Introduction

Over a time span of approximately nine months between 2022 and 2023, I engaged in self-study of All of Statistics from Larry Wasserman. Due to time constrains, I only devoted half an hour to an hour per day to studying. I believe that it is crucial to complete most of the exercises, as the book is consise and many details are left to the reader. In this document, you can find my solutions to most of the exercises of the book.

The first two parts of the book are well-organized and contain interesting exercises. In my opinion, the quality of the last part of the book inferior to the first two parts. This issue is particularly evident in the final section of the book. There are several issues with some of the Theorems and Algorithms presented in these chapters. Not all exercises are well-defined, which is due to missing details in the text as it tries to be a consise summary of advanced statistics. Another issue is that not all algorithms can be directly implemented due to floating point overflows. As a result of these issues, you may notice that my solutions become more sloppy in the last part of the book. I would also like to note that I typically write down the solution few details. The solutions are written in a way that I can understand them. While many mathematicians prefer more detail, I'm not one of them.

I admit that I did not fully comprehend Chapter 16, Causual Inference. I'm unsure if I will revisit this chapter and redo the exercises. Lastly, I'll acknowledge that I did look to other solutions as well from time to time. If I was stuck, I searched for (partial) solutions online.

Although this book has its issues, my overall experience was very positive. All of Statistics by Larry Wasserman is precisely what it claims to be on the cover: a concise course in statistical inference. A revised version of the book that addresses these mistakes would be beneficial. Nonetheless, it's a good book if you want to quickly become familiar with the topics, or to be used as a reference.

Part I

Probability

Chapter 1 - Probability of Finite Sample Spaces

Solution 1.1

Let $A_1 \subset A_2 \subset A_3 \subset ....$ Define $B_i = A_i \backslash \cup_{j<i} A_j$.

(a) Let $a \in A_n$. Let $i \leq n$ the smallest i such that $a \in A_n$. Then $a \in A_i \backslash \cup_{j<i} A_j = B_i$. Let $b \in B_i$ for some i. Then $b \in A_i \subset A_n$. Therefore $A_n = \cup_{i=1}^n B_i$.

(b) Take $a \in \cup_{i=1}^\infty A_i$. There is an n such $a \in A_n = \cup_{i=1}^n B_i$. So $a \in \cup_{i=1}^\infty B_i$. Let $b \in \cup_{i=1}^\infty B_i$, then $b \in B_n$ for some n. Therfore $b \in A_n$, hence $b \in \cup_{i=1}^\infty A_n$. So $\cup_{i=1}^\infty A_i = \cup_{i=1}^n B_i$.

Solution 1.2

(i) Let $A_1 = \emptyset$, $A_2 = \Omega$, then $A_1 \cap A_2 = \emptyset$ and $1 = P(\Omega) = P(A_1 \cup A_2) = P(A_1) + P(A_2) = P(\emptyset) + P(\Omega) = P(\emptyset) + 1$. Hence $P(\emptyset) = 1 - 1 = 0$.

(ii) $A \subset B$, let $A_1 = A$, $A_2 = B \backslash A$. Then $A_1 \cap A_2 = \emptyset$ and $P(A) + P(B \backslash A) = P(A_1) + P(A_2) = P(A_1 \cup A_2) = P(B)$. Because $P(X) \geq 0$ for all X, $P(A) = P(B) - P(B \backslash A) \leq P(B)$.

(iii) $A \subset \Omega$, so from (ii) $0 \leq P(A) \leq P(\Omega) = 1$.

(iv) $1 = P(\Omega) = P(A \cup \Omega \backslash A) = P(A) + P(\Omega \backslash A) = P(A) + P(A^c)$. So $P(A^c) = 1 - P(A)$.

(v) Take $A_1 = A$, $A_2 = B$, $A_i = \emptyset$ for $i > 2$.

Solution 1.3

(a) Let $b \in B_{n+1}$, then $b \in A_m$ for all $m \geq n+1 > n$, so $b \in B_n$. Hence $B_1 \supset B_2 \supset ....$ Let $c \in C_n$, then $c \in A_m$ for all $m \geq n$. So $c \in A_p$ for all $p \geq n+1 > n$, so $c \in C_{n+1}$. Hence $C_1 \subset C_2 \subset ....$

(b) $\rightarrow$, let $\omega \in \cap B_n$, then $\omega \in B_n$ for all n, hence there is an m such that for all $i \geq m$, $\omega \in A_i$. $\leftarrow$, let ω be in infinite A_i, but $\omega \notin \cap B_n$. Then there is a B_m such that $\omega \notin B_m = \cup_{i \geq m} A_m$. But then $\omega \notin A_i$ for all $i \geq m$, which contradics that there are infinite A_i containing ω.

(c) $\rightarrow$, let $\omega \in \cup C_i$, then there is an m such that $\omega \in C_m = \cap_{i \geq m} A_i$. So ω is contained in $A_m, A_{m+1}, ...$ $\leftarrow$, let $m > 0$ such that ω in $A_m, A_{m+1}, ...$ We can do this because there are only a finite number of A_i not containing ω. Then $\omega \in \cap_{i \geq m} A_i = C_m$, so $\omega \in \cup C_i$.

Solution 1.4

(a) $x \in (\cup A_i)^c$, iff $x \notin \cup A_i$, iff $x \notin A_i$ for all i, iff $x \in A_i^c$ for all i, iff $x \in \cap A_i^c$.

(b) $x \in (\cap A_i)^c$, iff $x \notin \cap A_i$, iff $x \notin A_i$ for some i, iff $x \in A_i^c$ for some i, iff $x \in \sum A_i^c$.

Solution 1.5

The sample space is $S = \{x_1 x_2 ... x_k : \exists i > n, x_i = H, x_n = H, \forall j \neq i, j \neq n, x_j = T\}$. The probability that k tosses are required is

$$\sum_{j<k-1} (1-p)^j p (1-p)^{k-j} p = (k-1)p^2(1-p)^{k-2} = \frac{k-1}{2^k}.$$

Solution 1.6

Assume there exists a uniform probability P on the discrete infinite sample space Ω. Because $\sum_{x \in \Omega} P(x) = 1$, there is a $y \in \Omega$ such that $P(y) = c > 0$. As $|\{y\}| = 1$, $1 = \sum_{x \in \Omega} P(x) = \sum_{x \in \Omega} P(y) = \sum_{x \in \Omega} c = c|\Omega|$. So Ω is a finite set. But this contradicts with the assumption that Ω is an infinite set. So there doesn't exist a uniform probability P on a discrete infinite sample space.

Solution 1.7

Define $B_n = A_n \backslash \cup_{i=1}^n A_i$. Then $B_i \cap B_j = \emptyset$ when $i \neq j$, and $\cup B_n = \cup A_n$. So $P(\cup A_n) = P(\cup B_n) = \sum P(B_n)$. As $B_n \subset A_n$, $P(B_n) \leq P(A_n)$. Therefore we have $P(\cup A_n) = \sum P(B_n) \leq \sum P(A_n)$.

Solution 1.8

Proving $P(\cap A_i) = 1$ is equivalent to proving $P((\cap A_i)^c) = P(\cup A_i^c) = 0$. Take disjoint sets $B_n = A_n^c \backslash \cup_{i<n} A_i^c$. Note that $B_n^c \subset A_n^c$, so $P(B_n^c) \leq P(A_n^c) = 1 - P(A_n) = 0$. So

$$P(\cup_{i=1}^n A_i^c) = P(\cup_{i=1}^n B_i) = \sum_{i=1}^n P(B_i) = 0.$$

Solution 1.9

1. $P(X, B) \geq 0$ and $P(B) \geq 0$, so $P(X|B) = P(X, B)/P(B) \geq 0$.
2. $P(\Omega|B) = P(\Omega \cap B)/P(B) = P(B)/P(B) = 1$.
3. Let $A_1, A_2, ...$ be disjoint, then $A_1 \cap B, A_2 \cap B, ...$ are disjoint, and

$$P(\cup_{i=1}^\infty A_i|B) = \frac{P(\cup_{i=1}^\infty A_i \cap B)}{P(B)} = \sum_{i=1}^\infty \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^\infty P(A_i|B).$$

Solution 1.10

When we always pick door 1, the sample space is $\Omega = \{(1,2), (1,3), (2,3), (3,2)\}$, with probabilities $P(1,2) = 1/6$, $P(1,3) = 1/6$, $P(2,3) = 1/3$, and $P(3,2) = 1/3$. We have

$$\begin{aligned}
P(\omega_1 = 1|\omega_2 = 2) &= \frac{P(\omega_1 = 1, \omega_2 = 2)}{P(\omega_2 = 2)} = \frac{1}{3}, \\
P(\omega_1 = 3|\omega_2 = 2) &= \frac{P(\omega_1 = 2, \omega_2 = 2)}{P(\omega_2 = 2)} = \frac{2}{3}, \\
P(\omega_1 = 1|\omega_2 = 3) &= \frac{P(\omega_1 = 1, \omega_2 = 2)}{P(\omega_2 = 3)} = \frac{1}{3}, \\
P(\omega_2 = 3|\omega_2 = 3) &= \frac{P(\omega_1 = 3, \omega_2 = 2)}{P(\omega_2 = 3)} = \frac{2}{3}.
\end{aligned}$$

We conclude that it's better to switch to doors.

Solution 1.11

Suppose $A \perp B$, i.e. $P(AB) = P(A)P(B)$, then

$$P(A^cB^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(A^c)P(B^c).$$

Solution 1.12

I think this question is not well-defined and similar to the boy-girl paradox. We could intrepid the questions in multiple ways and get both right answers $\frac{1}{2}$ and $\frac{1}{3}$.

Solution 1.13

(a) $\Omega = \{HH^kT, TT^kH : k \geq 0\}$.

(b) $P(HHT, TTH) = 2\frac{1}{2^3} = \frac{1}{4}$.

Solution 1.14

Let $A \subset \Omega$ such that $P(A) = 0$ or $P(A) = 1$.

(a) If $P(A) = 0$, $0 \leq P(AB) \leq P(A) = 0$, so $P(AB) = 0 = P(A)P(B)$. If $P(A) = 1$, then $P(A^c) = 0$, hence $P(A^cB^c) = P(A^c)P(B^c)$. So $A^c \perp B^c$, and by exercise 2.11, $A \perp B$.

(b) Let $A \perp A$, then $P(A) = P(A \cap A) = P(A)^2$. But this is only possible if $P(A) = 0$ or $P(A) = 1$.

Solution 1.15

We have the same problem as exercise 2.12. This question is not well-defined and similar to the boy-girl paradox. Depending on how you intrepid the question you can get different answers.

Solution 1.16

Follows almost from the definition. If $A \perp B$, then $P(AB) = P(A)P(B)$, and

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Solution 1.17

$P(ABC) = P(A|BC)P(BC) = P(A|BC)P(B|C)P(C)$.

Solution 1.18

Let $A_1, A_2, ..., A_k$ disjoint subsets such that $\cup A_i = \Omega$. Let $P(B) > 0$ such that $P(A_1|B) < P(A_1)$. We show that there is an A_i such that $P(A_i|B) > P(A_i)$. Suppose the oppostite, i.e. for all i, $P(A_i|B) \leq P(A_i)$. Then

$$1 = P(\cup_{i=1}^k A_i|B) = \sum_{i=1}^k P(A_i|B) = P(A_1|B) + \sum_{i=2}^k P(A_i|B) < P(A_1) + \sum_{i=2}^k P(A_i) = P(\cup_{i=1}^k A_i) = 1,$$

which is a contradiction. Therefore, there exists an A_i such that $P(A_i|B) > P(A_i)$.

Solution 1.19

$$P(W|I) = \frac{P(I|W)P(W)}{P(I)} = \frac{P(I|W)P(W)}{P(I|W)P(W) + P(I|M)P(M) + P(I|L)P(L)} \approx 0.58.$$

Solution 1.20

(a)

$$P(C_i|H) = \frac{P(H|C_i)P(C_i)}{\sum_i P(H|C_i)P(C_i)} = \frac{p_i}{\sum_i p_i}.$$

(b)

$$P(H_2|H_1) = \frac{P(H_1H_2)}{P(H_1)} = \frac{\sum_i P(H_1H_2|C_i)}{\sum_j P(H_1|C_j)} = \frac{\sum_i p_i^2}{\sum_j p_j}.$$

(c)

$$P(C_i|B_4) = \frac{P(B_4|C_i)P(C_i)}{P(B_4)} = \frac{P(B_4|C_i)P(C_i)}{\sum_j P(B_4|C_j)P(C_j)} = \frac{(1-p_i)^3 p_i}{\sum_j (1-p_j)^3 p_j}.$$

TODO: fill in values for p_1, p_2, p_3, and p_4.

Solution 1.21

See code.

Solution 1.22

See code.

Solution 1.23

Take $\Omega = \{1, 2\}$ with $P(\{1\}) = \frac{1}{2} = P(\{2\})$. For the dependent events, take $A = \{1\}$, $B = \{1\}$. Then $AB = \{1\}$ and $P(AB) = \frac{1}{2} \neq \frac{1}{4} = P(A)P(B)$. See code for simulations.

Chapter 2 - Random Variables

Solution 2.1

Let $x_1^+ > x_2^+ > ...$ and $x_1^- < x_2^- < ...$ such that $x_i^+ \downarrow x$ and $x_i^- \uparrow x$. Let $A_i = [x_i^-, x_i^+]$ for all $i > 0$. Note that $A_{i+1} \subset A_i$ for all $i > 0$ and $A_i \to \{x\}$ as $i \to \infty$. By theorem 1.8 (Continuity of Probability)

$$P(x) = \lim_{i\to\infty} P(A_i) = \lim_{i\to\infty} (F(x_i^+) - F(x_i^-)) = F(x^+) - F(x^-).$$

Solution 2.1 - Alternative

By lemma 2.15.1 $P(x) = F(x) - F(x^-)$. By Theorem 2.8.iii F is right continuous, i.e. $F(x) = F(x^+)$, so $P(x) = F(x^+) - F(x^-)$.

Solution 2.2

The CDF is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 2, \\ \frac{1}{10} & \text{if } 2 \leq x < 3, \\ \frac{2}{10} & \text{if } 3 \leq x < 5, \\ 1 & \text{if } 5 \leq x \end{cases}$$

So $P(2 \leq X \leq 4.8) = F(4.8) - F(2^-) = \frac{2}{10}$.

Solution 2.3

1. Let $x_1^- < x_2^- < ...$ such that $x_i^- \to x$ and $A_i = (x_i^-, x]$. Then $A_i \supset A_{i+1}$ for all i and $A_i \to \{x\}$. By the continuity of the probability function

$$P(x) = \lim_{i\to\infty} P(A_i) = \lim_{i\to\infty} F(x) - F(x_i^-) = F(x) - F(x^-).$$

2. $P(x < X \leq y) = P(X \leq y) - P(X \leq x) = F(y) - F(x)$.
3. $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$.
4. Follows directly from continuity of the probability function. $P(x < X) = P(x \leq X)$ if P is continuous.

Solution 2.4

(a)

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x/4 & \text{if } 0 < x < 1 \\ 1/4 & \text{if } 1 \leq x \leq 3 \\ 1/8(3x-7) & \text{if } 3 < x < 5 \\ 1 & \text{if } 5 \leq x \end{cases}$$

(b) For $Y = 1/X$, $F_Y(y) = P(1/X < y) = P(1/y < X) = 1 - P(X < 1/y)$, so

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 1/5 \\ 1 - 1/8(3/y - 7) & \text{if } 1/5 \leq y < 1/3 \\ 3/4 & \text{if } 1/3 \leq y \leq 1 \\ 1 - 1/(4y) & \text{if } 1 < y \end{cases}$$

Solution 2.5

Almost straight from the definition. As X and Y are discrete random variables:

$\rightarrow$) $f(x,y) = P(X \in \{x\}, Y \in \{y\}) = P(X \in \{x\})P(Y \in \{y\}) = f(x)f(y)$.

$\leftarrow$) $P(X \in A, Y \in B) = \sum_{(x,y)\in A\times B} f(x,y) = \sum_{x\in A}\sum_{y\in B} f(x)f(y) = P(X \in A)P(Y \in B)$.

Solution 2.6

$Y = I_A(X)$ takes the value 1 if $X \in A$ and 0 if $X \notin A$. So $P(Y = 0) = 1 - F_X(A)$ and $P(Y = 1) = F_X(A)$. This gives the CDF

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - F_X(A) & \text{if } 0 \leq y < 1 \\ 1 & \text{if } 1 \leq y \end{cases}$$

Solution 2.7

Let $Z = \min(X, Y)$, $X, Y \sim \text{Uniform}(0,1)$, then

$$\begin{aligned} F_Z(z) &= P(Z < z) \\ &= 1 - P(Z > z) \\ &= 1 - P(X > z \text{ and } Y > z) \\ &= 1 - P(X > z)P(Y > z) \\ &= 1 - (1 - F_X(z))(1 - F_Y(z)) = 1 - (1-z)^2. \end{aligned}$$

Taking the derivative gives the probability distribution function

$$f_Z(z) = 2(1-z).$$

Solution 2.8

Let X be a random variable with CDF F. Define $X^+ = \max(0, X)$. Note that $P(X^+ < 0) = 0$ and $P(X^+ = 0) = P(X \leq 0) = F(0)$. Therefore

$$F_+(x) = \begin{cases} 0 & \text{if } x < 0 \\ F(x) & \text{if } x \geq 0 \end{cases}$$

Solution 2.9

Let $X \sim \text{Exp}(\beta)$. We have $f(x) = \beta^{-1}\exp(-\beta^{-1}x)$. Then

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\beta}\exp\left(-\frac{t}{\beta}\right)dt = 1 - \exp\left(-\frac{x}{\beta}\right).$$

Solving x for $q = F_X(x) = 1 - \exp(-\beta^{-1}x)$ yields $x = F_X^{-1}(q) = -\beta\log(1-q)$.

Solution 2.10

Follows almost from the definition.

$$\begin{aligned} P(g(X) \in A, h(Y) \in B) &= P(X \in g^{-1}(A), Y \in h^{-1}(B)) \\ &= P(X \in g^{-1}(A))P(Y \in h^{-1}(B)) = P(g(X) \in A)P(h(Y) \in B). \end{aligned}$$

Solution 2.11

(a) $P(X = 1, Y = 0) = p$, but $P(X = 1) = p$ and $P(Y = 0) = p$, so $P(X = 1)P(Y = 1) = p^2 \neq p$ (assuming $p > 0$). Therefore X and Y are dependent.

(b) Let $N \sim \text{Poisson}(\lambda)$, flip N coins and let X be the number of heads, Y the number of tails. We have

$$\begin{aligned} P(X = i, Y = j) &= f_\lambda(N)\binom{N}{i}p^i(1-p)^{N-i} \\ &= f_\lambda(i+j)\binom{i+j}{i}p^i(1-p)^j \\ &= e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!}\frac{(i+j)!}{i!j!}p^i(1-p)^j \\ &= e^{-\lambda}\frac{\lambda^i p^i}{i!}\frac{\lambda^j(1-p)^j}{j!} = g(i)h(j), \end{aligned}$$

where $g(i) = e^{-\lambda}\lambda^i p^i/i!$ and $h(j) = \lambda^j(1-p)^j/j!$. By Theorem 2.33 X and Y are independent.

Here follows a remark, as the results of this exercise is counter intuitive (for me). Suppose you toss a coin 50 times. Let X be the number of heads and Y the number of tails. If $X = 25$, what is Y? Easy, $Y = 50 - X = 25$. If $X = 4$, $Y = 50 - 4 = 46$. Or $X = 48$, then $Y = 50 - 48 = 2$. X and Y are entirely correlated. If I know X, I also know Y.

Now let me toss a coin every time I receive an email in my mailbox. I count the number of heads and tails during the day. Let X be the number of heads and Y the number of tails of that day. I receive around 50 mails a day, and the exact number of mails per day follows a Poisson distribution with $\lambda = 50$. After one day I count $X = 25$ heads, how many tails do I expect to have counted? Intuitively I expect $Y = 25$ tails, which is correct. But suppose now that $X = 4$? Because X is a low number, I expect that I have not received many mails that day, and therefore flipped few coins, so $E(E(Y|X = 4)) = 4$ would be an intuitive guess. Similar, if $X = 48$, I probably received many mails, flipped a lot of coins, so $E(E(Y|X = 48)) = 48$ seems reasonable.

But this is not the case. As we saw in the exercise, X and Y are independent. Therefore

$$E(Y|X = x) = \int yf(y|x)dy = \int yf(y)dy = E(Y),$$

which shows that $E(E(Y|X = x)) = E(Y) = \frac{\lambda}{2} = 25$. In particular, $E(E(Y|X = 25)) = E(E(Y|X = 4)) = E(E(Y|X = 48)) = 25$. The number of heads will give you absolutely no information about the number of tails that occured during the day. This is an extreme case where a small change, going from $N = 50$ to $N \sim \text{Poisson}(50)$, turns entirely dependent random variables into completely independent random variables.

Solution 2.12

Let $f(x,y) = g(x)h(y)$ for all x, y. Note

$$1 = \lim_{x,y\to\infty} F(x,y) = \int f(x,y)dxdy = \int g(x)dx \int h(y)dy = A_g A_h.$$

Now

$$f_X(x) = \int f(x,y)dy = \int g(x)h(y)dy = A_h g(x),$$
$$f_Y(y) = \int f(x,y)dx = \int g(x)h(y)dx = A_g h(y).$$

We have $f(x,y) = g(x)h(y) = A_g A_h g(x)h(y) = (A_h g(x))(A_g h(y)) = f_X(x)f_Y(y)$, which shows that X and Y are independent.

