T)}{h} \;=\; \frac{1}{1-F(t)} \cdot \frac{F(t+h) - F(t)}{h}$$

Hence,

$$h(t) \;=\; \lim_{h\to 0} \frac{P(t \le T \le t+h \,|\, t \le T)}{h} \;=\; \frac{1}{1-F(t)} \cdot \lim_{h\to 0} \frac{F(t+h) - F(t)}{h} \;=\; \frac{1}{1-F(t)} \cdot F'(t) \;=\; \frac{f(t)}{S(t)}$$

□

Exercise 2.26(b)

Recall that $S(t) = 1 - F(t)$, hence $S'(t) = -F'(t) = -f(t)$. Consequently,

$$h(t) \;=\; \frac{f(t)}{S(t)} \;=\; -\frac{S'(t)}{S(t)} \;=\; -\,\frac{\mathrm{d}}{\mathrm{d}t}\Big(\log S(t)\Big)$$

Thus,

$$\int_0^t h(\tau)\,\mathrm{d}\tau \;=\; -\int_0^t \frac{\mathrm{d}}{\mathrm{d}\tau}\Big(\log S(\tau)\Big)\,\mathrm{d}\tau \;=\; -\,[\,\log S(\tau)\,]_0^t \;=\; -\log S(t) + \log S(0)$$

Now, $S(0) = 1 - F(0) = 1 - P(T < 0) = 1 - 0 = 1$; hence, $\log S(0) = \log(1) = 0$. We thus have:

$$\log\left(S(t)\right) \;=\; -\int_0^t h(\tau)\,\mathrm{d}\tau \quad\Longrightarrow\quad S(t) \;=\; \exp\left\{-\int_0^t h(\tau)\,\mathrm{d}\tau\right\} \;=\; \exp\left\{-H(t)\right\}$$

□

Exercise 2.26(c)

Assuming $\lim\limits_{t\to\infty} t \cdot S(t) \;=\; \lim\limits_{t\to\infty} t\,(1 - F(t)) \;=\; 0$, we want to prove that

$$E\,[\,T\,] \;=\; \int_0^\infty S(t)\,\mathrm{d}t$$

PROOF

$$\begin{aligned} E[T] &= \int_0^\infty t \cdot f(t)\,\mathrm{d}t = -\int_0^\infty t \cdot S'(t)\,\mathrm{d}t \\ &= -[t \cdot S(t)]_0^\infty + \int_0^\infty S(t)\,\mathrm{d}t, \quad \text{by integration by parts: } \int u\,\mathrm{d}v = uv - \int v\,\mathrm{d}u,\ u = t,\ \mathrm{d}v = S'(t)\,\mathrm{d}t \\ &= \int_0^\infty S(t)\,\mathrm{d}t \end{aligned}$$

□

Exercise 2.26(d)

We want to prove:

$$E[H(X)] = F_T(c),$$

where $X = \min\{T, c\}$, $H(t) = \int_0^t h(\tau)\,\mathrm{d}\tau$.

PROOF We first derive the cumulative probability function of the random variable $X := \min\{T, c\}$. First, note that $\text{range}(X) = (0, c]$.

$$\begin{aligned} P(X \le x) &= P(\min\{T, c\} \le x) = P(T \le x \text{ or } c \le x) \\ &= P(T \le x) + P(c \le x) - P(T \le x \text{ and } c \le x) \\ &= \begin{cases} P(T \le x) + 1 - P(T \le x), & \text{fox } c \le x \\ P(T \le x) + 0 - 0, & \text{fox } c > x \end{cases} \\ &= \begin{cases} 1, & \text{for } c \le x \\ P(T \le x), & \text{for } c > x \end{cases} \\ &= \begin{cases} P(T \le x), & \text{for } x < c \\ 1, & \text{for } x = c \end{cases} \end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t} P(X \le x) = \begin{cases} f_T(x), & \text{for } 0 < x < c \\ \text{undefined}, & \text{for } x = c \end{cases}$$

We thus see that the probability density μ_X of the random variable X is a probability measure on $\text{range}(X) = (0, c]$. Over the interval $x \in (0, c)$, μ_X is representable by a probability density function $f_X(x) = f_T(x)$ for $x \in (0, c)$. When restricted to the single point $x = c$, μ_X is the point mass measure with $\mu_X(c) = 1 - F_T(c)$. In other words, the probability density measure μ_X of X is given by:

$$\begin{aligned} \mu_X &= \begin{cases} f_T(x), & \text{for } 0 < x < c \\ P(X = c), & \text{for } x = c \end{cases} \\ &= \begin{cases} f_T(x), & \text{for } 0 < x < c \\ P(T \ge c), & \text{for } x = c \end{cases} \\ &= \begin{cases} f_T(x), & \text{for } 0 < x < c \\ 1 - F_T(c), & \text{for } x = c \end{cases} \end{aligned}$$

Thus,

$$E[H(X)] = \int_{[0,\,c]} H(x)\,\mathrm{d}\mu_X = \int_0^c H(x) \cdot f_T(x)\,\mathrm{d}x + H(c)\,\mu_X(c)$$

Now, note that

$$H(c) \cdot \mu_X(\,c\,) \;=\; -\log S(c) \cdot (\,1 - F_T(\,c\,)\,)$$

Next, recall from Exercise 2.26(b) that $H(t) = -\log S(t)$, $f_T(x) = -S'(x)$.

$$\begin{aligned}
\int_0^c H(x) \cdot f_T(x)\,\mathrm{d}x &= \int_0^c (-\log S(x)) \cdot (-S'(x))\;\mathrm{d}x \\
&= \int_0^c \frac{\mathrm{d}}{\mathrm{d}\,S(x)}\Big(S(x)\log S(x) - S(x)\Big) \cdot (S'(x))\;\mathrm{d}x \\
&= \int_0^c \frac{\mathrm{d}}{\mathrm{d}x}\Big(S(x)\log S(x) - S(x)\Big)\;\mathrm{d}x \\
&= [\,S(x)\log S(x) - S(x)\,]_0^c \\
&= S(c)\log S(c) - S(c) - S(0)\log S(0) + S(0) \\
&= S(c)\log S(c) - S(c) - 1 \cdot \log(S(0)) + 1 \\
&= S(c)\log S(c) - S(c) + 1 \\
&= S(c)\log S(c) + F_T(\,c\,)
\end{aligned}$$

Thus, we now have

$$\begin{aligned}
E[\,H(X)\,] &= \int_{[\,0\,,\,c\,]} H(x)\,\mathrm{d}\mu_X \;=\; \int_0^c H(x)\cdot f_T(x)\,\mathrm{d}x \;+\; H(c)\cdot\mu_X(\,c\,) \\
&= S(c)\log S(c) + F_T(\,c\,) - \log S(c)\cdot(\,1 - F_T(\,c\,)\,) \\
&= S(c)\log S(c) + F_T(\,c\,) - \log S(c) + \log(S(c))\cdot F_T(\,c\,) \\
&= -\log S(c)\cdot(1 - S(c)) + F_T(\,c\,)\cdot(1 + \log S(c)) \\
&= -\log S(c)\cdot F_T(\,c\,) + F_T(\,c\,)\cdot(1 + \log S(c)) \\
&= F_T(\,c\,)\cdot(-\log S(c) + 1 + \log S(c)) \\
&= F_T(\,c\,)
\end{aligned}$$

□

References