Solution 2.13

Let $X \sim \text{Normal}(0,1)$ and $Y = \exp(X)$.

(a) We first calculate F_Y,

$$F_Y(y) = P(Y < y) = P(X < \log(y)) = \Phi(\log(y)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\log(y)} \exp\left(-\frac{1}{2}t^2\right) dt$$

By the fundamental theorem of calculus

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\log(y)^2\right).$$

(b) See code.

Solution 2.14

Let $A_r = \{(x,y) : x^2 + y^2 \le r^2\}$ be the circle with radius r. We have $F_R(r) = \text{Opp}(A_r)/\text{Opp}(A_1) = r^2\pi/\pi = r^2$. So $f_R(r) = 2r$.

Solution 2.15

Let $X \sim F$ and F continuous, strictly increasing, i.e., $x_1 < x_2$, then $F(x_1) < F(x_2)$. Let $Y = F(X)$.

(a) $F_Y(y) = P(Y \le y) = P(F(X) \le y) = P(X \le F^{-1}(y)) = F(F^{-1}(y)) = y$. So $Y \sim \text{Uniform}(0,1)$.

(b) Let $U \sim \text{Uniform}(0,1)$ and $X = F^{-1}(U)$. Then $F_X(x) = P(X \le x) = P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x)$. So $X \sim F$.

(c) Take $F(x) = 1 - \exp(-\frac{x}{\beta}) = u$, then $F^{-1}(u) = -\beta\log(1-u)$. So if $U \sim \text{Uniform}(0,1)$, then $X = F^{-1}(U) = -\beta\log(1-U) \sim \text{Exp}(\beta)$. See code for an implementation.

Solution 2.16

$X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$. Let $n = X + Y$.

$$P(X = x|X+Y=n) \propto P(X=x, X+Y=n) = e^{-\lambda}\frac{\lambda^x}{x!}e^{-\mu}\frac{\mu^{n-x}}{(n-x)!} \propto \frac{\lambda^x\mu^{n-x}}{x!(n-x!)} \propto \binom{n}{x}\theta^x(1-\theta)^x,$$

where

$$\theta = \frac{\lambda}{\lambda+\mu}.$$

So $X|X+Y \sim \text{Binomial}(n,\theta)$.

Solution 2.17

We have

$$f_Y\left(\frac{1}{2}\right) = \int_0^1 f\left(x, \frac{1}{2}\right) dx = \int_0^1 c\left(x + \frac{1}{4}\right) dx = \frac{3}{4}c.$$

Using the definition of the conditional probability density function

$$P\left(X < \frac{1}{2} \middle| Y = \frac{1}{2}\right) = \int_0^{\frac{1}{2}} f(x|y = 1/2)dx = \int_0^{\frac{1}{2}} \frac{f(x, 1/2)}{f(1/2)}dx = \frac{\frac{1}{4}c}{\frac{3}{4}c} = \frac{1}{3}.$$

Solution 2.18

Note that $X = 4Z - 3$, where $Z \sim \text{Normal}(0, 1)$.

(a) $P(X < 7) = P(Z < 1) = \Phi(1) \approx 0.84$.

(b) $P(X > 2) = 1 - P(X \leq 2) = 1 - P(Z \leq -5/4) = 1 - \Phi(-5/4) \approx 0.89$.

(c) $0.05 = P(X > x) = 1 - P(X < x) = 1 - P(Z < (x - 3)/4) = 1 - \Phi((x - 3)/4)$ if and only if $x = 4\Phi^{-1}(0.95) + 3 \approx 9.58$.

(d) $P(0 \leq X < 4) = P(-3/4 \leq Z < 1/4) = \Phi(1/4) - \Phi(-3/4) \approx 0.37$.

(e) $P(|X| > |x|) = 0.05$ if and only if $0.025 = P(X < -x) = P(Z < -(x - 3)/4) = \Phi(-(x + 3)/4)$, so $x = \pm(4\Phi^{-1}(0.025) + 3) \approx \pm 4.84$.

Solution 2.19

Let r be strictly increasing. Let $s = r^{-1}$ and $Y = r(X)$. Note that $F_Y(y) = P(Y < y) = P(X < r^{-1}(y)) = F_X(s(y))$. So we have

$$f_Y(y) = \frac{dF_X(s(y))}{dy} = \frac{dF_x(s(y))}{ds(y)}\frac{ds(y)}{dy} = f_X(s(y))\left|\frac{ds(y)}{dy}\right|.$$

If r would be strictly decreasing the same calculation holds, but with a minus sign, as $F_Y(y) = 1 - F_X(s(y))$.

Solution 2.20

Let $Z_1 = X - Y$, $Z_2 = X/Y$. This is an exercise in drawing squares.

(a) If $z \leq 0$, then $P(Z_1 < z) = \frac{1}{2}(1 - z)^2$. If $z \geq 0$, then $P(Z_1 < z) = \frac{1}{2}(1 + z)^2$. So

$$f_{Z_1}(x) = \begin{cases} 1 - z, & \text{if } z \leq 0, \\ 1 + z, & \text{if } z > 0. \end{cases}$$

(b) Define $A_z = F_{Z_2}(z)$. If $0 \leq z \leq 1$, $|A_z| = z/2$. If $1 < z$, $|A_z| = 1 - \frac{1}{2z}$.

$$f_{Z_1}(x) = \begin{cases} \frac{1}{2}, & \text{if } 0 \leq z \leq 1, \\ \\ \frac{1}{2z^2}, & \text{if } 1 < z. \end{cases}$$

Chapter 3 - Expectation

Solution 3.1

Let X_n a random variable representing the amount of money after n turns. We set $X_0 = c$. We have $E(X_n|X_{n-1}) = \frac{1}{2}(2X_{n-1} + \frac{1}{2}X_{n-1}) = \frac{5}{4}X_{n-1}$. Using the tower property of conditional expectation we have

$$E(X_n) = E(E(X_n|X_{n-1})) = \frac{5}{4}E(E(X_{n-1}|X_{n-2})) = ... = \left(\frac{5}{4}\right)^n E(X_0) = \left(\frac{5}{4}\right)^n c.$$

Solution 3.2

→) TODO: This is a hand-wave proof. I should look for a better proof if I have some time. Suppose $V(X) = 0$, then

$$\int_{-\infty}^{\infty} (x - \mu_X)^2 dF(x) = 0.$$

But $(x - \mu_X)^2 \geq 0$ and continuous and F is right continuous, so for each x we must have $f(x) = 0$ or $(x - \mu_X)^2 = 0$. $f(x)$ cannot be zero everywhere, so there must be an x such that $(x - \mu_X)^2 = 0$, which is only once at $x = \mu_X$. In other words $P(X = \mu_X) = f(\mu_X) = 1$ and for all $x \neq \mu_X$ we must have $f(x) = 0$. Take $c = \mu_X$.

←) If $P(X = c) = 1$, then $E(X) = c$ and $E(X^2) = c^2$, and $V(X) = E(X^2) - E(X)^2 = c^2 - c^2 = 0$.

Solution 3.3

Let $X_1, X_2, ..., X_n \sim$ Uniform$(0, 1)$. Define $Y_n = \max(X_1, X_2, ..., X_n)$.

$$P(Y_n < y) = \prod_{i=1}^{n} P(X_i < y) = y^n.$$

So $f(y) = ny^{n-1}$ and

$$E(Y_n) = \int_0^1 ny^n dy = \frac{n}{n+1}.$$

Solution 3.4

Let $X_0 = 0$. Note that for $n > 0$, $X_n = \sum_{i=1}^{n}(1 - 2B_i)$, where $B_i \sim$ Bernoulli(p). We have $E(X_n) = n - 2\sum_i E(B_i) = n(1 - 2p)$ and $V(X_n) = 4\sum_i V(B_i) = 4np(1 - p)$.

Solution 3.5

The probability density function is $f(X = i) = p^{i-1}(1 - p) = p^i$, as $p = \frac{1}{2}$. We have

$$E(X) = \sum_{i=0}^{\infty} if(X = i) = \sum_{i=0}^{\infty} ip^i = \left(\frac{1}{1-p}\right)' = \frac{p}{(1-p)^2} = 2.$$

Solution 3.6

Write out the definition and do required book keeping.

$$E_Y(Y) = \sum_y yP(Y = y) = \sum_y yP(r(X) = y) = \sum_y yP(X \in r^{-1}(y)) = \sum_y \sum_{x \in r^{-1}(y)} yP(X = x)$$
$$= \sum_y \sum_{x \in r^{-1}(y)} r(x)P(X = x) = \sum_x r(x)P(X = x) = E_X(r(X)).$$

Solution 3.7

We first prove a lemma.

$$xP(X > x) = x \int_x^\infty f(t)dt \leq \int_x^\infty tf(t)dt = E(X) - \int_{-\infty}^x tf(t)dt \to 0,$$

as $x \to \infty$. So $\lim_{x\to\infty} x(1 - F(x)) = \lim_{x\to\infty} xP(X > x) = 0$.

For the solution, use integration by parts.

$$\int_0^\infty P(X > x)dx = \int_0^\infty (1 - F(x))dx = x(1 - F(x))|_0^\infty + \int_0^\infty xf(x)dx = E(X).$$

Solution 3.8

Let $X_1, X_2, ..., X_n$ i.i.d. with $\mu = E(X_i)$ and $\sigma^2 = V(X_i)$. Let the sample mean be $\overline{X}_n = \frac{1}{n}\sum X_i$. We have

$$\begin{aligned} E(\overline{X}_n) &= \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{n}{n}\mu = \mu, \\ V(\overline{X}_n) &= \frac{1}{n^2}\sum_{i=1}^n V(X_i) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Define the sample variance by $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Note that

$$\begin{aligned} E(X_i^2) &= V(X_i) + E(X_i)^2 = \sigma^2 + \mu^2, \\ E(\overline{X}_n^2) &= V(\overline{X}_n) + E(\overline{X}_n)^2 = \frac{\sigma^2}{n} + \mu^2, \\ E(X_i\overline{X}_n) &= \frac{1}{n}\sum_{j=1}^n E(X_iX_j) = \frac{1}{n}\left(E(X_i^2) + \sum_{i\neq j} E(X_i)E(X_j)\right) = \mu^2 + \frac{\sigma^2}{n}. \end{aligned}$$

We calculate

$$\begin{aligned} E(S_n^2) &= \frac{1}{n-1}\sum_{i=1}^n \left(E(X_i^2) - 2E(X_i\overline{X}_n) + E(\overline{X}_n^2)\right) \\ &= \frac{1}{n-1}\sum_{i=1}^n \left(\sigma^2 + \mu^2 - 2\frac{\sigma^2}{n} - 2\mu^2 + \mu^2 + \frac{\sigma^2}{n}\right) \\ &= \frac{1}{n-1}\sum_{i=1}^n \frac{n-1}{n}\sigma^2 = \sigma^2. \end{aligned}$$

Solution 3.9

See code. Note that the Cauchy distribution doesn't have moments. Therefore, we cannot "sample" the mean of the Cauchy distribution.

Solution 3.10

Let $X \sim$ Normal$(0, 1)$ and let $Y = e^X$. Note that for the moment generating function for X we have $\exp(\frac{t^2}{2}) = \phi_X(t) = E(e^{tX}) = E(Y^t)$. So $E(Y) = \phi_X(1) = \sqrt{e}$, and $E(Y^2) = \phi_X(2) = e^2$ so that $V(Y) = E(Y^2) - E(Y)^2 = e^2 - e = e(e - 1)$.

Solution 3.11

(a) See solution 3.4 with $p = \frac{1}{2}$. $E(X_n) = 0$ and $V(X_n) = n$.

(b) See code. Note that $V(X_n) = n \to \infty$ as $n \to \infty$. So, although $E(X_n) = 0$, the variance increases as n increases, and the random walks will be different from each other.

Solution 3.12

We calculate the expected value and variance for all distributions in section 3.4.

(a) Point mass distribution. $E(X) = \sum xp(x) = ap(a) = a$ and $V(X) = E(X^2) - E(X)^2 = a^2 - a^2 = 0$.

(b) Bernoulli, $X \sim \text{Bernoulli}(p)$. $E(X) = 1p + 0(1-p) = p$. Note that $E(X^2) = 1^2 p + 0^2(1-p) = p$, so $V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$.

(c) Binomial, $X \sim \text{Binomial}(n, p)$. Write $X = \sum_{i=1}^n X_i$, where $X_i \sim \text{Bernoulli}(p)$. Then $E(X) = E(\sum X_i) = np$ and $V(X) = V(\sum X_i) = np(1-p)$.

(d) Geometric, $X \sim \text{Geometric}(p)$.

$$E(X) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = p\left(-\sum_{x=0}^{\infty}(1-p)^x\right)' = p\left(-\frac{1}{p}\right)' = \frac{p}{p^2} = \frac{1}{p},$$

and

$$E(X(X-1)) = \sum_{x=2}^{\infty} x(x-1)p(1-p)^{x-1} = p(1-p)\left(\sum_{x=0}^{\infty}(1-p)^x\right)'' = p(1-p)\left(\frac{1}{p}\right)'' = \frac{2(1-p)}{p^2}.$$

So $E(X^2) = E(X(X-1)) + E(X) = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2-p}{p^2}$. Such that $V(X^2) = E(X^2) - E(X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$.

(e) Poisson, $X \sim \text{Poisson}(\lambda)$.

$$E(X) = \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{x=0}\frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Note that

$$E(X(X-1)) = \sum_{x=0}^{\infty} x(x-1)\frac{\lambda^x}{x!}e^{-\lambda} = \lambda^2 e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{x!} = \lambda^2.$$

So $E(X^2) - E(X)^2 = E((X-1)X) + E(X) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

(f) Uniform, $X \sim \text{Uniform}(a, b)$. First take $U \sim \text{Uniform}(0, 1)$, then

$$E(U) = \int_0^1 xf(x)dx = \frac{1}{2}x^2\Big|_0^1 = \frac{1}{2},$$
$$E(U^2) = \int_0^1 x^2 f(x)dx = \frac{1}{3}x^3\Big|_0^1 = \frac{1}{3}.$$

So

$$E(X) = E((b-a)U + a) = (b-a)\frac{1}{2} + a = \frac{1}{2}(a+b),$$
$$V(X) = V((b-a)U + a) = (b-a)^2 V(U) = (b-a)^2(E(U^2) - E(U)) = \frac{1}{12}(b-a)^2.$$

(g) Normal, $X \sim \text{Normal}(\mu, \sigma^2)$. First take $Z \sim \text{Normal}(0, 1)$. Because $xf_z(x) = x\phi(x)$ is anti-symmetric $E(Z) = 0$. To calculate $V(Z) = E(Z^2)$ we need the following lemma.

Let ϕ be monotonic decreasing function with finite moments on some interval (a, ∞) such that $\phi(z) \to 0$ if $z \to \infty$, then $z\phi(z) \to 0$ if $z \to \infty$. Indeed

$$\int_{z/2}^{\infty} \phi(t)dt \geq \int_{z/2}^{z} \phi(t)dt \geq \int_{z/2}^{z} \phi(z)dt = \frac{1}{2}z\phi(z) > z\phi(z),$$

when z is large enough such that $\phi(t)$ is monotonic decreasing. As $\phi \to 0$ as $z \to \infty$, we have

$$z\phi(z) < \int_{\frac{z}{2}}^{\infty} \phi(t)dt \to 0.$$

So $z\phi(z) \to 0$ as $z \to \infty$.

Now we calculate

$$E(Z^2) = \int_{-\infty}^{\infty} z^2\phi(z)dz = 2\int_{0}^{\infty} z^2\phi(z)dz = [\Phi(z) - z\phi(z)]_0^{\infty} = 2\lim_{z\to\infty} \Phi(z) - 2\Phi(0) = 2 - 1 = 1$$

Where we use the lemma $z\phi(z) \to 0$ if $z \to \infty$. So $V(Z) = E(Z^2) = 1$.

Now $X = \sigma Z + \mu$, so $E(X) = \sigma E(Z) + \mu = \mu$ and $V(X) = \sigma^2 V(Z) = \sigma^2$.

(h) Exponential, $X \sim \text{Exp}(\beta)$. Using integration by parts

$$\begin{aligned} E(X) &= \int_0^{\infty} \frac{x}{\beta}\exp\left(-\frac{x}{\beta}\right) dx \\ &= -\, x\exp\left(-\frac{x}{\beta}\right)\Big|_0^{\infty} + \frac{1}{\beta}\int_0^{\infty} \beta\exp\left(-\frac{x}{\beta}\right) dx \\ &= -\beta\exp\left(-\frac{x}{\beta}\right)\Big|_0^{\infty} = \beta. \end{aligned}$$

And

$$E(X^2) = \int_0^{\infty} \frac{x^2}{\beta}\exp\left(-\frac{x}{\beta}\right) dx = -\, x\exp\left(-\frac{x}{\beta}\right)\Big|_0^{\infty} + 2\int_0^{\infty} x\exp\left(-\frac{x}{\beta}\right) dx = 2\beta E(X) = 2\beta^2.$$

So $V(X) = E(X^2) - E(X)^2 = 2\beta^2 - \beta^2 = \beta^2$.

(i) Gamma, $X \sim \text{Gamma}(\alpha, \beta)$.

$$\begin{aligned} E(X) &= \int_0^{\infty} \frac{\beta^{\alpha}x^{\alpha}}{\Gamma(\alpha)}\exp(-\beta x)dx \\ &= \frac{\alpha}{\beta}\int_0^{\infty} \frac{\beta^{\alpha+1}x^{(\alpha+1)-1}}{\Gamma(\alpha+1)}\exp(-\beta x)dx \\ &= \frac{\alpha}{\beta}\int_0^{\infty} f(x;\alpha+1,\beta)dx = \frac{\alpha}{\beta}. \end{aligned}$$

And

$$\begin{aligned} E(X^2) &= \int_0^{\infty} \frac{\beta^{\alpha}x^{\alpha+1}}{\Gamma(\alpha)}\exp(-\beta x)dx \\ &= \frac{\alpha(\alpha+1)}{\beta^2}\int_0^{\infty} \frac{\beta^{\alpha+2}x^{(\alpha+2)-1}}{\Gamma(\alpha+2)}\exp(-\beta x)dx \\ &= \frac{\alpha(\alpha+1)}{\beta^2}\int_0^{\infty} f(x;\alpha+2,\beta)dx = \frac{\alpha(\alpha+1)}{\beta^2}. \end{aligned}$$

So $V(X) = E(X^2) - E(X)^2 = \frac{\alpha}{\beta^2}$.

(j) Beta, $X \sim \text{Beta}(\alpha, \beta)$.

$$\begin{aligned} E(X) &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1}dx \\ &= \frac{\alpha}{\alpha+\beta}\int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)} x^{(\alpha+1)-1}(1-x)^{\beta-1}dx = \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

For the variance

$$\begin{aligned} E(X^2) &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha+1}(1-x)^{\beta-1}dx \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\int_0^1 \frac{\Gamma(\alpha+2+\beta)}{\Gamma(\alpha+2)\Gamma(\beta)} x^{(\alpha+2)-1}(1-x)^{\beta-1}dx = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}, \end{aligned}$$

so that

$$V(X) = E(X^2) - E(X)^2 = \frac{\alpha}{\alpha+\beta} + \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

(k) Student-t, $X \sim t(\nu)$. For $\nu > 1$, $xf(x;\nu)$ is odd, so $E(X) = 0$. Let $\nu > 2$, because $E(X) = 0$ we have $V(X) = E(X^2)$. The pdf is given by

$$f(x;\nu) = \frac{1}{\sqrt{\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Note that we have the equality

$$\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

We'll rewrite $E(X^2)$ into this integral using substition $y = \left(1 + \frac{x^2}{\nu}\right)^{-1}$, such that $x = \sqrt{\nu}y^{-\frac{1}{2}}(1-y)^{\frac{1}{2}}$ and $xdx = -\frac{\nu}{2}y^2dy$. The variance is calculated with the following tedious calculation

$$\begin{aligned} E(X^2) &= \frac{1}{\sqrt{\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \int_{-\infty}^{\infty} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx \\ &= \frac{2}{\sqrt{\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \int_{0}^{\infty} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx \\ &= \frac{2}{\sqrt{\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \int_{1}^{0} \sqrt{\nu}y^{-\frac{1}{2}}(1-y)^{\frac{1}{2}}y^{\frac{\nu+1}{2}}(-1)\frac{\nu}{2}y^2dy \\ &= \nu \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \int_{0}^{1} y^{\frac{\nu-2}{2}-1}(1-y)^{\frac{3}{2}-1}dy \\ &= \nu \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \frac{\Gamma(\frac{\nu-2}{2})\Gamma(\frac{3}{2})}{\Gamma(\frac{\nu+1}{2})} \\ &= \frac{\nu}{2}\frac{1}{\frac{\nu}{2}-1} = \frac{\nu}{\nu-2}. \end{aligned}$$

(ℓ) χ^2 distribution, $X \sim \chi^2_p$. We calculate all moments of X.

$$\begin{aligned} E(X^k) &= \frac{1}{2^{p/2}\Gamma(p/2)} \int_0^\infty x^{k+p/2-1} e^{-x/2} dx \\ &= \frac{2}{2^{p/2}\Gamma(p/2)} \int_0^\infty (2y)^{k+p/2-1} e^{-y} dy \\ &= \frac{2^{k+p/2}}{2^{p/2}\Gamma(p/2)} \int_0^\infty y^{k+p/2-1} e^{-y} dy = \frac{2^k}{\Gamma(p/2)} \Gamma\left(k + \frac{p}{2}\right). \end{aligned}$$

So $E(X) = p$, $E(X^2) = p(p+1)$, and $V(X) = E(X^2) - E(X)^2 = p(p+1) - p^2 = p$.

(m) Multinomial is explained in the book.

(n) Multi-Normal, $X \sim \text{Normal}(\mu, \Sigma)$. By Theorem 2.44, if $Z \sim \text{Normal}(0, I)$, then $E(Z) = 0$ and $V(Z) = I$. By Theorem 2.43, $X = \Sigma^{\frac{1}{2}}(Z - \mu)$. By lemma 3.21, $E(X) = E(\Sigma^{1/2}Z + \mu) = \mu$ and $V(X) = V(\Sigma^{1/2}Z + \mu) = \Sigma^{1/2}V(Z)\Sigma^{1/2} = \Sigma$.

Solution 3.13

Let $X = BU_1 + (1-B)U_2$ a random variable, where $B \sim \text{Bernoulli}(\frac{1}{2})$, $U_1 \sim \text{Uniform}(0,1)$, and $U_2 \sim \text{Uniform}(3,4)$.

(a) $E(X) = E(B)E(U_1) + E(1-B)E(U_2) = 2$.

(b) Note that $E(U_1^2) = \frac{1}{3}$ and $E(U_2^2) = \frac{37}{3}$. We have

$$\begin{aligned} E(X^2) &= E(B^2U_1^2 + BU_1(1-B)U_2 + (1-B)^2U_2^2) \\ &= E(B^2U_1^2) + E((1-B)^2U_2^2) \\ &= E(B^2)E(U_1^2) + E((1-B)^2)E(U_2^2) \\ &= \frac{1}{2\cdot 3} + \frac{37}{2\cdot 3} = \frac{19}{3}, \end{aligned}$$

such that $V(X) = E(X^2) - E(X)^2 = \frac{7}{3}$.

Solution 3.14

Let $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_m$ be random variables, and $a_1, a_2, ..., a_m$ and $b_1, b_2, ..., b_n$ be constants. We have

$$\begin{aligned} \text{Cov}\left(\sum_{i=1}^m a_iX_i, \sum_{j=1}^n b_jY_j\right) &= E\left(\sum_{i=1}^m a_iX_i \sum_{j=1}^n b_jY_j\right) - E\left(\sum_{i=1}^m a_iX_i\right) E\left(\sum_{j=1}^n a_jY_j\right) \\ &= \sum_{i=1}^m \sum_{j=1}^n a_ib_j \left(E(X_iY_j) - E(X_i)E(Y_j)\right) = \sum_{i=1}^m \sum_{j=1}^n a_ib_j \text{Cov}(X_i, Y_j). \end{aligned}$$

Solution 3.15

We have

$$E(2X - 3Y) = \frac{1}{3}\int_0^2 \int_0^1 (2x - 3y)(x+y) dx dy = \frac{1}{3}\int_0^2 (\frac{2}{3} - \frac{1}{2}y - 3y^2) dy = \frac{1}{3}(\frac{4}{3} - 1 - 8) = -\frac{23}{9},$$

and

$$\begin{aligned}E((2X-3Y)^2) &= \frac{1}{3}\int_0^2\int_0^2 (2x-3y)^2(x+y)dxdy\\ &= \frac{1}{3}\int_0^1\int_0^2 (4x^3-8x^2y-3xy^2+9y^3)dydx\\ &= \frac{1}{3}\int_0^1 (8x^3-16x^2-8x+36)dx\\ &= \frac{1}{3}(2-\frac{16}{3}-4+36)=\frac{86}{3}.\end{aligned}$$

So that $V(2X-3Y)=E((2X-3Y)^2)-E(2X-3Y)^2=\frac{245}{81}$.

Solution 3.16

This follows almost by definition,

$$E(r(X)s(Y)|X)=\int r(x)s(y)f(y|x)dy=r(x)\int s(y)f(y|x)dy=r(X)E(s(Y)|X).$$

Take $s(Y)=1$ so that $E(r(X)|X)=E(r(X)s(Y)|X)=r(X)E(s(Y)|X)=r(X)$.

Solution 3.17

This is an algebraic manipulation of symbols.

$$\begin{aligned}E(V(Y|X))+V(E(Y|X)) &= E(E(Y^2|X)-E(Y|X)^2)+E(E(Y|X)^2)-E(E(Y|X))^2\\ &= E(E(Y^2|X))-E(E(Y|X)^2)+E(E(Y|X)^2)-E(E(Y|X))^2\\ &= E(E(Y^2|X))-E(E(Y|X))^2\\ &= E(Y^2)-E(Y)^2=V(Y).\end{aligned}$$

Solution 3.18

Suppose $E(X|Y)=c$. We have $E(XY)=E(E(XY|Y))=E(E(X|Y)Y)=cE(Y)$. The covariance of X and Y is

$$\text{Cov}(X,Y)=E(XY)-E(X)E(Y)=(c-E(X))E(Y)=(c-E(E(X|Y)))E(Y)=(c-c)E(Y)=0.$$

If the covariance of X and Y is zero, $X\perp Y$.

Solution 3.19

We have $f_X(x)=1$ for $0\leq x\leq 1$. As $E(X_i)=1$ and $V(X_i)=\frac{1}{12}$, we have $E(\overline{X}_n)=\frac{1}{2}$ and $V(\overline{X}_n)=\frac{1}{12n}$.

Solution 3.20

Let a be a vector and X a random variable with mean μ and variance Σ. We have

$$E(a^tX)=E\left(\sum_{i=1}^n a_iX_i\right)=\sum_{i=1}^n E(a_iX_i)=\sum_{i=1}^n a_iE(X_i)=a^t\mu.$$

Note that in the same way $E(Xa)=\mu a$. For the variance we calculate

$$V(a^tX)=E((a^tX-E(a^tX))^2)=E(a^t(X-E(X))(X-E(X))^ta)=a^tE((X-E(X))(X-E(X))^t)a=a^t\Sigma a.$$

For the matrix A the calculations are similar,

$$E(AX)_k = E\left(\sum_{i=1}^{n} a_{ki}X_i\right) = \sum_{i=1}^{n} a_{ki}E(X_i) = (A\mu)_k,$$

so $E(AX) = A\mu$, and similar $E(XA) = \mu A$. The variance is

$$V(AX) = E((AX - E(AX))^2) = E(A(X-\mu)(X-\mu)^t A^t) = AE((X-\mu)(X-\mu)^t)A^t = A\Sigma A^t.$$

Solution 3.21

Let X and Y be random variables. Suppose that $E(Y|X) = X$. We have

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E(XY) - E(X)E(Y)\\ &= E(E(XY|X)) - E(X)E(E(Y|X))\\ &= E(XE(Y|X)) - E(X)E(X)\\ &= E(X^2) - E(X)^2 = \mathrm{Var}(X).\end{aligned}$$

Solution 3.22

Let $0 < a < b < 1$ and $X \sim \text{Uniform}(0,1)$. Define $Y = 1$ if $0 \le x \le b$ else $Y = 0$, and $Z = 1$ if $a \le x \le 0$ else $Z = 0$.

(a) Y and Z are dependent. Suppose Y and Z are independent, then $P(Y = 1|Z = 1) = P(Y = 1) = P(X < b) = b$. But we calculate

$$P(Y=1|Z=1) = \frac{P(Y=1, Z=1)}{P(Z=1)} = \frac{P(a < X < b)}{P(a < X)} = \frac{b-a}{1-a} \neq b.$$

(b) If $Z = 1$, then $a \le x \le 1$, so $E(Y|Z = 1) = \frac{b-a}{1-a}$. If $Z = 0$, then $x < a$, so $E(Y|Z = 0) = 1$. Similar $E(Z|Y = 1) = \frac{b-a}{a}$ and $E(Z|Y = 0) = 1$.

Solution 3.23

We do the whole list.

(a) $X \sim \text{Bernoulli}(p)$, then

$$\psi_X(t) = E(e^{tX}) = \sum_{x=0}^{1} e^{tx}f(x) = pe^t + (1-p).$$

(b) $X \sim \text{Binomial}(n,p)$. If we write $X = \sum_{i=0}^{n} B_i$, where $B_i \sim \text{Bernoulli}(p)$, we have

$$\psi_X(t) = E(e^{t\sum B_i}) = \prod_{i=0}^{n} E(e^{tB_i}) = (pe^t + (1-p))^n.$$

(c) $X \sim \text{Poisson}(\lambda)$, then

$$\psi_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx}\frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = \exp(\lambda(e^t - 1)).$$

(d) $X \sim \text{Normal}(\mu, \sigma)$. Note that $X = \sigma Z + \mu$ and $E(e^{tX}) = e^{t\mu}E(e^{t\sigma Z})$. So we only have to calculate

$$\begin{aligned} E(e^{t\sigma Z}) &= \int e^{t\sigma z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dz \\ &= \int \frac{1}{\sqrt{2\pi}}\exp\left(\sigma tz - \frac{z^2}{2}\right)dz \\ &= \int \frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}(z-\sigma t)^2 + \frac{\sigma^2t^2}{2}\right)dz \\ &= \exp\left(\frac{\sigma^2t^2}{2}\right)\int \frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}(z-\sigma t)^2\right)dz = \exp\left(\frac{\sigma^2t^2}{2}\right). \end{aligned}$$

Which gives $\psi_X(t) = E(e^{tX}) = \exp\left(t\mu + \frac{\sigma^2t^2}{2}\right)$.

(e) $X \sim \Gamma(\alpha, \beta)$. This is a lenghty calculation. The idea is to rewrite the integral to the Gamma function $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}du$. Use substitution $u = (\frac{1}{\beta} - t)x$, such that $dx = \frac{\beta}{1-t\beta}du$. We calculate

$$\begin{aligned} E(e^{tX}) &= \int_0^\infty e^{xt}\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}dx \\ &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-(\frac{1}{\beta}-t)x}dx \\ &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty \left(\frac{\beta}{1-t\beta}\right)^{\alpha-1}u^{\alpha-1}e^{-u}\frac{\beta}{1-t\beta}du \\ &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\left(\frac{\beta}{1-t\beta}\right)^{\alpha}\int_0^\infty u^{\alpha-1}e^{-u}du = \left(\frac{1}{1-t\beta}\right)^{\alpha}. \end{aligned}$$

Solution 3.24

Let $X_1, X_2, ..., X_n \sim \text{Exp}(\beta)$. We have

$$\begin{aligned} \psi_{X_i}(t) &= E(e^{X_it}) \\ &= \int_0^\infty e^{xt}\frac{1}{\beta}e^{-\frac{x}{\beta}}dx \\ &= \frac{1}{1-\beta t}\int_0^\infty \frac{1-\beta t}{\beta}\exp\left(-\left(\frac{1-\beta t}{\beta}x\right)\right)dx \\ &= \frac{1}{1-\beta t}\left[-\exp(-\left(\frac{1-\beta t}{\beta}\right)x\right]_{x=0}^\infty = \frac{1}{1-\beta t}. \end{aligned}$$

Let $Y = \sum_{i=1}^n X_i$. The moment generating function of Y is $\psi_Y(t) = E(e^{Yt}) = \prod E(e^{tX_i}) = \left(\frac{1}{1-\beta t}\right)^n$. Because $\psi_Y(t) = (1-\beta t)^{-n}$ is the moment generating function for the Gamma distribution, we have $Y \sim \text{Gamma}(n, \beta)$.

Chapter 4 - Inequalities

Solution 4.1

Let $X \sim \text{Exp}(\beta)$. We have $E(X) = \beta$ and $\sigma^2 = V(X) = \beta^2$. So

$P(|X-\mu| \geq k\sigma) = 1 - P(|X-\beta| < k\beta) = 1 - F((1+k)\beta) + F((1-k)\beta) = 1 - \exp(-(k+1)) + \exp(k-1).$

From Chebyshev's inequality

$$P(|X-\mu| \geq k\sigma) \leq \frac{\sigma^2}{(k\sigma)^2} = \frac{1}{k^2}.$$

Note that $1 - \exp(-(k+1)) + \exp(k-1) \leq k^{-2}$.

Solution 4.2

Let $X \sim \text{Poisson}(\lambda)$. Note that $E(X) = \lambda$, $\sigma^2 = V(X) = \lambda$, and $X > 0$. From Chebyshev's inequality

$$P(X > 2\lambda) = P(|X-\lambda| > \lambda) \leq \frac{\sigma^2}{\lambda^2} = \frac{1}{\lambda}.$$

Solution 4.3

Let $X_1, X_2, ..., X_n \sim \text{Bernoulli}(p)$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_n$. We have $E(\overline{X}) = p$ and $V(\overline{X}) = \frac{1}{n}p(1-p)$. From Chebyshev's inequality

$$P(|\overline{X} - p| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}.$$

From Hoeffding's inequality

$$P(|\overline{X} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Note that $e^{-n}/n \to 0$ when $n \to \infty$, so Hoeffding's inequality is a more strict upperbound than Chebyshev's inequality for $P(|\overline{X} - p| > \epsilon)$ when n is large.

Solution 4.4

Let $X_1, X_2, ..., X_n \sim \text{Bernoulli}(p)$.

(a) Let $\alpha > 0$ and

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

Let $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $C_n = (\hat{p_n}_n - \epsilon_n, \hat{p}_n + \epsilon_n)$. By Hoeffdings inequality

$$P(p \in C_n) = 1 - P(p \notin C_n) = 1 - P(|\hat{p}_n - p| > \epsilon) \geq 1 - 2e^{-2n\epsilon^2} = 1 - \alpha.$$

(b) See code. It seems that Hoeffding's inequality is a weak lower bound, with a factor 10 error margin.

(c) See code.

Solution 4.5

Let $Z \sim \text{Normal}(0,1)$. Note that if $x > t$, then $\frac{x}{t} > 1$. We have

$$P(|Z| > t) = 2P(Z > t) = 2\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx < \frac{2}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-\frac{x^2}{2}} dx = \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2}}}{t}.$$

Solution 4.6

Let $Z \sim \text{Normal}(0,1)$. We calculate the moments of the absolute normal random variable $|Z|$.

$$E(|Z|^k) = \int_{-\infty}^\infty |z|^k \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{2}{\sqrt{2\pi}} \int_0^\infty z^k e^{-\frac{z^2}{2}} dz = \frac{2}{\sqrt{2\pi}} \int_0^\infty 2^{\frac{k-1}{2}} x^{\frac{k-1}{2}} e^{-x} dx = \sqrt{\frac{2^k}{\pi}} \Gamma\left(\frac{k+1}{2}\right),$$

were we use substitution $x = z^2/2$. For the rest of the solution see code.

Solution 4.7

Let $X_1, X_2, ..., X_n \sim \text{Normal}(0,1)$. Define $\overline{X}_n = \frac{1}{2}\sum X_i$. Note that $E(\overline{X}) = 0$ and $V(\overline{X}) = \frac{1}{n}$, so $\sqrt{n}\overline{X} \sim \text{Normal}(0,1)$. Using Mill's inequality, we have

$$P(|\overline{X}_n| \geq t) = P(|\sqrt{n}\,\overline{X}_n| \geq \sqrt{n}t) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{nt}{2}}}{\sqrt{n}t}.$$

By Chebyshev's inequality, we have

$$P(|\overline{X}_n| \geq t) = P(|\sqrt{n}\,\overline{X}_n| \geq \sqrt{n}t) \leq \frac{1}{\sqrt{n}t}$$

Note that Mill's inequality is a shaper bound that Chebychev's inequality.

Chapter 5 - Convergence of Random Variables

Solution 5.1

Let $X_1, X_2, ..., X_n$ i.i.d. with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$.

(a) See solution 3.8.

(b) We write

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\overline{X}_n X_i + \overline{X}_n^2) \\ &= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{n-2}{n-1}\overline{X}_n^2 = c_n\frac{1}{n}\sum_{i=1}^{n}X_i^2 - d_n\overline{X}_n^2, \end{aligned}$$

where $c_n, d_n \to 1$ as $n \to \infty$. By the weak law of large numbers $\frac{1}{n}\sum_{i=1} X_i^2 \xrightarrow{P} E(X_i^2) = \mu^2 + \sigma^2$ and $\overline{X}_n^2 \xrightarrow{P} \mu^2$. Using Theorem 5.5, $S_n^2 \xrightarrow{P} \sigma^2$.

Solution 5.2

Let $X_1, X_2, ...$ be a sequence of random variables.

$\rightarrow$) Suppose $X_n \xrightarrow{qm} b$, i.e. $E((X_n - b)^2) \to 0$ as $n \to \infty$. We calculate

$$E((X_n - b)^2) = E(X_n^2) - 2bE(X_n) + b^2 = V(X_n) + (E(X_n) - b)^2 \to 0, \tag{1}$$

as $n \to \infty$. Both $V(X_n)$ and $(E(X_n) - b)^2$ are non-negative, so we must have $V(X_n) \to 0$ and $(E(X_n) - b)^2 \to 0$ as $n \to \infty$. Finally, if $(E(X_n) - b)^2 \to 0$, then $E(X_n) \to b$, as $n \to \infty$.

$\leftarrow$) Suppose $E(X_n) \to b$ and $V(X_n) \to 0$ as $n \to \infty$. Using (1), $E((X_n - b)^2) = V(X_n) + (E(X_n) - b)^2 \to 0$ as $n \to \infty$.

Solution 5.3

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables. Let $\mu = E(X)$ and $\sigma^2 = V(X)$ be finite. Take the sample mean $\overline{X}_n = \frac{1}{n}\sum X_i$. We have

$$E((\overline{X}_n - \mu)^2) = V(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{\sigma^2}{n} \to 0,$$

when $n \to \infty$. Therefore $X_n \xrightarrow{qm} \mu$ as $n \to \infty$.

Solution 5.4

Let $X_1, X_2, ...$ be i.i.d. random variables defined by $P(X_n = \frac{1}{n}) = 1 - \frac{1}{n^2}$ and $P(X_n = n) = \frac{1}{n^2}$.

(a) Note that $E(X_n) = \frac{1}{n}(1 - \frac{1}{n^2}) + n\frac{1}{n^2} = 2\frac{1}{n} - \frac{1}{n^3} \to 0$ as $n \to \infty$. But $V(X_n) = E(X_n^2) - E(X_n)^2 = E(X_n^2) = \frac{1}{n^2}(1 - \frac{1}{n^2}) + n^2\frac{1}{n^2} = 1 + \frac{1}{n^2} - \frac{1}{n^4} \to 1$ as $n \to \infty$. Therefore, by exercise 5.2, X_n cannot converge in quadratic mean.

(b) Let $\epsilon > 0$, choose n large enough such that $\frac{1}{n} < \epsilon$. We have $P(|X_n| > \epsilon) = \frac{1}{n^2} \to 0$ when $n \to \infty$. Therefore $X_n \xrightarrow{P} 0$.

Solution 5.5

Let $X_1, X_2, ... \sim$ Bernoulli(p) i.i.d. Note

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) = \frac{1}{n}\sum_{i=1}^{n} p = p,$$

and

$$V(\frac{1}{n}\sum_{i=1}^{n} X_i^2) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i^2) = \frac{1}{n^2}\sum_{i=1}^{n}(E(X_i^2) - E(X_i)^2) = \frac{1}{n^2}\sum_{i=1}^{n} p(p-1) = \frac{p(p-1)}{n}.$$

So $E(\frac{1}{n}\sum X_i^2) \to p$ and $V(\frac{1}{n}\sum X_i^2) \to 0$ as $n \to \infty$. By Exercise 5.2 we have $\frac{1}{n}\sum X_i^2 \xrightarrow{qm} p$. By Theorem 5.4 we have $\frac{1}{n}\sum X_i^2 \xrightarrow{P} p$.

Solution 5.6

Let $X_1, X_2, ..., X_{100}$ be i.i.d. random samples with $E(X_i) = 68$ and $V(X_i) = 26^2$. By the Central Limit Theorem

$$Z_{100} = \sqrt{100}\,\frac{\overline{X}_{100} - 68}{26} \approx \text{Normal}(0, 1).$$

Therefore

$$P(\overline{X}_{100} > 68) = P\left(\frac{26}{\sqrt{100}} Z_{100} + 26 > 26\right) = P(Z_{100} > 0) \approx 0.5.$$

Solution 5.7

Let $\lambda_n = \frac{1}{n}$ and $X_n \sim$ Poisson(λ_n) for $n = 1, 2, ...$

(a) $P(|X_n| > \epsilon) = 1 - P(|X_n| \leq \epsilon) < 1 - P(X_n = 0) = 1 - e^{-\frac{1}{n}} \to 0$ as $n \to \infty$.

(b) Almost the same, $P(|Y_n| > \epsilon) = P(|X_n| > \frac{\epsilon}{n}) = 1 - P(|X_n| \leq \frac{\epsilon}{n}) < 1 - P(X_n =) \to 0 = 1 - e^{-\frac{1}{n}}$ when $n \to \infty$

Solution 5.8

Let $X_1, X_2, ..., X_{100}$ be i.i.d. random variables such that $X_i \sim \text{Poisson}(1)$. Let $Y = \sum X_i = n\overline{X}_n$. By the Central Limit Theorem $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \approx Z \sim \text{Normal}(0, 1)$. We have

$$P(Y < 90) = P(\overline{X}_n < 0.9) \approx P(\frac{\sigma}{\sqrt{n}}Z + \mu < 0.9) = P(Z < \frac{\sqrt{n}}{\sigma}(0.9 - \mu)) = \Phi(-1).$$

Solution 5.9

Let X be a discrete random variable such that $P(X = 1) = P(X = -1) = \frac{1}{2}$. Define X_n by $P(X_n = X) = 1 - \frac{1}{n}$ and $P(X_n = e^n) = \frac{1}{n}$.

(a) (Quadratic convergence) $E(X_n) = (1 - \frac{1}{n})E(X) + \frac{1}{n}E(e^n) = \frac{1}{n}e^n \to \infty$ when $n \to \infty$. By exercise 5.2 X_n doesn't converge quadratic to any distribution.

(b) (Probability convergence) Let $\epsilon > 0$ and take n such that $\frac{1}{n} < \epsilon$. Then $P(|X_n| > \epsilon) = 1$. So $P(X_n > \epsilon)$ does not converge to 0 when $n \to \infty$. Therefore X_n doesn't converge in probability.

(c) (Distribution convergence) Let

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < -1, \\ \frac{1}{2} - \frac{1}{2n} & \text{if} \quad -1 \le x < 1, \\ 1 - \frac{1}{n} & \text{if} \quad 1 \le x < e^n, \\ 1 & \text{if} \quad e^n < x, \end{cases} \qquad F(x) = \begin{cases} 0 & \text{if} \quad x < -1, \\ \frac{1}{2} & \text{if} \quad -1 \le x < 1, \\ 1 & \text{if} \quad 1 \le x. \end{cases}$$

We have $X_n \sim F_n$. Let $X \sim \text{Uniform}(-1, 1)$, then $X \sim F$. Note that $\lim_{n\to\infty} F_n(x) = F(x)$ for all x. Therefore $X_n \xrightarrow{D} X$.

Solution 5.10

Let $Z \sim \text{Normal}(0, 1)$. Let $t > 0$ and $k \ge 1$ (I think this is wrong in the book, we cannot have $0 \le k < 1$). From Markov's inequality

$$P(|Z| > t) = P(|Z|^k > t^k) \le \frac{E(|Z|^k)}{t^k}.$$

From Mill's inequality we have

$$P(|Z| > t) \le \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}.$$

Note that $\frac{E(|Z|)}{t} \le \frac{E(|Z|^k)}{t^k}$. So we only have to compare $E(|Z|)$ with $\sqrt{\frac{2}{\pi}}e^{-t^2/2}$. We have

$$E(|Z|) = \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \int_0^\infty z e^{-z^2/2} dz \ge \frac{2}{\sqrt{2\pi}} \int_t^\infty z e^{-z^2/2} dz = \sqrt{\frac{2}{\pi}} \left[-e^{-z^2/2} \right]_t^\infty = \sqrt{\frac{2}{\pi}} e^{\frac{t^2}{2}}.$$

So Mill's inequality is always sharper than Markov's inequality.

Solution 5.11

Let $X_1, X_2, ...$ be i.i.d. random variables defined by $X_n \sim \text{Normal}(0, \frac{1}{n})$. Let $X \sim F$ where the cumulative distribution function is defined by $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x \ge 0$. In other words $X = 0$. Let $\epsilon > 0$, then by Mill's inequality

$$P(|X_n - X| > \epsilon) = P(|\sqrt{n}Z| > \epsilon) = P(|Z| > \frac{\epsilon}{\sqrt{n}}) \le \sqrt{\frac{2}{\pi}} \frac{1}{\epsilon} \frac{\sqrt{n}}{e^{-\frac{1}{2\epsilon}\sqrt{n}}} \to 0,$$

when $n \to \infty$. So $X_n \xrightarrow{P} X$ as $n \to \infty$. By Theorem 5.4 we have $X_n \xrightarrow{D} X$ as $n \to \infty$.

Solution 5.12

Let X and $X_1, X_2, ...$ be positive integer values random variables.

($\rightarrow$) Suppose $X_n \xrightarrow{D} X$, then $\lim_{n\to\infty} F_n(k) = F(k)$ for all $k \in \mathbb{N}$. Note that $P_n(X_n = k) = F_n(k) - F_n(k-1)$. So we have

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} (F_n(k) - F_n(k-1)) = F(k) - F(k-1) = P(X = k).$$

($\leftarrow$) Suppose that $\lim_{n\to\infty} P(X_n = k) = P(X = k)$ for all k. We have

$$F(k) - F(k-1) = P(X = k) = \lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} (F_n(k) - F_n(k-1)) = \lim_{n\to\infty} F_n(k) - \lim_{n\to\infty} F_n(k-1).$$

Rewriting the equation gives us $F(k) - \lim_{n\to\infty} F_n(k) = F(k-1) - \lim_{n\to\infty} F_n(k-1)$. Recursively applying the equation above gives

$$F(k) - \lim_{n\to\infty} F_n(k) = ... = F(-1) - \lim_{n\to\infty} F_n(-1) = 0.$$

So $F(k) = \lim_{n\to\infty} F_n(k)$.

Solution 5.13

Let $Z_1, Z_2, ...$ be i.i.d. random variables with probability density function f. Suppose that $P(Z_i > 0) = 1$ and $\lambda = \lim_{x\downarrow 0} f(x) > 0$. Let $X_n = n \min(Z_1, Z_2, ..., Z_n$. Note that

$$F(x) = \int_{-\infty}^{x} f(t)dt = 1 - \int_{x}^{\infty} f(t)dt \quad \rightarrow \quad F'(0) = \left(1 - \int_{x}^{\infty} f(t)dt\right)' = \lim_{x\downarrow 0} f(x) = \lambda > 0.$$

By the Taylor expension of F in 0 we have

$$P(Z_i \geq \frac{x}{n}) = 1 - P(Z_i \leq \frac{x}{n}) = 1 - F(\frac{x}{n}) = 1 - (F(0) + \frac{x}{n}F'(0) + O(\frac{x^2}{n^2})) = 1 - \frac{x}{n}\lambda + O(\frac{x^2}{n^2}).$$

Such that

$$\begin{aligned} F_n(x) &= P\left(X_n \leq x\right) \\ &= P\left(\min(Z_1, Z_2, ..., Z_n) \leq \frac{x}{n}\right) \\ &= 1 - \prod_{i=1}^{n} P\left(Z_i \geq \frac{x}{n}\right) \\ &= 1 - \prod_{i=1}^{n} \left(1 - \frac{x\lambda}{n} + O(\frac{x^2}{n^2})\right) \\ &= 1 - \left(1 - \frac{x\lambda}{n}\right)^n + O\left(\frac{x^2}{n^2}\right) \rightarrow 1 - e^{-\lambda x}, \end{aligned}$$

as $n \to \infty$. So $X_n \xrightarrow{D} Z \sim \text{Exp}(\frac{1}{\lambda})$.

Solution 5.14

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(0,1)$. Define $Y_n = (\overline{X}_n)^2$. From the weak law of large numbers $\overline{X}_n \xrightarrow{P} \mu = \frac{1}{2}$, hence $\overline{X}_n \xrightarrow{D} \frac{1}{2}$. Let $g(x) = x^2$, by Theorem 5.5.(g), $Y_n = g(\overline{X}_n) \xrightarrow{D} g(\frac{1}{2}) = \frac{1}{4}$.

Solution 5.15

Let

$$\begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix}, \begin{pmatrix} X_{21} \\ X_{22} \end{pmatrix}, ..., \begin{pmatrix} X_{n1} \\ X_{n2} \end{pmatrix},$$

be i.i.d. random vectors with mean $\mu = (\mu_1, \mu_2)$ and variance Σ. Define

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \quad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i},$$

and $Y_n = \overline{X}_1 / \overline{X}_2$.

We'll use the delta method. Let $g(x_1, x_2) = x_1/x_2$ such that $\nabla g(x_1, x_2) = \begin{pmatrix} 1/x_2 \\ -x_1/x_2^2 \end{pmatrix}$. We have

$$\nabla g(\mu)^t \Sigma \nabla g(\mu) = \frac{\sigma_{11}}{\mu_1^2} - \frac{\sigma_{21}}{\mu_2^2} - \frac{\sigma_{12}}{\mu_2^2} + \frac{\mu_1^2}{\mu_2^4}\sigma_{22} = \frac{\sigma_{11}}{\mu_1^2} - 2\frac{\sigma_{12}}{\mu_2^2} + \frac{\mu_1^2}{\mu_2^4}\sigma_{22}.$$

So we have

$$\sqrt{n}\left(Y_n - \frac{\mu_1}{\mu_2}\right) \xrightarrow{D} \text{Normal}\left(0, \frac{\sigma_{11}}{\mu_1^2} - 2\frac{\sigma_{12}}{\mu_2^2} + \frac{\mu_1^2}{\mu_2^4}\sigma_{22}\right),$$

and

$$Y_n \xrightarrow{D} \text{Normal}\left(\frac{\mu_1}{\mu_2}, \frac{1}{n}\left(\frac{\sigma_{11}}{\mu_1^2} - 2\frac{\sigma_{12}}{\mu_2^2} + \frac{\mu_1^2}{\mu_2^4}\sigma_{22}\right)\right).$$

Solution 5.16

Let $X_1, X_2, ... \sim \text{Normal}(0,1)$ and $X \sim \text{Normal}(0,1)$. Define $Y_n = -X_n$. Then $X_n \xrightarrow{D} Z$ and $Y_n \xrightarrow{D} Z$, but $X_n + Y_n = 0$ does not converge to Z.

Part II

Statistical Inference

Chapter 6 - Models, Statistical Inference and Learning

Solution 6.1

Let $X_1, X_2, ..., X_n \sim \text{Poisson}(\lambda)$. Estimate $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$. We have $E(\hat{\lambda}) = \lambda$, so $\text{bias}(\hat{\lambda}) = E(\hat{\lambda}) - \lambda = 0$. Moreover, $V(\hat{\lambda}) = \frac{\lambda}{n}$, so $\text{se} = \sqrt{\frac{\lambda}{n}}$ and $\text{MSE} = \text{bias}^2 + V(\hat{\lambda}) = \frac{\lambda}{n}$.

Solution 6.2

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $\hat{\theta} = \max(X_1, X_2, ..., X_n)$. The cumulative density function is $F(t) = P(\hat{\theta} < t) = \prod P(X_i < t) = t^n$. So the probability distribution function is $f(t) = F'(t) = nt^{n-1}$. This gives

$$E(\hat{\theta}) = \int_0^\theta tf(t)dt = \frac{n}{n+1}\theta, \quad \text{bias} = E(\hat{\theta}) - \theta = -\frac{1}{n+1}\theta.$$

So $\hat{\theta}$ is not unbiased ($\frac{n+1}{n}\hat{\theta}$ would be unbiased). Note that

$$\begin{aligned} E(\hat{\theta}^2) &= \int_0^\theta t^2 f(t)dt = \frac{n}{n+2}\theta^2, \\ V(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2 &= \frac{n}{n+1}\theta^2 - \frac{1}{(n+1)^2}\theta^2 = \frac{n}{(n+1)^2(n+2)^2}\theta^2. \end{aligned}$$

We have $\text{se} = \sqrt{V(\hat{\theta})}$. And the mean squared error is given by

$$\text{MSE} = \text{bias}^2 + V(\hat{\theta}) = \frac{1}{(n+1)^2}\theta^2 + \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{n^2+2n+1}{(n+1)^2(n+2)}\theta^2 = \frac{1}{n+2}\theta^2.$$

Solution 6.3

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $\hat{\theta} = \frac{2}{n}\sum X_i$. We have

$$E(\hat{\theta}) = \frac{2}{n}\sum_{i=0}^n E(X_i) = \theta, \quad V(\hat{\theta}) = \frac{2^2}{n^2}\sum_{i=0}^n V(X_i) = \frac{\theta^2}{3n^2}.$$

Therefore, $\text{bias} = 0$, so $\hat{\theta}$ is unbiased, $\text{se} = \frac{\theta}{\sqrt{3}n}$, and $\text{MSE} = \frac{\theta^2}{3n^2}$.

Chapter 7 - Estimating the CDF and Statistical Functionals

Solution 7.1

Let $X_1, X_2, ..., X_n$ random variables with cumulative distribution function F. Let $\hat{F}_n(x) = \frac{1}{n}\sum I(X_i \leq x)$ be the emperical distribution function of F.

(a) Note that

$$E(I(X_i \leq x)) = \int_{-\infty}^{\infty} I(t \leq x)f(t)dt = \int_{-\infty}^{x} f(t)dt = F(x).$$

Therefore, $E(\hat{F}_n(x)) = \frac{1}{n}\sum E(I(X_i \leq x)) = F(x)$.

(b) As in (a), we calculate $E(I(X_i \leq x)^2) = F(x)$. So

$$V(\hat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^n V(I(X_i \leq x)) = \frac{1}{n^2}\sum_{i=1}^n (E(I(X_i \leq x)^2) - E(I(X_i \leq x))^2) = \frac{1}{n}F(x)(1 - F(x)).$$

(c) Using (a) we have bias $= E(\hat{F}_n(x)) - F(x) = 0$, so from (b), MSE $= V(\hat{F}_n(x)) = \frac{1}{n}F(x)(1-F(x)) \to 0$ when $n \to \infty$.

(d) As MSE $= E((\hat{F}_n(x) - F(x))^2) \to 0$ as $n \to \infty$, $\hat{F}_n(x) \xrightarrow{qm} F(x)$ and by Theorem 5.4.a $\hat{F}_n(x) \xrightarrow{P} F(x)$.

Solution 7.2

Let $X_1, X_2, ..., X_n \sim$ Bernoulli(p) and $Y_1, Y_2, ..., Y_m \sim$ Bernoulli(q).

(a) The plug-in estimator for p is $\hat{p} = E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^n X_i$. The plug-in estimator for the standard error on $\hat{p}$ is $\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$.

(b) The 90% confidence interval for plug-in estimator $\hat{p}$ is $\hat{p} \pm z_{0.05}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$.

(c) The plug-in estimator for $p - q$ is $\hat{p} - \hat{q}$. The plug-in estimator for the standard error on $\hat{p} - \hat{q}$ is $\sqrt{V(\hat{p}) + V(\hat{q})} = \sqrt{\frac{1}{n}\hat{p}(1-\hat{p}) + \frac{1}{m}\hat{q}(1-\hat{1})}$.

(d) The 90% confidence interval for plug-in estimator $\hat{p} - \hat{q}$ is $\hat{p} - \hat{q} \pm z_{0.05}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p}) + \frac{1}{m}\hat{q}(1-\hat{q})}$.

Solution 7.3

See code.

Solution 7.4

Let $X_1, X_2, ..., X_n \sim F$. Let $\hat{F}(x) = \frac{1}{n}\sum I(X_i \leq x)$. Denote $Y_i = I(X_i \leq x)$. Note that $\overline{Y}_n = \hat{F}(x)$, and $E(\hat{F}(x)) = F(x)$ and $V(\hat{F}(x)) = \frac{1}{n}F(x)(1-F(x))$. By the central limit theorem

$$\text{Normal}(0,1) \approx \frac{\overline{Y}_n - E(\overline{Y}_n)}{\sqrt{V(\overline{Y}_n)}} = \sqrt{n}\frac{\hat{F}(x) - F(x)}{\sqrt{F(x)(1-F(x))}}.$$

In other words, when $n \to \infty$, $\hat{F}(x)$ behaves as a random variable from the distribution Normal$(F(x), \frac{1}{n}F(x)(1-F(x)))$.

Solution 7.5

Let $x \neq y$. We have

$$\text{Cov}(\hat{F}(x), \hat{F}(y)) = \frac{1}{n^2}\text{Cov}(\sum_{i=1}^n I(X_i \leq x), \sum_{j=1}^n I(X_j \leq y)) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \text{Cov}(I(X_i \leq x), I(X_j \leq y)).$$

Now, to find

$$\begin{aligned}\text{Cov}(I(X_i \leq x), I(X_j \leq y)) &= E(I(X_i \leq x)I(X_j \leq y)) - E(I(X_i \leq x))E((X_j \leq y)) \\ &= E(I(X_i \leq x)I(X_j \leq y)) - F(x)F(y).\end{aligned}$$

If $i \neq j$, we have

$$\begin{aligned}E(I(X_i \leq x)I(X_j \leq y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(s \leq x)I(t \leq y)f(s)f(t)dsdt \\ &= \int_{-\infty}^{\infty} I(s \leq x)f(s)ds\int_{-\infty}^{\infty} I(t \leq y)f(t)dt = F(x)F(y),\end{aligned}$$

and if $i = j$,

$$E(I(X_i \leq x)I(X_i \leq y)) = E(I(X_i \leq \min(x,y))^2) = F(\min(x,y)).$$

Therefore, we ahve

$$\text{Cov}(\hat{F}(x), \hat{F}(y)) = \begin{cases} 0 & \text{if} \quad i \neq j, \\ \frac{1}{n}(F(\min(x,y)) - F(x)F(y)) & \text{if} \quad i = j. \end{cases}$$

Solution 7.6

Let $X_1, X_2, ..., X_n \sim F$, and define the emperical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$. Take $a < b$ and $\theta = T(F) = F(b) - F(a)$. Let $\hat{\theta} = T(\hat{F}) = \hat{F}(b) - \hat{F}(a)$ be the plug-in estimator of θ. We have

$$\begin{aligned}
E(\hat{\theta}) &= E(\hat{F}(b) - \hat{F}(a)) = F(b) - F(a) = \theta, \\
E(\hat{F}(a)^2) &= V(\hat{F}(a)) + E(\hat{F}(a))^2 = \frac{1}{n}F(a)(1 - F(a)) + F(a)^2 = \frac{1}{n}F(a)(1 + (n-1)F(a)), \\
E(\hat{F}(b)^2) &= \frac{1}{n}F(b)(1 + (n-1)F(b)), \\
E(\hat{F}(a)\hat{F}(b)) &= \text{Cov}(\hat{F}(a), \hat{F}(b)) + E(\hat{F}(a))E(\hat{F}(b)) \\
&= \frac{1}{n}(F(a) - F(a)F(b)) + F(a)F(b) = \frac{1}{n}F(a) + \frac{n-1}{n}F(a)F(b),
\end{aligned}$$

where we used Solution 7.5 to calculate the covariance in the last equation. We now have

$$\begin{aligned}
E(\hat{\theta}^2) &= E(F(b)^2 - 2F(b)F(a) + F(a)^2) \\
&= \frac{1}{n}F(b)(1 + (n-1)F(b)) - \frac{2}{n}F(a)(1 - (n-1)F(b)) + \frac{1}{n}F(a)(1 + (n-1)F(a)) \\
&= \frac{1}{n}(F(b) - F(a)) + \frac{n-1}{n}(F(b) - F(a))^2 \\
&= \frac{1}{n}\theta + \frac{n-1}{n}\theta^2,
\end{aligned}$$

so that

$$V(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2 = \frac{1}{n}\theta + \frac{n-1}{n}\theta^2 - \theta^2 = \frac{1}{n}\theta(1 - \theta).$$

The plugin-estimator for the standard error becomes

$$\hat{\text{se}}(\hat{\theta}) = \sqrt{\frac{1}{n}\hat{\theta}(1 - \hat{\theta})},$$

and the $1 - \alpha$ confidence interval is given by

$$\hat{\theta} \pm z_{\frac{\alpha}{2}} \hat{se}(\hat{\theta}) = \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{\theta}(1 - \hat{\theta})}.$$

Solution 7.7

See code.

Solution 7.8

See code.

Solution 7.9

Let $X_1, X_2, ..., X_{100} \sim \text{Bernoulli}(p_1)$ and $Y_1, Y_2, ..., Y_{100} \sim \text{Bernoulli}(p_2)$. The plugin-estimators are $\hat{p}_1 = \frac{90}{100} = 0.9$ and $\hat{p}_2 = \frac{85}{100} = 0.85$. Recall that the variance on $\hat{p}_i$ is

$$V(\hat{p}_i) = \frac{1}{100^2} \sum_{i=1}^{n} V(X_i) = \frac{1}{100} p_i(1 - p_i).$$

Let $\theta = p_1 - p_2$, then $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = 0.9 - 0.85 = 0.05$. We have

$$V(\hat{\theta}) = V(\hat{p}_1) + V(\hat{p}_2) = \frac{1}{100}(p_1(1 - p_1) + p_2(1 - p_2)),$$

so that we estimate $\text{se}(\hat{\theta})$ with the plugin-estimator

$$\hat{\text{se}}(\hat{\theta}) = \frac{1}{10}\sqrt{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)} \approx 0.05.$$

The 80% interval is given by

$$\hat{\theta} \pm z_{0.1}\hat{\text{se}}(\hat{\theta}) \approx (-0.01, 0.11),$$

and the 95% interval is

$$\hat{\theta} \pm z_{0.05}\hat{\text{se}}(\hat{\theta}) \approx (-0.04, 0.14).$$

Solution 7.10

See code.

Chapter 8 - The Bootstrap

Solution 8.1

See code.

Solution 8.2

See code.

Solution 8.3

See code.

Solution 8.4

Regard $X_1, X_2, ..., X_n$ as n different buckets. A bootstrap sample $X_1*, X_2^*, ..., X_n^*$ can be seen as a non-negative integer valued vector $(y_1, y_2, ..., y_n)$ where y_i represents the number of times X_i is chosen. Or, in other words, y_i is the number of times a ball is put in bucket X_i. We look for all non-negative valued vector solutions $(y_1, y_2, ..., y_n)$ such that $y_1 + y_2 + ... + y_n = n$ and $y_i \geq 0$. Or, equivalent, all $(\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n)$ such that $\tilde{y}_1 + \tilde{y}_2 + ... + \tilde{y}_n = 2n$ and $\tilde{y}_i = y_1 + 1 \geq 1$. This is the stars and bars problem. Consider $2n$ stars $\star\star\ldots\star$ which we want to partition into n non-empty sets. There are $2n - 1$ places between the stars where we can put n bars to split the stars. Therefore, there are $\binom{2n-1}{n}$ possible n non-empty partitions of $2n$ stars. Equivalent, there are $\binom{2n-1}{n}$ unique vectors $(y_1, y_2, ..., y_n)$ such that $y_1 + y_2 + ... + y_n = n$ with $y_n \geq 0$. Which gives the number of unique bootstrap draws of $Y_1, Y_2, ..., Y_n$ with replacement.

Solution 8.5

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Let $X_1^*, X_2^*, ..., X_n^*$ be the bootstrap sample. Define $\overline{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$. Note that

$$E(X_i^*|X_1, X_2, ..., X_n) = \sum_{i=1}^n \frac{1}{n} X_i = \overline{X}_n.$$

Therefore,

$$E(\overline{X}_n^*|X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^n E(X_i^*|X_1, X_2, ..., X_n) = \frac{1}{n}\sum_{i=1}^n \overline{X}_n = \overline{X}_n.$$

By the rule of iterated expectations, Theorem 3.24,

$$E(\overline{X}_n^*) = E(E(\overline{X}_n^*|X_1, X_2, ..., X_n)) = E(\overline{X}_n) = \mu.$$

Next we have

$$V(\overline{X}_n^*|X_1, X_2, ..., X_n) = \frac{1}{n^2}\sum_{i=1}^n V(X_i^*|X_1, X_2, ..., X_n) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{1}{n}(X_j - E(X_j))^2 = \frac{1}{n^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

By Theorem 3.27 we have

$$\begin{aligned} V(\overline{X}_n^*) &= E(V(\overline{X}_n^*|X_1, X_2, ..., X_n)) + V(E(\overline{X}_n^*|X_1, X_2, ..., X_n)) \\ &= E\left(\frac{1}{n^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right) + V(\overline{X}_n) \\ &= \frac{n-1}{n^2}\sigma^2 + \frac{\sigma^2}{n} = \frac{2n-1}{n^2}\sigma^2, \end{aligned}$$

where we used Theorem 3.17 to calculate

$$E\left(\frac{1}{n^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right) = \frac{n-1}{n^2} E\left(\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right) = \frac{n-1}{n^2}\sigma^2.$$

Solution 8.6

See code. For the true distribution of $\hat{\theta} = e^{\overline{X}_n}$ note that $\overline{X}_n \sim$ Normal(μ, $\frac{1}{n}$), and

$$F(t) = P(\hat{\theta} < t) = P\left(e^{\overline{X}_n} < t\right) = P(\overline{X} < \log(t)) = P(Z < \sqrt{n}(\log(t) - \mu)) = \Phi(\sqrt{n}(\log(t) - \mu)),$$

such that

$$f(t) = F(t)' = \frac{\sqrt{n}}{t}\phi(\sqrt{n}(\log(t) - \mu)).$$

Solution 8.7

Let $X_1, X_2, ..., X_n \sim$ Uniform$(0, \theta)$. Let $th\hat{e}ta = \max(X_1, X_2, ..., X_n)$. Let $\theta = 1$.

(a) We have $P(\hat{\theta} < t) = \prod_i P(X_i < t) = t^n$. So the probability density function is given by $f(t) = nt^{n-1}$. See code for simulation with $\theta = 1$.

(b) Note that the pdf for $\hat{\theta}$ is continuous, hence $P(\hat{\theta}^* = \hat{\theta}) = 0$. We have

$$P(\hat{\theta}^* = \hat{\theta}) = 1 - \prod_{i=1}^n P(X_i^* \neq \hat{\theta}) = 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - \frac{1}{e} \approx 0.632,$$

as $n \to \infty$.

Solution 8.8

I have no idea what this exercise is about or how to solve it.

Chapter 9 - Parametric Inference

Solution 9.1

Let $X_1, X_2, ..., X_n \sim \text{Gamma}(\alpha, \beta)$. We have

$$\begin{aligned}
E(X^n) &= \int_0^\infty x^n \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-\frac{x}{\alpha}} dx \\
&= \frac{(\alpha+n-1)!}{(\alpha-1)!} \beta \int_0^\infty \frac{x^{\alpha+n-1}}{\beta^{\alpha+n}\Gamma(\alpha+n)} e^{-\frac{x}{\alpha}} dx \\
&= \frac{(\alpha+n-1)!}{(\alpha-1)!} \beta \int_0^\infty f(x; \alpha+n, \beta+n) dx = \frac{(\alpha+n-1)!}{(\alpha-1)!} \beta.
\end{aligned}$$

In particular, $\alpha_1 = E(X) = \alpha\beta$ and $\alpha_2 = E(X^2) = (\alpha+1)\alpha\beta^2$. Solving these two equations leads to

$$\alpha = \frac{\alpha_1^2}{\alpha_2 - \alpha_1^2}, \quad \beta = \frac{\alpha_2 - \alpha_1^2}{\alpha_1^2}.$$

When we take $\hat{\alpha}_1 = \frac{1}{n}\sum X_i$ and $\hat{\alpha}_2 = \frac{1}{n}\sum X_i^2$, the methods of moment estimators for α and β become

$$\hat{\alpha} = \frac{\alpha_1^2}{\alpha_2 - \alpha_1^2}, \quad \hat{\beta} = \frac{\alpha_2 - \alpha_1^2}{\alpha_1^2}.$$

Solution 9.2

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(a, b)$.

(a) We calculate the moments

$$E(X^n) = \int_a^b x^n f(x; a, b) dx = \frac{1}{b-a} \frac{x^{n+1}}{n+1}\bigg|_a^b = \frac{1}{n+1} \frac{b^{n+1} - a^{n+1}}{b-a}.$$

In particular, $\alpha_1 = E(X) = \frac{a+b}{2}$ and $\alpha_2 = E(X^2) = \frac{a^2+ab+b^2}{3}$. From α_1 we have $\hat{b} = 2\hat{\alpha}_1 - \hat{a}$. Combined with α_2 we get a second order equation

$$3\hat{\alpha}_2 = \hat{a}^2 + \hat{a}(2\hat{\alpha}_1 - \hat{a}) + (2\hat{\alpha}_1 - \hat{a})^2 = \hat{a}^2 - 2\hat{a}\hat{\alpha}_1 + 4\hat{\alpha}_1^2,$$

with the (only possible) solution

$$\hat{a} = \hat{\alpha}_1 - \sqrt{3(\hat{\alpha_2} - \hat{\alpha_1}^2)}, \quad \hat{b} = \hat{\alpha}_1 + \sqrt{3(\hat{\alpha_2} - \hat{\alpha_1}^2)}.$$

(b) The likelihood estimator is

$$\mathcal{L}_n(a, b) = \prod_{i=1}^n f(X_i; a, b) = \begin{cases} (b-a)^{-n} & \text{if} \qquad a \leq X_1, X_2, ..., X_n \leq b \\ 0 & \text{otherwise} \end{cases}$$

We cannot differentiate the likelihood estimator. However, let $\hat{a} = \min(X_1, X_2, ..., X_n)$ and $\hat{b} = \max(X_1, X_2, ..., X_n)$. If $a' < \hat{a}$ or $\hat{b} < b'$, then $\mathcal{L}_n(a', b') = 0 \leq \mathcal{L}_n(\hat{a}, \hat{b})$. If $\hat{a} \leq a'$ and $b' \leq \hat{b}$, then $\mathcal{L}_n(a', b') = (b' - a')^{-n} \leq (\hat{b} - \hat{a})^{-n} = \mathcal{L}_n(\hat{a}, \hat{b})$. Therefore, $\mathcal{L}_n(a, b)$ is maximized in $\hat{a}, \hat{b}$.

(c) Let $\tau = \int x dF(x) = E(X) = \frac{a+b}{2}$. The MLE is $\hat{\tau} = \frac{\hat{a}+\hat{b}}{2}$.

(d) The plugin-estimator for τ is $\tilde{\tau} = \frac{1}{n}\sum X_i$, also see examples 7.10 and 7.11. We have

$$\begin{aligned} E((\tilde{\tau}-\tau)^2) &= E((\tilde{\tau}-E(\tilde{\tau}))^2) \\ &= V(\tilde{\tau}) \\ &= E(\tilde{\tau}^2) - E(\tilde{\tau}^2) \\ &= \frac{1}{n^2}E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i^2) - \sum_{i\neq j} E(X_i)E(X_j)\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n}(V(X_i)+E(X_i)^2) - \sum_{i\neq j}\left(\frac{a+b}{2}\right)^2\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{1}{n^2}\left(\frac{n}{12}(b-a)^2 + n\left(\frac{a+b}{2}\right)^2 + n(n-1)\left(\frac{a+b}{2}\right)^2\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{(b-a)^2}{12n} = \frac{V(X)}{n}. \end{aligned}$$

Solution 9.3

Let $X_1, X_2, ..., X_n \sim \text{Normal}(\mu, \sigma^2)$. Let τ be the 95% percentile, i.e., $P(X > \tau) = 0.95$.

(a) Note that

$$0.95 = P(X < \tau) = P(\sigma Z + \mu < \tau) = P\left(Z < \frac{\tau-\mu}{\sigma}\right) = \Phi\left(\frac{\tau-\mu}{\sigma}\right).$$

Hence $\tau = z_{0.95}\sigma + \mu$. Therefore, the MLE for τ is $\hat{\tau} = z_{0.95}\hat{\sigma} + \hat{\mu}$.

(b) We use the Delta method. $\tau = g(\mu,\sigma) = z_{0.95}\sigma + \mu$, such that $\triangledown g = (1, z_{0.95})^t$. The Fisher information matrix becomes

$$I_n(\mu,\sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}.$$

Let $J_n = I_n^{-1}$ be the inverse of the Fisher information matrix. Then

$$\hat{se}(\hat{\tau})^2 = \hat{\triangledown} g^t \hat{J}_n \hat{\triangledown} g = \frac{\sigma^2}{n}\left(1 + \frac{1}{2}z_{0.95}^2\right).$$

Therefore, the (approximated) $1-\alpha$ confidence interval for $\hat{\tau}$ is

$$(\hat{\mu} + z_{0.95}\hat{\sigma}) \pm z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{2+z_{0.95}}{2}}.$$

(c) See code.

Solution 9.4

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(0,\theta)$. The MLE of θ is given by $\hat{\theta} = \max(X_1, X_2, ..., X_n)$. For every $\epsilon > 0$, we have

$$P(|\theta - \hat{\theta}| < \epsilon) = P(X_1 < \theta-\epsilon, X_2 < \theta-\epsilon, ..., X_n < \theta-\epsilon) = \prod_{i=1}^{n}\frac{\theta-\epsilon}{\theta} = \left(1-\frac{\epsilon}{\theta}\right)^n \to 0,$$

when $n \to \infty$, so $\hat{\theta} \xrightarrow{P} \theta$ and $\hat{\theta}$ is consistant.

Solution 9.5

Let $X_1, X_2, ..., X_n \sim \text{Poisson}(\lambda)$.

(a) Let $X \sim \text{Poisson}(\lambda)$. We have $\alpha_1 = E(X) = \lambda$, so the moments estimator is $\hat{\lambda} = \hat{\alpha}_1$.

(b) The Likelihood function is

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{X_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n X_i}}{x_1! x_2! \ldots x_n!}.$$

Let $F_n(\lambda) = e^{-n\lambda} \lambda^{\sum X_i}$. To find the maximum likelihood estimator $\hat{\lambda}$ it's sufficient to maximize $F_n(\lambda)$, as the rest of $\mathcal{L}_n(\theta)$ is independent of θ. We have

$$f_n(\lambda) = \log(F_n(\lambda)) = -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i = 0,$$

if and only if $\lambda = \frac{1}{n} \sum_{i=1}^n X_i$ $(= \hat{\alpha}_1)$.

(c) To calculate the Fisher matrix we first calculate the score function

$$\frac{\partial^2 \ell_n(\lambda)}{\partial \lambda^2} = \frac{\partial^2 f_n(\lambda)}{\partial \lambda^2} = -\frac{\sum_{i=1}^n X_i}{\lambda^2},$$

such that

$$I_n(\lambda) = -E_\lambda \left(\frac{\partial^2 \ell_n(\lambda)}{\partial \lambda^2} \right) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

In particular, the Fisher Information matrix is given by $I(\lambda) = \frac{1}{n} I_n(\lambda) = \frac{1}{\lambda}$.

Solution 9.6

Let $X_1, X_2, ..., X_n \sim \text{Normal}(\theta, 1)$. Define $Y_i = 1$ if $X_i \geq 0$ and $Y_i = 0$ otherwise. Let $\psi = P(Y_1 = 1)$.

(a) $\psi = P(Y_1 = 1) = P(X_1 \geq 0) = P(Z \geq -\theta) = P(Z < \theta) = \Phi(\theta)$. So the MLE of ψ is $\hat{\psi} = \Phi(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ.

(b) We use Delta method of Theorem 9.24. $\hat{\psi} = g(\hat{\theta}) = \Psi(\hat{\theta})$, so $g'(\hat{\theta}) = \phi(\hat{\theta})$. Moreover, $\hat{se}(\hat{\theta})^2 = \frac{\sigma^2}{n} = \frac{1}{n}$. So the 95% confidence interval is

$$\Phi(\hat{\theta}) \pm z_{0.025} \frac{\phi(\hat{\theta})}{\sqrt{n}}.$$

(c) Let $\tilde{\psi}_n = \frac{1}{n} \sum_{i=1}^n Y_i$. By the weak law of large numbers $\tilde{\psi}_n \xrightarrow{P} E(Y_1) = P(X > 0) = \psi$.

(d) We have $\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{D} \text{Normal}(0, \phi(\theta))$. Note that $V(\tilde{\psi}_n) = \frac{1}{n^2} \sum V(Y_i) = \frac{1}{n} \psi(1-\psi)$. So $\sqrt{n}(\tilde{\psi}_n - \psi) \xrightarrow{D} \text{Normal}(0, \psi(1-\psi))$. Which leads to

$$\text{ARE}(\hat{\psi}, \tilde{\psi}) = \frac{\phi(\theta)}{\psi(1-\psi)}.$$

(e) $\hat{\psi} = \Phi(\hat{\theta})$ will converge to $\Psi(\mu)$ as $\hat{\theta}$ will converge to mean μ.

Solution 9.7

Let $X_1 \sim \text{Binomial}(n_1, p_1)$ and $X_2 \sim \text{Binomial}(n_2, p_2)$. Take $\psi = p_1 - p_2$.

(a) The probability density function for X_1 and X_2 is $f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$. The Likelihood function is given by

$$\mathcal{L}_1(p) = f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

So $\ell_1(p) = \log\binom{n}{x} + x\log(p) + (n-x)\log(p)$. Maximizing $\ell_1(p)$ yields $\hat{p} = x/n$. The MLE of $\psi = p_1 - p_2$ is $\hat{\psi} = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$.

(b) The probability density function is $f(x; p_1, p_2) = \binom{n_1}{x} p_1^x (1-p_1)^{n-x} \binom{n_2}{x} p_2^x (1-p_2)^{n_2-x}$. We have $\ell(p_1, p_2) = x\log(p_1) + (n_1 - x)\log(1-p_1) + x\log(p_2) + (n_2 - x)\log(1-p_2) + C$, where C is a constant independent of p_1 and p_2. The second order differentials are

$$\frac{\partial^2 \ell(p_1,p_2)}{\partial p_i^2} = -\frac{X_i}{p_i^2} + \frac{X_i - n_i}{(1-p_i)^2}, \quad \frac{\partial^2 \ell(p_1,p_2)}{\partial p_1 \partial p_2} = 0.$$

Which gives

$$-E\left(\frac{\partial^2 \ell(p_1,p_2)}{\partial p_i^2}\right) = \frac{n_i p_i}{p_i^2} - \frac{n_i p_i - n_i}{(1-p_i)^2} = \frac{n_i}{p_i} - \frac{n_i}{(1-p_i)} = \frac{n_i}{p_i(1-p_i)}.$$

The Fisher Information matrix is

$$I(p_1,p_2) = \begin{pmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{pmatrix}.$$

(c) Let $g(p_1, p_2) = p_1 - p_2$. We have $\nabla g = (1, -1)^t$. The inverse Information matrix is given by

$$J_n(p_1,p_2) = I_n^{-1}(p_1,p_2) = \frac{1}{n} I(p_1,p_2) = \frac{1}{n}\begin{pmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{pmatrix}.$$

Set $n = 1$, with the Delta method,

$$\hat{se}(\hat{\psi})^2 = \nabla \hat{g}^t \hat{J}_1 \nabla \hat{g} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}.$$

(d) See code.

Solution 9.8

Let $X_1, X_2, ..., X_n \sim \text{Normal}(\mu, \sigma^2)$. The Likelihood estimator is given by

$$\mathcal{L}_n(\mu,\sigma) = C\frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right).$$

Therefore,

$$\ell_n(\mu,\sigma) = \log(C) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

The partial derivatives are

$$\begin{aligned}
\frac{\partial \ell_n}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu), \quad \frac{\partial^2 \ell_n}{\partial \mu^2} = -\frac{1}{\sigma^2}\sum_{i=1}^n X_i,\\
\frac{\partial^2 \ell_n}{\partial \mu \partial \sigma} &= -\frac{2}{\sigma^3}\sum_{i=1}^n (X_i-\mu) = \frac{\partial^2 \ell_n}{\partial \sigma \partial \mu},\\
\frac{\partial \ell_n}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (X_i-\mu)^2, \quad \frac{\partial^2 \ell_n}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^n (X_i-\mu)^2.
\end{aligned}$$

The expected values become

$$E\left(-\frac{\partial^2 \ell_n}{\partial \mu^2}\right) = \frac{n}{\sigma^2}, \quad E\left(-\frac{\partial^2 \ell_n}{\partial \mu \partial \sigma}\right) = 0, \quad E\left(-\frac{\partial^2 \ell_n}{\partial \sigma^2}\right) = -\frac{n}{\sigma^2} + \frac{3}{\sigma^3} n\sigma^2 = \frac{2n}{\sigma^2}.$$

The Fisher information matrix is

$$I_n(\mu,\sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}.$$

Solution 9.9

See code. It seems like delta, parametric, and non-pametric bootstrap are equally close to the truth.

Solution 9.10

See solution 8.7 and code.

10 Hypothesis Testing and p-values

Solution 10.1

By the definition of the power $\beta(\theta_*) = P_{\theta_*}(X \in R)$. The Wald statistic is defined by

$$W = \frac{\hat{\theta} - \theta_0}{\hat{se}},$$

and the rejection area R is defined by $|W| > z_{\alpha/2}$. For large sample size $W \xrightarrow{D} \text{Normal}(0,1)$, or in other words $\hat{\theta} \xrightarrow{D} \theta_*$. Therefore, we have

$$\begin{aligned}
\beta(\theta_*) &= P_{\theta_*}(X \in R)\\
&= P_{\theta_*}(|W| > z_{\frac{\alpha}{2}})\\
&\approx P_{\theta_*}\left(\left|\frac{\theta_* - \theta_0}{\hat{se}}\right| > z_{\frac{\alpha}{2}}\right)\\
&= 1 - \Phi\left(\left|\frac{\theta_* - \theta_0}{\hat{se}}\right| + z_{\frac{\alpha}{2}}\right) + \Phi\left(\left|\frac{\theta_* - \theta_0}{\hat{se}}\right| - z_{\frac{\alpha}{2}}\right).
\end{aligned}$$

Solution 10.2

Some definitions are missing in the book. Let T be the test statistic with the continuous cumulative distribution function F. We can consider the p-value (of a left-sided one-tailed hypothesis) as a distribution taking $P = F(T)$, such that when we observe t_{obs}, the p-value is $p = F(t_{\text{obs}})$. With this definition, we have

$$P_{\theta_0}(P < t) = P_{\theta_0}(F(T) < t) = P_{\theta_0}(T < F^{-1}(t)) = F(F^{-1}(t)) = t.$$

So P has the identify function as cumulative distribution function. This is only possible if $P \sim \text{Uniform}(0,1)$, proving Theorem 10.14.

Solution 10.3

Theorem 10.10 follows directly from the definition. We reject H_0 if and only if $|W| > z_{\frac{\alpha}{2}}$ which is equivalent to saying

$$\theta_0 \notin (\hat{\theta} - \hat{se} z_{\frac{\alpha}{2}}, \hat{\theta} + \hat{se} z_{\frac{\alpha}{2}}).$$

Solution 10.4

We reject H_0 if and only if $T(X^n) \geq c_\alpha$. From the definition

$$\begin{aligned} p &= \inf_\alpha(\alpha : T(x^n) \in R_\alpha) \\ &= \inf_\alpha(\alpha : T(x^n) \geq c_\alpha) \\ &= \inf_\alpha(\sup_\theta \beta(\theta) : T(x^n) \geq c_\alpha) \\ &= \inf_\alpha(\sup_\theta P_\theta(T(X^n) \geq c_\alpha) : T(x^n) \geq c_\alpha) \\ &= \sup_\theta P_\theta(T(x^n) \geq T(X^n)), \end{aligned}$$

as $P_\theta(T(X^n) \geq c_\alpha)$ is smalles when c_α is as large as possible, which happens when $c_\alpha = T(x^n)$. In particular, when $\Theta_0 = \{\theta_0\}$, we have $p = P_{\theta_0}(T(X^n) \geq T(x^n))$.

Solution 10.5

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $Y = \max(X_1, X_2, ..., X_n)$. We want to test $H_0 : \theta = \frac{1}{2}$ against $H_1 : \theta \neq \frac{1}{2}$.

(a) We generalize $H_0 : \theta = c$ against $H_1 : \theta \neq c$. The power function is given by $\beta(\theta) = P_\theta(Y > c) = 1 - P_\theta(Y \leq c) = 1 - \left(\frac{c}{\theta}\right)^n$.

(b) We calculate $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = 1 - \left(\frac{c}{\theta_0}\right)^n$, which leads to $c_\alpha = \theta_0(1-\alpha)^{\frac{1}{n}}$. With a test size of 0.05 under $H_0 : \theta = \frac{1}{2}$ we have $c_{0.05} = \frac{1}{2} 0.95^{\frac{1}{n}}$.

(c) If $n = 20$ and $Y = 0.48$, we have $p = P_{\theta_0}(Y \geq 0.48) = 1 - \left(\frac{0.48}{0.5}\right)^{20} \approx 0.56$. We would not reject the null hypothesis $H_0 : \theta = \frac{1}{2}$.

(d) If $n = 20$ and $Y = 0.52$, then $Y > \frac{1}{2} = \theta$, so we would reject the null hypothesis $H_0 : \theta = \frac{1}{2}$ immediately.

Solution 10.6

I'm not sure if I understand the question correctly. Phillips and King are interested in testing the null hypothesis $\theta \leq \frac{1}{2}$, not $\theta = \frac{1}{2}$. The first one is statistically significant, the latter is not as you will see in the solution.

We can use the Wald test. The null hypothesis $H_0 : \theta_0 = \frac{1}{2}$. Let $n = 1919$. We estimate $\hat{\theta} = \frac{922}{n} \approx 0.48$. Note that

$$\text{Var}(\theta_0) = \text{Var}(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} n\theta_0(1-\theta_0) = \frac{1}{4n}$$

as $X_i \sim \text{Bin}(n, \theta_0)$ and $\theta_0 = \frac{1}{2}$. Wald's test statistic is then given by

$$W = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\theta_0)}} = 2\sqrt{n}(\hat{\theta} - \theta_0) \approx -1.75.$$

The p-value is $2\Phi(|W|) \approx 0.08$. So we do not reject the null-hypothesis.

The confidence interval for which we reject the null-hypothesis is

$$C = (\hat{\theta} - \hat{se}z_{0.025}, \hat{\theta} + \hat{se}z_{0.025}) = (0.480, 0.503),$$

as $\hat{se} = \sqrt{\hat{\theta}(1 - \hat{\theta})/n}$.

Solution 10.6 - Alternative solution

We can calculate the probability directly from the biomial distribution. The null hypothese $H_0 : \theta_0 = \frac{1}{2}$. Under H_0, $X \sim \text{Bin}(n, \theta_0)$, where $n = 1919$. Then the p-value is $P(X \geq 922) + P(X \leq 1919/2 - 37.5) \approx 0.087$. So we don't reject the null-hypothesis.

We reject the null-hypothesis if

$$X \notin (z_{0.025}, z_{0.975}) = (917, 1002).$$

Solution 10.7

See code for explicit calculations.

(a) We use the Wald test on $H_0 : \overline{\Delta} = \overline{X} - \overline{Y} = 0$, where X are the measurements of Twain and Y the measurements of Snodgrass. The Wald test gives us a p-value of approximately 0.00008, so we reject the null hypothesis that the essays of Twain and Snodgrass are the same. The confidence interval of $\overline{\Delta}$ is aproximately $(0.011, 0.033)$.

(b) Using the permutation test on $T(X_1, ..., X_8, Y_1, ..., Y_{10}) = |\overline{X} - \overline{Y}|$. The approximate p-value is 0.00089. We reject the null hypothesis that the distributions of X and Y are the same.

Solution 10.8

(a) We have to find c such that

$$\alpha = P_0\left(\frac{1}{n}\sum_{i=1}^{n} X_i > c\right).$$

Note that under H_0, $E(\frac{1}{n}\sum_{i=1}^{n} X_i) = 0$ and $\text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{\sqrt{n}}$. So

$$P_0\left(\frac{1}{n}\sum_{i=1}^{n} X_i > c\right) = P(Z > \sqrt{n}c) = \Phi(-\sqrt{n}c).$$

Take $c = -\Phi^{-1}(\alpha)/\sqrt{n}$.

(b) Under H_1, $E(\frac{1}{n}\sum_{i=1}^{n} X_i) = 1$ and $\text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{\sqrt{n}}$. So

$$\beta(1) = P_1\left(\frac{1}{n}\sum_{i=1}^{n} X_i > c\right) = P(Z > (c - 1)\sqrt{n}) = \Phi((1 - c)\sqrt{n}).$$

(c) Note that for fixed α,

$$\beta(1) = \Phi(\sqrt{n} - c\sqrt{n}) = \Phi(\sqrt{n} + \Phi^{-1}(\alpha)).$$

As $\Phi^{-1}(\alpha)$ is fixed $\beta(1) \to 1$ as $n \to \infty$.

Solution 10.9

$$\begin{aligned}\beta(\theta_1) &= P_{\theta_1}(|Z| < z_{\frac{\alpha}{2}}) \\ &= 1 - P_{\theta_1}(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) \\ &= 1 - P_{\theta_1}\left(-z_{\frac{\alpha}{2}} < \frac{\hat{\theta} - \theta_0}{\hat{se}} < z_{\frac{\alpha}{2}}\right) \\ &= 1 - P_{\theta_1}\left(-z_{\frac{\alpha}{2}} < \frac{\hat{\theta} - \theta_1}{\hat{se}} + \frac{\theta_1 - \theta_0}{\hat{se}} < z_{\frac{\alpha}{2}}\right).\end{aligned}$$

As $n \to \infty$, $\frac{\hat{\theta}-\theta_1}{\hat{se}} \to 0$ and $\frac{\theta_1-\theta_0}{\hat{se}} \to \infty$, because $\hat{se} \to 0$ and $\theta_1 > \theta_0$. Therefore $\beta(\theta_1) \to 1$ as $n \to \infty$.

Solution 10.10

See code. We use three different tests: Walk test per week, binomial test per week, and the χ^2-test. All three of them suggest that in week -1 and 1 a significant change in elderly Chinese woman took place.

Solution 10.11

See code. The only drugs that seems to have any significant effect is Chlorpromazine.

Solution 10.12

(a) Let $X_1, X_2, ..., X_n \sim \text{Poison}(\lambda_0)$. Let $\hat{\lambda} = \overline{X}$ be the MLE of X. Note that

$$\text{Var}(\hat{\lambda}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\lambda_0}{n}.$$

So the Wald estimate is

$$W = \frac{\hat{\lambda} - \lambda_0}{\hat{se}} = \sqrt{n}\frac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0}}.$$

We reject $H_0 : \lambda = \lambda_0$ if $|W| > z_{\frac{\alpha}{2}}$.

(b) See code. For $n = 20$ we approximate that 5.62% of all tests is rejected. This is not exactly equal to 5% as the Poisson distribution is only approximately similar to the normal distribution. When $n \to \infty$ the rejection rate will go to exactly 5%.

Solution 10.13

Let $X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$. We will construct a likelihood ratio test for null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Let $\hat{\mu} = \overline{X}$ be the maximum likelihood estimator for μ. Note that

$$\ell(\mu) = -\frac{n}{2} - \log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

So

$$\begin{aligned}\lambda &= 2\ell(\hat{\mu}) - 2\ell(\mu_0) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(X_i - \mu_0)^2 - (X_i - \hat{\mu})^2\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu_0^2 - 2(\mu_0 - \hat{\mu}) - \hat{\mu}^2) \\ &= \frac{1}{\sigma^2}(n\mu_0^2 - 2n(\mu_0 - \hat{\mu})\hat{\mu} - n\hat{\mu}^2) = n\left(\frac{\hat{\mu} - \mu_0}{\sigma}\right)^2.\end{aligned}$$

Observe that the likelihood ration and Wald statistics are related, because $\lambda = W^2$.

Solution 10.14

Let $X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$. We will construct a likelihood ratio test for null hypothesis $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma \neq \sigma_0$. Let $\hat{\sigma} = \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2$ be the maximum likelihood estimator for σ. Note that

$$\ell(\sigma) = -\frac{n}{2} - \log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

So

$$\begin{aligned}\lambda &= 2\ell(\hat{\sigma}) - 2\ell(\sigma_0) \\ &= 2\log(\hat{\sigma}) - 2\log(\sigma_0) + 2\left(\frac{1}{2\sigma_0} - \frac{1}{2\hat{\sigma}}\right)\sum_{i=1}^{n}(X_i - \mu)^2 \\ &= 2\log\left(\frac{\hat{\sigma}}{\sigma_0}\right) + \frac{\hat{\sigma}^2 - \sigma_0^2}{\sigma_0^2\hat{\sigma}^2}n\hat{\sigma}^2 = 2\log\left(\frac{\hat{\sigma}}{\sigma_0}\right) + n - n\left(\frac{\sigma_0}{\hat{\sigma}}\right)^2.\end{aligned}$$

Chapter 11 - Bayesian Inference

Solution 11.1

Let $X_1, X_2, ..., X_n \sim \text{Normal}(\theta, \sigma^2)$ with σ known. As prior take $\theta \sim \text{Normal}(a, b^2)$. By Bayes' Theorem we have

$$f(\theta|X^n) = \frac{f(X^n|\theta)f(\theta)}{\int f(X^n|\theta)f(\theta)d\theta} \propto \mathcal{L}_n(\theta)f(\theta),$$

where

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i|\theta) = \frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2\right),$$

$$f(\theta) = \frac{1}{b\sqrt{2\pi}}\exp\left(-\frac{1}{2b^2}(\theta - a)^2\right).$$

With a tedious calculation we get

$$\begin{aligned}\mathcal{L}_n(\theta)f(\theta) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2 - \frac{1}{2b^2}(\theta - a)^2\right) \\ &\propto \exp\left(-\frac{(nb^2 + \sigma^2)\theta^2 - 2(n\overline{X}b^2 + a\sigma^2)\theta}{2\sigma^2 b^2}\right) \\ &= \exp\left(-\frac{(nb^2 + \sigma^2)}{2\sigma^2 b^2}\left(\theta^2 - 2\frac{nb^2\overline{X} + a\sigma^2}{nb^2 + \sigma^2}\theta\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)\left(\theta - \frac{b^2\overline{X} + a\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}}\right)^2\right) \\ &= \exp\left(-\frac{1}{2\tau^2}(\theta - \overline{\theta})^2\right).\end{aligned}$$

Therefore, $\theta|X^n \sim \text{Normal}(\theta, \overline{\theta})$.

Solution 11.2

Let $X_1, X_2, ..., X_n \sim \text{Normal}(\mu, 1)$.

(a) See code.

(b) Let $f(\mu) = 1$, then

$$f(\mu|X^n) \propto f(X^n|\mu)f(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \mu)^2\right) \propto \exp\left(-\frac{n}{2}(\mu - \overline{X})^2\right),$$

such that we have $\mu|X^n \sim \text{Normal}(\overline{X}, \frac{1}{n})$.

(c) See code.

(d) Let $\theta = e^\mu$, we calculate the cumulative distribution function

$$\begin{aligned} P_\theta(\theta < T|X^n) &= P_\theta(e^\mu < T|X^n) \\ &= P_\theta(\mu < \log(T)|X^n) \\ &= P_\theta(Z < \sqrt{n}(\log(T) - \overline{X})) \\ &= \Phi(\sqrt{n}(\log(T) - \overline{X})). \end{aligned}$$

Differentiate with respect to T gives the probability density function

$$f(\theta|X^n) = \frac{\sqrt{n}}{\theta}\phi\left(\sqrt{n}(\log(\theta) - \overline{X})\right).$$

For the simulation see code.

(e) We have $\mu|X^n \sim \text{Normal}(\overline{X}, \frac{1}{n})$, so the 95% posterior interval is

$$\left(\overline{X} - \frac{z_{0.025}}{\sqrt{n}}, \overline{X} + \frac{z_{0.025}}{\sqrt{n}}\right).$$

(f) We use the Delta method. Let $\theta = g(\mu) = e^\mu$, then $g'(\mu) = e^\mu$, so $\hat{se}(\hat{\theta}) = |g'(\hat{\mu})|\hat{se}(\hat{\mu})$. The 95% confidence interval is given by

$$\left(e^{\overline{X}} - z_{0.025}\frac{e^{\overline{X}}}{\sqrt{n}}, e^{\overline{X}} + z_{0.025}\frac{e^{\overline{X}}}{\sqrt{n}}\right) = \left(e^{\overline{X}}(1 - \frac{z_{0.025}}{\sqrt{n}}), e^{\overline{X}}(1 + \frac{z_{0.025}}{\sqrt{n}})\right).$$

Solution 11.3

Let $X_1, X_2, ..., X_n \sim \text{Uniform}(0, \theta)$ and prior $f(\theta) \propto 1/\theta$. The likelihood is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(X_i \leq \theta) = \frac{1}{\theta^n} I(X^+ \leq \theta).$$

The posterior becomes

$$\theta|X^n \propto \mathcal{L}_n(\theta)f(\theta) \propto \frac{1}{\theta^{n+1}} I(X^+ \leq \theta).$$

Finally, to calculate the normalization coefficient

$$\int_{-\infty}^{\infty} \frac{1}{\theta^{n+1}} I(X^+ \leq \theta) d\theta = \int_{X^+}^{\infty} \frac{1}{\theta^{n+1}} d\theta = -\frac{1}{n\theta^n}\Big|_{\theta = X^+}^{\infty} = \frac{1}{n(X^+)^n}.$$

So, the probability density function is

$$f(\theta|X^n) = \frac{1}{n}\frac{1}{(X^+)^n}\frac{1}{\theta^{n+1}},$$

for $\theta \geq X^+$.

Solution 11.4

Let $n_1 = 50 = n_2$, $x_1 = 30$ and $x_2 = 40$. Let $X_1 \sim \text{Binomial}(n_1, p_1)$ and $X_2 \sim \text{Binomial}(n_2, p_2)$. Define $\tau = p_1 - p_2$.

(a) To find MLE $\hat{\tau} = g(\hat{p}_1, \hat{p}_2) = \hat{p}_1 - \hat{p}_2$ we first find MLE $\hat{p}_1$ and $\hat{p}_2$. Calculate the likelihood estimator for p_i,

$$\mathcal{L}_n(p_i) = \binom{n_i}{X_i} p_i^{X_i}(1-p_i)^{n_i - X_i}.$$

The logarithm of the likelihood estimator is

$$\ell_n(p_i) = X_i \log(p_i) - (n_i - X_i)\log(1-p_i) + C,$$

where C is independent of p_i. Maximize $\ell_n(p_i)$ by taking setting the derivative to zero,

$$0 = \ell_n'(p_i) = \frac{X_i}{p_i} - \frac{n_i - X_i}{1-p_i} \rightarrow p_i = \frac{X_i}{n_i}.$$

Therefore, the MLE is $\hat{p}_1 = \frac{x_1}{n_1} = \frac{3}{5}$ and $\hat{p}_2 = \frac{x_2}{n_2} = \frac{4}{5}$, and the MLE for τ is given by $\hat{\tau} = g(\hat{p}_1, \hat{p}_2) = \hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2} = -\frac{1}{5}$. The Fisher information matrix for p_i is

$$I(p_i) = E(-s(p_i)) = E\left(-\frac{\partial^2 \ell(p_i)}{\partial p_i^2}\right) = E\left(\frac{X_i}{p_i^2} - \frac{n_i - X_i}{(1-p_i)^2}\right) = \frac{n_i p_i}{p_i^2} - \frac{n_i - n_i p_i}{(1-p_i)^2} = \frac{n}{p_i(1-p_i)}.$$

So we have $\hat{se}(\hat{p}_i)^2 = I(p_i)^{-1} = \frac{p_i(1-p_i)}{n}$. With the Delta method, $\tau = g(p_1, p_2) = p_1 - p_2$. $\nabla g = (1, -1)^t$ and

$$I(p_1, p_2) = \begin{pmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{pmatrix}, \quad J(p_1, p_2) = \begin{pmatrix} \frac{p_1(1-p_1)}{n_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} \end{pmatrix}.$$

We have

$$\hat{se}(\hat{\tau})^2 = \nabla\hat{g}^t \hat{J}(\hat{p}_1, \hat{p}_2)\nabla\hat{g} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}.$$

The 90% confidence interval is therefore given by

$$(\hat{\tau} - z_{0.05}\hat{se}(\hat{\tau}), \hat{\tau} + z_{0.05}\hat{se}(\hat{\tau})) \approx (-0.35, -0.05).$$

(b) See code.

(c) Use $f(p_1, p_2) = 1$ as prior, then

$$\begin{aligned} f(p_1, p_2 | X_1, X_2) &\propto f(X_1|p_1) f(X_2|p_2) f(p_1, p_2) \\ &\propto p_1^{X_1}(1-p_1)^{n_1 - X_1} p_2^{X_2}(1-p_2)^{n_2 - X_2} \\ &\propto f(p_1|X_1) f(p_2|X_2), \end{aligned}$$

so $p_1|X_1 \sim \text{Beta}(X_1 + 1, n_1 - X_1 + 1)$ and $p_2|X_2 \sim \text{Beta}(X_2 + 1, n_2 - X_2 + 1)$. See code for simulation.

(d) Let

$$\psi = g(p_1, p_2) = \log\left(\left(\frac{p_1}{1-p_1}\right) / \left(\frac{p_2}{1-p_2}\right)\right).$$

The MLE is given by $\hat{psi} = g(\hat{p}_1, \hat{p}_2)$. We use the Delta method to calculate the confidence interval. The gradient of ψ is (skipping the details)

$$\frac{\partial g(p_1, p_2)}{\partial p_1} = \frac{1}{p_1(1-p_1)}, \quad \frac{\partial g(p_1, p_2)}{\partial p_2} = -\frac{1}{p_2(1-p_2)}.$$

Such that

$$\mathrm{se}(\psi)^2 = \nabla g^t J(p_1, p_2) \nabla g = \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}.$$

The 90% confidence interval is given by

$$(\hat{\psi} - z_{0.05}\hat{se}(\hat{\psi}), \hat{psi} + z_{0.05}\hat{se}(\hat{\psi})) \approx (-1.44, 0.06).$$

(e) See code. We can use the algorithm above, replacing τ with ψ.

Solution 11.5

Let $X_1, X_2, ..., X_n \sim \mathrm{Bernoulli}(p)$. Take prior $p \sim \mathrm{Beta}(a, b)$. The posterior $p|X^n$ is

$$f(p|X^n) \sim \mathcal{L}_n(p) f(p) = \left(\prod_{i=1}^n p^{X_i}(1-p)^{1-X_i}\right) p^{a-1}(1-p)^{b-1} = p^{\sum X_i + a - 1}(1-p)^{n - \sum X_i + b - 1},$$

such that $p|X^n \sim \mathrm{Beta}(\sum X_i + a, \sum X_i + b)$. See code for probability density function plots of $p|X^n$.

Solution 11.6

Let $X_1, X_2, ..., X_n \sim \mathrm{Poisson}(\lambda)$.

(a) Take prior $\lambda \sim \mathrm{Gamma}(a, b)$. The likelihood is

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^n \frac{\lambda^{X_i}}{X_i!} e^{-\lambda} \propto \lambda^{\sum X_i} e^{-n\lambda}.$$

The posterior becomes

$$\lambda | X^n \propto \mathcal{L}_n(\lambda) f(\lambda) \propto \lambda^{\sum X_i} e^{-n\lambda} \lambda^{\alpha - 1} e^{-\lambda\beta} = \lambda^{\sum X_i + \alpha - 1} e^{-\lambda(n+\beta)},$$

so that $\lambda|X^n \sim \mathrm{Gamma}(\sum X_i + \alpha, n + \beta)$. The posterior mean is

$$E(\lambda|X^n) = \frac{\sum X_i + \alpha}{n + \beta}.$$

(b) Let $f(\lambda) \propto \sqrt{I(\lambda)}$ be Jeffrey's prior. We calculate

$$\ell_n(\lambda) = \sum_{i=1}^n \log(\lambda) X_i - n\lambda, \quad \ell_n'(\lambda) = \sum_{i=1}^n \frac{X_i}{\lambda} - n, \quad \ell_n''(\lambda) = -\sum_{i=1}^n \frac{X_i}{\lambda^2}.$$

The Fisher information matrix becomes

$$I(\lambda) = \frac{1}{n} I_n(\lambda) = -\frac{1}{n} E_\lambda(\ell_n''(\lambda)) = \frac{1}{\lambda}.$$

So we take prior $f(\lambda) \propto \frac{1}{\sqrt{\lambda}}$, and get posterior

$$\lambda|X^n \propto \mathcal{L}_n(\lambda) f(\lambda) \propto \lambda^{\sum X_i - \frac{1}{2}} e^{-n\lambda},$$

so that $\lambda|X^n \sim \mathrm{Gamma}\left(\sum X_i + \frac{1}{2}, n\right)$.

Solution 11.7

Let $\theta = (\theta_1, \theta_2, ..., \theta_n) \in \mathbb{R}^n$ and $\xi = (\xi_1, \xi_2, ..., \xi_n) \in \mathbb{R}^n$. Let $X_i \sim \text{Uniform}(\{1, 2, ..., B\})$ with $B >> 0$, $R_i \sim \text{Bernoulli}(\xi_{X_i})$, and $Y_i \sim \text{Bernoulli}(\theta_{X_i})$ if $R_i = 1$ and else don't draw Y_i. We frequentist approximate $\psi = P(Y = 1)$ with

$$\hat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\xi_i}.$$

Note that

$$\begin{aligned} E\left(\frac{R_i Y_i}{\xi_{X_i}}\right) &= E\left(E\left(\frac{R_i Y_i}{\xi_{X_i}} \middle| X_i\right)\right) \\ &= \sum_{j=1}^{B} \frac{1}{\xi_j} P(R_i|X_i = j) P(Y_i|X_i = j) P(X_i = j) \\ &= \sum_{j=1}^{B} \frac{1}{\xi_j} \xi_j \theta_j \frac{1}{B} \\ &= \frac{1}{B} \sum_{j=1}^{B} \theta_j = \psi. \end{aligned}$$

Furthermore,

$$\begin{aligned} E\left(\left(\frac{R_i Y_i}{\xi_{X_i}}\right)^2\right) &= E\left(E\left(\left(\frac{R_i Y_i}{\xi_{X_i}}\right)^2 \middle| X_i\right)\right) \\ &= \sum_{j=1}^{B} \frac{1}{\xi_j^2} P(R_i^2|X_i = j) P(Y_i^2|X_i = j) P(X_i = j) \\ &= \sum_{j=1}^{B} \frac{1}{\xi_j^2} \xi_j \theta_j = \frac{1}{B} \sum_{j=1}^{B} \frac{\theta_j}{\xi_j}. \end{aligned}$$

Take δ such that $\frac{\theta_j}{\xi_j} \leq \frac{1}{\delta^2}$ for all j. The variance is bounded by

$$\begin{aligned} V(\hat{\psi}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{n} \frac{R_i Y_i}{\xi_{X_i}}\right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \left(E\left(\left(\frac{R_i Y_i}{\xi_{X_i}}\right)^2\right) - E\left(\frac{R_i Y_i}{\xi_{X_i}}\right)^2\right) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{1}{B} \sum_{j=1}^{B} \frac{\theta_j}{\xi_j} - \left(\frac{1}{B} \sum_{j=1}^{B} \theta_j\right)^2\right) \\ &\leq \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{B} \sum_{j=1}^{B} \frac{1}{\delta^2} = \frac{1}{n\delta^2}. \end{aligned}$$

Solution 11.8

Let $X \sim \text{Normal}(\mu.1)$. We will test $H_0 : \mu = 0$ against the alternative hypothesis $H_1 : \mu \neq 0$. As prior of the tests take $P(H_0) = P(H_1) = \frac{1}{2}$. And let the prior of μ under H_1 be $\mu \sim \text{Normal}(0, b^2)$. From Bayes'

Theorem we have

$$P(H_0|X^n) = \frac{\mathcal{L}_n(\mu_0)}{\mathcal{L}_n(0) + \int \mathcal{L}_n(\mu) f(\mu) d\mu} = \frac{1}{1 + \int \frac{\mathcal{L}_n(\mu)}{\mathcal{L}_n(0)} f(\mu) d\mu}.$$

We calculate the nominator and denominator in the integral seperately

$$\mathcal{L}_n(0) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2\right),$$

and

$$\mathcal{L}_n(\mu) f(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \frac{1}{b} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^n (X_i - \mu)^2 + \frac{\mu^2}{b^2}\right)\right).$$

Expanding and factorizing μ in the exponential term in $\mathcal{L}_n(\mu)f(\mu)$ gives

$$\sum_{i=1}^n (X_i - \mu)^2 + \frac{\mu^2}{b^2} = \sum_{i=1}^n X_i^2 - 2n\overline{X}\mu + \left(n + \frac{1}{b^2}\right)\mu^2 = \frac{1}{\sigma^2}\left(\mu - \sigma^2 n\overline{X}\right)^2 - \sigma^2 n^2 \overline{X}^2 + \sum_{i=1}^n X_i^2,$$

where $\sigma^2 = \frac{1}{n+1/b^2} = \frac{b^2}{1+nb^2}$. This brings us to

$$\mathcal{L}_n(\mu) f(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \frac{1}{b} \exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2 + \frac{1}{2}\sigma^2 n^2 \overline{X}^2\right) \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \sigma^2 n\overline{X}\right)^2\right).$$

With these expressions we can calculate the integral

$$\begin{aligned}
\int \frac{\mathcal{L}_n(\mu)}{\mathcal{L}_n(0)} f(\mu) d\mu &= \int \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \frac{1}{b} \exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2 + \frac{1}{2}\sigma^2 n^2 \overline{X}^2\right) \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \sigma^2 n\overline{X}\right)^2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n X_i^2\right)} d\mu \\
&= \int \frac{1}{\sqrt{2\pi}} \frac{1}{b} \exp\left(\frac{1}{2}\sigma^2 n^2 \overline{X}^2\right) \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \sigma^2 n\overline{X}\right)^2\right) d\mu \\
&= \frac{\sigma}{b} \exp\left(\frac{1}{2}\sigma^2 n^2 \overline{X}^2\right) \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\mu - \sigma^2 n\overline{X}\right)^2\right) d\mu \\
&= \frac{\sigma}{b} \exp\left(\frac{1}{2}\sigma^2 n^2 \overline{X}^2\right) \\
&= \sqrt{\frac{b}{1+n^2b^2}} \exp\left(\frac{bn^2\overline{X}^2}{2(1+n^2b^2)}\right) \\
&= \sqrt{\frac{b}{1+n^2b^2}} \exp\left(\frac{b\overline{X}^2}{2(\frac{1}{n^2} + b^2)}\right).
\end{aligned}$$

Therefore,

$$P(H_0|X^n) = \frac{1}{1 + \sqrt{\frac{b}{1+n^2b^2}} \exp\left(\frac{b\overline{X}^2}{2(\frac{1}{n^2}+b^2)}\right)} \longrightarrow 1,$$

when $n \to \infty$. In other words, $P(H_0|X^n)$ goes to 1 when n goes to infinity, no matter if H_0 is true or false. This is called the Jeffreys-Lindley paradox.

Frequentist would use the Wald's test instead, i.e.,

$$W = \frac{\overline{X}_n}{\hat{se}} = \sqrt{n}\,\overline{X}_n,$$

and reject H_0 when $|W| \geq z_{\alpha/2}$.

Chapter 12 - Statistical Decision Theory

Solution 12.1

Note that the definition for Bayes risk is wrong. The Bayes risk should be

$$r(f, \hat{p}) = \int L(p, \hat{p}) f(p) dp.$$

(a) $X \sim \text{Binomial}(n, p)$, $p \sim \text{Beta}(\alpha, \beta)$. Bayes risk is given by

$$\begin{aligned}
r(f, \hat{p}) &= \int L(p, \hat{p}) f(p) dp \\
&= \int (p - \hat{p})^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int (p^2 - 2p\hat{p} + \hat{p}^2) p^{\alpha-1}(1-p)^{\beta-1} dp \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} - 2\hat{p}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} + \hat{p}^2 \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \right) \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{2\alpha}{\alpha+\beta}\hat{p} + \hat{p}^2 \\
&= \frac{\alpha(\alpha+1) - 2\alpha(\alpha+\beta+1)\hat{p} + (\alpha+\beta+1)(\alpha+\beta)\hat{p}^2}{(\alpha+\beta)(\alpha+\beta+1)}.
\end{aligned}$$

According to Theorem 12.8, the Bayes estimator is given by

$$\hat{\theta}(x^n) = \int \theta f(\theta | x^n) dx^n = E(\theta | X^n = x^n).$$

We calculate

$$\begin{aligned}
\theta | X^n \propto \mathcal{L}_\theta(X^n) f(\theta) &= \left(\prod_{i=1}^m \binom{n}{X_i} p^{X_i}(1-p)^{n-X_i} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{\alpha + \sum X_i - 1}(1-p)^{\beta + mn - \sum X_i - 1} \\
&\sim \text{Gamma}\left(\alpha + \sum_{i=1}^n X_i, \beta + mn - \sum_{i=1}^n X_i \right).
\end{aligned}$$

As the mean of $X \sim \text{Gamma}(\alpha, \beta)$ is $E(X) = \frac{\alpha}{\alpha+\beta}$, the Bayes estimator is

$$\hat{\theta}(x^n) = E(\theta | X^n = x^n) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + mn}.$$

(b) $X \sim \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(\alpha, \beta)$. We use

$$\int \lambda^{a-1} e^{-b\lambda} d\lambda = \int b^{-a+1} y^{a-1} e^{-y} \frac{1}{b} dy = \frac{1}{b^a} int y^{a-1} e^{-y} dy = \frac{\Gamma(a)}{b^a},$$

where we substituted $y = b\lambda$, to show that the Bayes risk is given by

$$\begin{aligned}
r(f, \hat{\lambda}) &= \int L(\lambda, \hat{\lambda}) f(\lambda) d\lambda \\
&= \int (\lambda - \hat{\lambda})^2 \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \left[\int \lambda^{2+\alpha-1} e^{-\beta\lambda} d\lambda - 2\hat{\lambda} \int \lambda^{1+\alpha-1} e^{\beta\lambda} d\lambda + \int \lambda^{\alpha-1} e^{\beta\lambda} d\lambda \right] \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \left[\frac{\Gamma(\alpha+2)}{\beta^{\alpha+2}} - 2\hat{\lambda} \frac{\Gamma(\alpha+1)}{\beta^{\alpha+1}} + \hat{\lambda}^2 \frac{\Gamma(\alpha)}{\beta^\alpha} \right] \\
&= \frac{\alpha(\alpha+1)}{\beta^2} - 2\hat{\lambda}\frac{\alpha}{\beta} + \hat{\lambda}^2.
\end{aligned}$$

For the Bayes estimator we use Theorem 12.8 such that

$$\begin{aligned}
f(\lambda | X^n) \propto \mathcal{L}_n(\lambda) f(\lambda) &= \left(\prod_{i=1}^n \frac{1}{X_i!} \lambda^{X_i} e^{-\lambda} \right) \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
&\propto \lambda^{\sum X_i + \alpha - 1} e^{-(\beta+n)\lambda} \sim \text{Gamma}\left(\sum_{i=1}^n X_i + \alpha, \beta + n \right).
\end{aligned}$$

(c) $X \sim \text{Normal}(\theta, \sigma^2)$, σ^2 known, $\theta \sim \text{Normal}(a, b^2)$. Bayes risk is

$$\begin{aligned}
r(f, \hat{\theta}) &= \int (\hat{\theta} - \theta)^2 f(\theta) d\theta \\
&= \int (\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) \frac{1}{2\pi} \frac{1}{b} \exp\left(-\frac{1}{2b^2} (\theta - a)^2 \right) d\theta \\
&= \hat{\theta}^2 - 2\hat{\theta} E(\theta) + E(\theta^2) \\
&= \hat{\theta}^2 - 2\hat{\theta} a + a^2 + b^2,
\end{aligned}$$

as $V(\theta) = E(\theta^2) - E(\theta)^2$, we have $E(\theta^2) = V(\theta) - E(\theta)^2 = a^2 + b^2$. Bayes estimator is, using Theorem 12.8,

$$\begin{aligned}
f(\theta | X^n) &\propto \mathcal{L}_n(\theta) f(\theta) \\
&= \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2} (X_i - \theta)^2 \right) \right) \frac{1}{\sqrt{2\pi}} \frac{1}{b} \exp\left(-\frac{1}{2b^2} (\theta - a)^2 \right) \\
&\propto \exp\left(-\frac{1}{2} \left(\left(\frac{n}{\sigma^2} + \frac{1}{b^2} \right) \theta^2 - (2n\overline{X} - 2a)\theta \right) \right) \\
&\propto \exp\left(-\frac{1}{2\hat{\sigma}^2} \left(\theta - \hat{\sigma}^2 (n\overline{X} - a) \right)^2 \right) \sim \text{Normal}(\hat{\mu}, \hat{\sigma}^2),
\end{aligned}$$

where

$$\frac{1}{\hat{\sigma}^2} = \left(\frac{n}{\sigma^2} + \frac{1}{b^2} \right), \quad \hat{\mu} = \hat{\sigma}^2 (n\overline{X} - a).$$

Solution 12.2

Let $X_1, X_2, ..., X_n \sim \text{Normal}(\theta, \sigma^2)$, and θ is estimated with loss function $L(\theta, \hat{\theta}) = \frac{1}{\sigma^2}(\theta - \hat{\theta})^2$. Note that

$$R(\theta, \hat{\theta}) = E\left(\frac{1}{\sigma^2} (\theta - \hat{\theta})^2 \right) = \frac{1}{\sigma^2} E\left((\theta - \hat{\theta})^2 \right) = \frac{1}{\sigma^2} \left(V_\theta(\hat{\theta}) + \text{bias}_\theta^2(\hat{\theta}) \right) = \frac{1}{\sigma^2} R_{\text{MSE}}(\theta, \hat{\theta}).$$

By Theorem 12.20, $\overline{X}$ is admissible under $R_{\mathcal{MSE}}$, i.e., There is no $\hat{\theta}' \neq \overline{X}$ such that

$$R_{\text{MSE}}(\theta, \hat{\theta}') \leq R_{\text{MSE}}(\theta, \overline{X}),$$

for all θ, and

$$R_{\text{MSE}}(\theta, \hat{\theta}') < R_{\text{MSE}}(\theta, \overline{X}),$$

for at least one θ. When we replace $R_{\text{MSE}} \leftarrow \frac{1}{\sigma^2} R_{\text{MSE}} = R$ in the two equations above, we see that R is admissible. Moreover,

$$R(\theta, \overline{X}) = \frac{1}{\sigma^2} R_{\text{MSE}}(\theta, \overline{X}) = \frac{1}{\sigma^2} \frac{\sigma^2}{n} = \frac{1}{n},$$

is constant. Therefore, by Theorem 12.21, $\overline{X}$ is minimax.

Solution 12.3

Let $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$ be a finite parameter space. The zero-one loss is given by $L(\theta, \hat{\theta}) = 0$ if $\theta = \hat{\theta}$ and $L(\theta, \hat{\theta}) = 1$ otherwise. The posterior risk is defined by

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) dx = \int \left(1 - I_{\theta, \hat{\theta}(x)}\right) f(\theta|x) dx = 1 - \sum_{\theta = \hat{\theta}(x)} f(\theta|x) dx = 1 - f(\hat{\theta}(x)|x),$$

where we use that Θ is finite. The Bayes estimator $\hat{\theta}$ is the minimal value of the posterior risk $r(\hat{\theta}|x)$. By Theorem 12.8, the Bayes estimator, under zero-one loss L, is the mode of the posterior $f(\theta|x)$. The mode of the posterior is not defined in the book, so I give it here. In the discrete case, the mode of the posterior is the value $\hat{\theta}_{\text{MAP}}(x)$ at wich the probability mass function of the posterior $f(\theta|x)$ takes it's maximum value, i.e.,

$$\hat{\theta}_{\text{MAP}}(x) = \text{argmax}_{\theta} f(\theta|x),$$

where MAP estimator stands for Maximum A Posteriori estimator. Notice that $\hat{\theta}_{\text{MAP}}(x)$ is the minimal value for $1 - f(\theta|x)$, so $\hat{\theta}_{\text{MAP}}(x)$ is the Bayes estimator.

Solution 12.4

Let $X_1, X_2, ..., X_n$ be samples from a distribution with variance σ^2. Consider the estimator bS^2, where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_i)^2.$$

Define loss

$$L(\sigma^2, \hat{\sigma}^2) = \frac{\hat{\sigma}^2}{\sigma^2} - \log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) - 1.$$

To find the optimal b that minimizes the risk of L simplify

$$L(\sigma^2, bS^2) = b\frac{S^2}{\sigma^2} - \log(b) - \log\left(\frac{S^2}{\sigma^2}\right) - 1.$$

The risk becomes

$$R(\sigma^2, bS^2) = E_{\sigma^2}(L(\sigma^2, bS^2)) = b - \log(b) + C,$$

where C is a term independent of b. Differentiating $f(x) = x - \log(x)$ and setting the differential to zero gives $b = 1$. So $R(\sigma^2, bS^2)$ is minimal if $b = 1$.

Solution 12.5

Let $X \sim \text{Binomial}(n, p)$. Define loss function

$$L(p, \hat{p}) = \left(1 - \frac{\hat{p}}{p}\right)^2,$$

where $0 < p < 1$. Take estimator $\hat{p}(X) = 0$. We calculate the risk

$$R(p, \hat{p}) = E_p(L(p, \hat{p})) = E_p\left(\left(1 - \frac{\hat{p}}{p}\right)^2\right) = E_p\left(1 - 2\frac{\hat{p}}{p} + \frac{\hat{p}^2}{p^2}\right) = 1 - \frac{2}{p}E_p(\hat{p}) + \frac{1}{p^2}E(\hat{p}^2).$$

When $\hat{p}(X) = 0$ we have $R(p, \hat{p}(X)) = 1$. Let $\hat{p}'(X) > 0$, then $E(\hat{p}'(X)) > 0$. Take p such that

$$0 < p < \frac{1}{2}\frac{E(\hat{p}'(X)^2)}{E(\hat{p}'(X))}.$$

In particular we have

$$\frac{1}{p^2}E(\hat{p}'(X)^2) - \frac{2}{p}E(\hat{p}'(X)) > 0,$$

such that $R(p, \hat{p}'(X)) > 1 = R(p, \hat{p}(X))$. In other words, $\hat{p}(X)$ is minimax.

Solution 12.6

See code.

Part III

Statistical Models and Methods

Chapter 13 - Linear and Logistic Regression

Solution 13.1

I tried to calculate the equations directly from the formula, but it's significantly easier to do it with matrix differentiation. Therefore, we develop a set of matrix differentiation tools.

For the rest of this section, let $\psi : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$, such that $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}) = \psi(\boldsymbol{x})$. We define

$$\frac{\partial \psi(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \left(\frac{\partial y_i}{\partial x_j}\right)_{\substack{1\le i\le m\\ 1\le j\le n}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

Theorem 1. *If $\psi = A \in \mathbb{R}^{m\times n}$, such that $\boldsymbol{y}(\boldsymbol{x}) = A\boldsymbol{x}$, and A is independent of $\boldsymbol{x}$, then $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = A$.*

Proof. We have $y_i = \sum_{j=1}^{n} a_{ij}x_j$, so $\frac{\partial y_i}{\partial x_j} = a_{ij}$, and therefore, $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = A$. □

Theorem 2. *Let $\boldsymbol{y}(\boldsymbol{x},\boldsymbol{z}) = A\boldsymbol{x}(\boldsymbol{z})$, where $\boldsymbol{x}$ is a vector values function in vector $\boldsymbol{z}$, and A is independent of $\boldsymbol{x}$ and $\boldsymbol{z}$, then $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} = A\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$.*

Proof. We have $y_i = \sum_{j=1}^{n} a_{ij}x_j(\boldsymbol{z})$, so $\frac{\partial y_i}{\partial z_j} = \sum_{k=1}^{n} a_{ik}\frac{\partial x_k}{\partial z_j} = \left(A\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}\right)_{ij}$. Hence, $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} = A\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$. □

Theorem 3. *Let $\alpha(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y}^tA\boldsymbol{x}$ be a real valued function, where A is a matrix indepedent of $\boldsymbol{x}$ and $\boldsymbol{y}$. We have $\frac{\partial \alpha}{\partial \boldsymbol{x}} = \boldsymbol{y}^tA$ and $\frac{\partial \alpha}{\partial \boldsymbol{y}} = (A\boldsymbol{x})^t$.*

Proof. Let $\boldsymbol{z}^t = \boldsymbol{y}^tA$, so that $\alpha = \boldsymbol{z}^t\boldsymbol{x}$. By Theorem 1 we have $\frac{\partial \alpha}{\partial \boldsymbol{x}} = \boldsymbol{z}^t = \boldsymbol{y}^tA$. Because α is a scalar, $\alpha = \alpha^t = \boldsymbol{x}^tA\boldsymbol{y}$. Applying the same method for $\boldsymbol{x}^tA$ we get $\frac{\partial \alpha}{\partial \boldsymbol{y}} = \frac{\partial \alpha^t}{\partial \boldsymbol{y}} = \boldsymbol{x}^tA^t = (A\boldsymbol{x})^t$. □

Theorem 4. *Let $\alpha(\boldsymbol{x})\boldsymbol{x}^tA\boldsymbol{x}$, where A is independent of $\boldsymbol{x}$, then $\frac{\partial \alpha}{\partial \boldsymbol{x}} = \boldsymbol{x}^t(A + A^t)$.*

Proof. We have $\alpha(\boldsymbol{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i\alpha_{ij}x_j$, so $\frac{\partial \alpha}{\partial x_k} = \sum_{i=1}^{n} x_i\alpha_{ik} + \sum_{j=1}^{n} x_j\alpha_{kj} = \left(\boldsymbol{x}^t(A + A^t)\right)_k$. □

Directly from Theorem 4 we have,

Theorem 5. *If A is symmetic, i.e., $A^t = A$, then $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^tA\boldsymbol{x} = 2\boldsymbol{x}^tA$.*

Theorem 6. *Let $\alpha(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z})^t\boldsymbol{x}(\boldsymbol{z})$, then $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^t\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} + \boldsymbol{y}^t\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$.*

Proof. We have $\alpha(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \sum_{i=1}^{n} y_i(\boldsymbol{z})x_i(\boldsymbol{z})$, so $\frac{\partial \alpha}{\partial z_j} = \sum_{i=1}^{n}\frac{\partial y_j}{\partial z_j}x_i + \sum_{i=1}^{n} y_i\frac{\partial x_i}{\partial z_j}$. Hence $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}^t\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} + \boldsymbol{y}^t\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$. □

Directly from Theorem 6 we have,

Theorem 7. *Let $\alpha(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}(\boldsymbol{z})^t\boldsymbol{x}(\boldsymbol{z})$, then $\frac{\partial \alpha}{\partial \boldsymbol{z}} = 2\boldsymbol{x}^t\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$.*

Theorem 8. *Let $\alpha(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z})^tA\boldsymbol{x}(\boldsymbol{z})$, where A is independent of $\boldsymbol{x}$ and $\boldsymbol{y}$, then $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}(\boldsymbol{z})^tA^t\frac{\partial \boldsymbol{y}}{\boldsymbol{z}} + \boldsymbol{y}(\boldsymbol{z})^tA\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$.*

Proof. Let $\boldsymbol{w}(\boldsymbol{z}) = A^t\boldsymbol{y}(\boldsymbol{z})$, such that $\alpha(\boldsymbol{x},\boldsymbol{z},\boldsymbol{w}) = \boldsymbol{w}(\boldsymbol{z})^t\boldsymbol{x}(\boldsymbol{z})$. By Theorem 2, $\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{z}} = A^t\frac{\partial \boldsymbol{y}^t}{\partial \boldsymbol{z}}$. By Theorem 6, $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}(\boldsymbol{z})^t\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{z}} + \boldsymbol{w}(\boldsymbol{z})^t\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}} = \boldsymbol{x}(\boldsymbol{z})^tA^t\frac{\partial \boldsymbol{y}(\boldsymbol{z})}{\boldsymbol{z}} + \boldsymbol{y}(\boldsymbol{z})^tA\frac{\partial \boldsymbol{x}(\boldsymbol{z})}{\partial \boldsymbol{z}}$. □

Theorem 9. *Let $\alpha(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}(\boldsymbol{z})^tA\boldsymbol{x}$, where A is independent of $\boldsymbol{x}$, then $\frac{\partial \alpha}{\partial \boldsymbol{z}} = \boldsymbol{x}(\boldsymbol{z})(A + A^t)\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{z}}$.*

Proof. Directly from Theorem 8. □

Theorem 10. *Let $\alpha(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}(\boldsymbol{z})^t A\boldsymbol{x}$, where A is independent of $\boldsymbol{x}$. If A is symmetric, i.e., $A^t = A$, then $\frac{\partial\alpha}{\partial\boldsymbol{z}} = 2\boldsymbol{x}^t A\frac{\partial\boldsymbol{x}}{\partial\boldsymbol{z}}$.*

Proof. Directly from Theorem 9 □

Now we can follow Hasty, with a lot of extra calculations. Let

$$\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta}).$$

We want to minimize RSS($\boldsymbol{\beta}$). To do so we differentiate with respect to $\boldsymbol{\beta}$, using matrix differentiation.

$$\frac{\partial\text{RSS}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = (\boldsymbol{y} - X\boldsymbol{\beta})^t(-X) + (\boldsymbol{y} - X\boldsymbol{\beta})^t(-X) = -2(\boldsymbol{y} - X\boldsymbol{\beta})^t X.$$

Solving the equation to zero gives estimate $\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$.

In the case of $n = 1$, we have

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \dots & \dots \\ 1 & X_m \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_m \end{pmatrix}, \quad X^t\boldsymbol{y} = \begin{pmatrix} \sum Y_i \\ \sum X_iY_i \end{pmatrix}, \quad X^tX = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix},$$

so that

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X^tX)^{-1}X^t\boldsymbol{y} \\ &= \frac{1}{n\sum X_i^2 - \left(\sum X_i\right)^2}\begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}\left(\sum Y_i \sum X_iY_i\right) \\ &= \frac{1}{n\sum X_i^2 - \left(\sum X_i\right)^2}\begin{pmatrix} \left(\sum X_i^2\right)\left(\sum Y_i\right) - \left(\sum X_i\right)\left(\sum X_iY_i\right) \\ n\left(\sum X_iY_i\right) - \left(\sum X_i\right)\left(\sum Y_i\right) \end{pmatrix}.\end{aligned}$$

In particular,

$$\begin{aligned}\hat{\beta}_1 &= \frac{n\left(\sum X_iY_i\right) - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum X_iY_i - \frac{1}{n}\left(\sum X_i\right)\left(\sum Y_i\right)}{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)\left(\sum X_i\right)} = \frac{\sum(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum(X_i - \overline{X}_n)^2}, \\ \hat{\beta}_0 &= \frac{\left(\sum X_i^2\right)\left(\sum Y_i\right) - \left(\sum X_i\right)\left(\sum X_iY_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \frac{\left(\sum X_i^2\right)\left(\sum Y_i\right) - \left(\sum X_i\right)\left(\sum X_iY_i\right) + \frac{1}{n}(\sum X_i)(n\sum X_iY_i - (\sum X_i)(\sum Y_i))}{n\sum X_i^2 - \left(\sum X_i\right)^2} - \hat{\beta}_1\overline{X}_n \\ &= \frac{\left(\sum X_i^2\right)\left(\sum Y_i\right) - \left(\sum X_i\right)^2\frac{1}{n}\sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} - \hat{\beta}_1\overline{X}_n \\ &= \overline{Y}_n - \hat{\beta}_1\overline{X}_n.\end{aligned}$$

The residual is defined by $\boldsymbol{\epsilon} = Y - X\boldsymbol{\beta}$ and $\hat{\boldsymbol{\epsilon}} = Y - X\hat{\boldsymbol{\beta}}$. Note that

$$(I - X(X^tX)^{-1}X^t)X\boldsymbol{\beta} = X\boldsymbol{\beta} - X\boldsymbol{\beta} = 0, \quad (X\boldsymbol{\beta})^t(I - X(X^tX)^{-1}X^t) = \boldsymbol{\beta}^tX^t - \boldsymbol{\beta}^tX^t = 0.$$

Therefore, to find the relationship of $\hat{\boldsymbol{\epsilon}}$ with $\boldsymbol{\epsilon}$, note that

$$E(\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}|X^n) = \frac{1}{n}\sum_{i=1}^n E(\epsilon_i^2|X_i) = \frac{1}{n}\sum_{i=1}^n(V(\epsilon_i|X_i) - E(\epsilon_i|X_i)^2) = \sigma^2,$$

and

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}} &= (Y - X\boldsymbol{\beta})^t(Y - X\boldsymbol{\beta}) \\
&= Y^t(I - X(X^tX)^{-1}X^t)^t(I - X(X^tX)^{-1}X^t)Y \\
&= Y^t(I - X(X^tX)^{-1}X^t)^2Y \\
&= Y^t(I - 2X(X^tX)^{-1}X^t + (X(X^tX)^{-1}X^t)^2)Y \\
&= Y^t(I - X(X^tX)^{-1}X^t)Y \\
&= (X\boldsymbol{\beta} + \boldsymbol{\epsilon})^t(I - X(X^tX)^{-1}X^t)(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\epsilon}^t(I - X(X^tX)^{-1}X^t)\boldsymbol{\epsilon}.
\end{aligned}$$

We apply a trick, used somewhere further up in the book as well, to calculate $\hat{\boldsymbol{\epsilon}}^t\boldsymbol{\epsilon}$. Note that $\hat{\boldsymbol{\epsilon}}^t\boldsymbol{\epsilon} \in \mathbb{R}$, so $\operatorname{tr}(\hat{\boldsymbol{\epsilon}}^t\boldsymbol{\epsilon}) = \hat{\boldsymbol{\epsilon}}^t\boldsymbol{\epsilon}$ and $E(\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}}|X^n) = \hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}}$. So,

$$\begin{aligned}
E(\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}}|X^n) &= E(\operatorname{tr}(\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}})|X^n) \\
&= E(\operatorname{tr}(\boldsymbol{\epsilon}^t(I - X(X^tX)^{-1}X^t)\boldsymbol{\epsilon})|X^n) \\
&= E(\operatorname{tr}(\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}(I - X(X^tX)^{-1}X^t)|X^n) \\
&= \sigma^2E(\operatorname{tr}(I - X(X^tX)^{-1}X^t)|X^n) \\
&= \sigma^2(\operatorname{tr}(I_n) - \operatorname{tr}(X(X^tX)^{-1}X^t)) \\
&= \sigma^2(\operatorname{tr}(I_n) - \operatorname{tr}(I_k)) \\
&= \sigma^2(n - k),
\end{aligned}$$

and therefore an unbias estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-k}\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}} = \frac{1}{n-k}\sum_{i=1}^{k}\hat{\epsilon}_i^2.$$

In particular for this exercise, $k = 2$, which gives the final result.

Solution 13.2

Note that $E(\hat{\boldsymbol{\beta}}|X) = \boldsymbol{\beta}$, so from the definition of the variance and $X\boldsymbol{\beta} + \boldsymbol{\epsilon} = Y$,

$$\begin{aligned}
V(\hat{\boldsymbol{\beta}}|X) &= E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t|X^n) \\
&= E(((X^tX)^{-1}X^tY - \boldsymbol{\beta})((X^tX)^{-1}X^tY - \boldsymbol{\beta})^t|X^n) \\
&= E(((X^tX)^{-1}X^t(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta})((X^tX)^{-1}X^t(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta})^t|X^n) \\
&= E((X^tX)^{-1}X^t\boldsymbol{\epsilon}\boldsymbol{\epsilon}^tX(X^tX)^{-1}|X^n) \\
&= (X^tX)^{-1}X^tE(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t|X^n)X(X^tX)^{-1} \\
&= \sigma^2(X^tX)^{-1} \\
&= \frac{\sigma^2}{nS_X^2}\begin{pmatrix} \frac{1}{n}\sum X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}.
\end{aligned}$$

Solution 13.3

With all the previous work this exercise is almost trivial. We have

$$\hat{\beta} = (X^tX)^{-1}X^tY = \frac{\sum_{i=1}^n X_iY_i}{\sum_{i=1}^n X_i^2}.$$

The standard error is given by

$$\mathrm{se}(\hat{\beta}) = \sqrt{V(\hat{\beta}|X^n)} = \sqrt{\sigma^2(X^tX)^{-1}} = \frac{\sigma}{||X||_2}.$$

Solution 13.4

The bias is defined by $\mathrm{bias}(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$, so by definition

$$\mathrm{bias}(\hat{R}_{\mathrm{tr}}(S)) = E_{R(S)}(\hat{R}_{\mathrm{tr}}(S)) - R(S).$$

Firstly, as $\epsilon^* \perp \epsilon_{\mathrm{tr}}$, we have $\hat{Y}_i(S) = X_i\beta(S) + \epsilon_{\mathrm{tr}} \perp X_i\beta + \epsilon^* = Y_i^*$, hence $E(\hat{Y}_i(S)Y_i^*) = E(\hat{Y}_i(S))E(Y_i^*)$. Secondly, $E(Y_i^*) = E(X_i\beta + \epsilon^*) = X_i\beta + E(\epsilon^*) = X_i\beta + E(\epsilon) = E(X_i\beta + \epsilon) = E(Y_i)$, and similar $E((Y_i^*)^2) = E(Y_i^2)$. Therefore,

$$\begin{aligned}
E_{R(S)}(\hat{R}_{\mathrm{tr}}(S)) - R(S) &= \sum_{i=1}^n \Big(E((\hat{Y}_i(S) - Y_i)^2) - E((\hat{Y}_i(S) - Y_i^*)^2)\Big) \\
&= \sum_{i=1}^n \Big(E(\hat{Y}_i(S)^2) - 2E(\hat{Y}_i(S)Y_i) + E(Y_i^2) - E(\hat{Y}_i(S)^2) + 2E(\hat{Y}_i(S)Y_i^*) - E((Y_i^*)^2)\Big) \\
&= \sum_{i=1}^n \Big(-2E(\hat{Y}_i(S)Y_i) + 2E(\hat{Y}_i(S))E(Y_i)\Big) \\
&= -2\sum_{i=1}^n \Big(E(\hat{Y}_i(X)Y_i) - E(\hat{Y}_i(S))E(Y_i)\Big) \\
&= -2\sum_{i=1}^n \mathrm{Cov}(\hat{Y}_i(S), Y_i).
\end{aligned}$$

Solution 13.5

Let generalize the exercise a little bit. Let $a \neq 0$, we test $H_0 : \beta_1 = a\beta_0$ against $H_1 : \beta_1 \neq a\beta_0$. Take $\theta = \beta_1 - a\beta_0$, then $\hat{\theta} = \hat{\beta}_1 - a\hat{\beta}_0$, and

$$V(\hat{\theta}) = V(\hat{\beta}) + a^2V(\hat{\beta}_0) = \frac{\sigma^2}{nS_X^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 + a^2\right).$$

So $\hat{\mathrm{se}}(\hat{\theta}) = \sqrt{V(\hat{\theta})}$. For the Wald test we reject H_0 when $|W| > z_{\frac{\alpha}{2}}$, where

$$W = \frac{\hat{\beta}_1 - a\hat{\beta}_0}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 + a^2}}\frac{\sqrt{n}S_X}{\hat{\sigma}}.$$

Solution 13.6

See code.

Solution 13.7

See code. Note that for BIC,

$$\mathrm{BIC}(S) = \ell_n(S) - \frac{1}{2}|S|\log(n) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n(\hat{Y}_i(S) - Y_i)^2 - \frac{1}{2}|S|\log(n).$$

So minimizing BIC(S) is equivalent to maximizing

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i(S)-Y_i)^2+|S|\log(n).$$

Solution 13.8

As σ is known, $\hat{\sigma}=\sigma$. Mallow's C_p statistic and AIC are connected through,

$$\begin{aligned}\hat{R}(S)&=\hat{R}_{\text{tr}}(S)+2|S|\sigma^2\\&=\sum_{i=1}^{n}(\hat{Y}_i(S)-Y_i)^2+2|S|\sigma^2\\&=-2\sigma^2\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i(S)-Y_i)^2-|S|\right)\\&=-2\sigma^2\left(\ell_n(S)+\frac{n}{2}\log(2\pi\sigma^2)-|S|\right)\\&\propto -\text{AIC}(S).\end{aligned}$$

So maximizing $\hat{R}(S)$ is equivalent to minimizing AIC(S).

Solution 13.9

Let $X_1,X_2,...,X_n$ i.i.d. random variables. Consider two models: $\mathcal{M}_0$ assumes $X_i\sim\text{Normal}(0,1)$, $\mathcal{M}_1$ assumes $X_i\sim\text{Normal}(\theta,1)$ where $\theta\neq 0$. We have AIC$(\mathcal{M}_0)=\ell_n(0)$ and AIC$(\mathcal{M}_1)=\ell_n(\hat{\theta})-1$, as $\mathcal{M}_1$ has one extra parameter (i.e., θ). Define

$$J_n=\begin{cases}0 & \text{if}\quad \text{AIC}(\mathcal{M}_0)>\text{AIC}(\mathcal{M}_1),\\ 1 & \text{if}\quad \text{AIC}(\mathcal{M}_0)\leq\text{AIC}(\mathcal{M}_0).\end{cases}$$

(a) We calculate

$$\text{AIC}(\mathcal{M}_0)-\text{AIC}(\mathcal{M}_1)=-\frac{1}{2}\sum_{i=1}^{n}X_i^2+\frac{1}{2}\sum_{i=1}^{n}(X_i-\hat{\theta})^2+1=\frac{n}{2}\hat{\theta}^2-\hat{\theta}\sum_{i=1}^{n}X_i+1=-\frac{n}{2}\hat{\theta}^2+1,$$

where we've used that $\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}X_i$. So for AIC$(\mathcal{M}_0)>$ AIC$(\mathcal{M}_1)$, we need $\hat{\theta}^2<\frac{2}{n}$. We know that $\hat{\theta}\sim\text{Normal}(\theta,\frac{1}{n})$, and therefore,

$$\begin{aligned}P(\text{AIC}(\mathcal{M}_0)>\text{AIC}(\mathcal{M}_1))&=P\left(-\sqrt{\frac{2}{n}}<\hat{\theta}<\sqrt{\frac{2}{n}}\right)\\&=P\left(-\sqrt{\frac{2}{n}}<\frac{Z+\theta}{\sqrt{n}}<\sqrt{\frac{2}{n}}\right)\\&=\Phi(\sqrt{2}-\sqrt{n}\theta)-\Phi(-\sqrt{2}-\sqrt{n}\theta).\end{aligned}$$

Under the assumption of $\mathcal{M}_0$ we have $\theta=0$ and

$$P(\text{AIC}(\mathcal{M}_0)>\text{AIC}(\mathcal{M}_1))=\Phi(\sqrt{2})-\Phi(-\sqrt{2})\approx 0.84.$$

Under the assumption of $\mathcal{M}_1$ we have $\theta=0$ and

$$P(\text{AIC}(\mathcal{M}_0)>\text{AIC}(\mathcal{M}_1))=\Phi(\sqrt{2}-\sqrt{n}\theta)-\Phi(-\sqrt{2}-\sqrt{n}\theta)<\Phi(\sqrt{2}-\sqrt{n}\theta)\to 0,$$

as $n\to\infty$.

(b) Define

$$\hat{f}_n(x) = \begin{cases} \phi_0(x) & \text{if} \quad J_n = 0, \\ \phi_{\hat{\theta}}(x) & \text{if} \quad J_n = 1. \end{cases}$$

Let $D(f, g)$ be the Kullback-Leibner distance. We have

$$D(\phi_\theta, \hat{f}_n) = \int \phi_\theta(x) \log\left(\frac{\phi_\theta(x)}{\hat{f}_n(x)}\right) dx = \int \phi_\theta(x) \left(\log(\phi_\theta(x)) - \log(\hat{f}_n(x))\right) dx.$$

Suppose $\theta \neq 0$, then from (a) we know $J_n \xrightarrow{P} 1$, hence $\hat{f}_n \xrightarrow{P} \phi_\theta$. But, because log is continuous almost everywhere, $\log(\hat{f}_n) \xrightarrow{P} \log(\phi_\theta)$. Hence, $D(\phi_\theta, \hat{f}_n) \xrightarrow{P} 0$ as $n \to \infty$.

(c) We have $\text{BIC}(\mathcal{M}_0) = \ell_n(0)$ and $\text{BIC}(\mathcal{M}_1) = \ell_n(\hat{\theta}) - \frac{1}{2}\log(n)$. We calculate

$$\begin{aligned} \text{BIC}(\mathcal{M}_0) - \text{BIC}(\mathcal{M}_1) &= \ell_n(0) - \ell_n(\hat{\theta}) + \frac{1}{2}\log(n) \\ &= -\frac{1}{2}\sum_{i=1}^n X_i^2 + \frac{1}{2}\sum_{i=1}^n (X_i - \hat{\theta})^2 + \frac{1}{2}\log(n) \\ &= -\frac{n}{2}\hat{\theta}^2 + \frac{1}{2}\log(n). \end{aligned}$$

So $\text{BIC}(\mathcal{M}_0) > \text{BIC}(\mathcal{M}_1)$ if and only if $\hat{\theta}^2 < \frac{1}{n}\log(n)$. Therefore,

$$\begin{aligned} P(\hat{\theta}^2 < \frac{1}{n}\log(n)) &= P\left(-\sqrt{\frac{1}{n}\log(n)} < \frac{Z}{\sqrt{n}} + \theta < \sqrt{\frac{1}{n}\log(n)}\right) \\ &= \Phi\left(\sqrt{\log(n)} - \sqrt{n}\theta\right) - \Phi\left(-\sqrt{\log(n)} - \sqrt{n}\theta\right) \to \begin{cases} 1 & \text{if} \quad \theta = 0, \\ 0 & \text{if} \quad \theta \neq 0. \end{cases} \end{aligned}$$

Solution 13.10

Let $\theta = \beta_0 + \beta_1 X_*$, and $th\hat{e}ta = \hat{\beta}_0 + \hat{\beta}_1 X_*$, such that $Y_* = \theta + \epsilon_*$ and $\hat{Y}_* = \hat{\theta}$.

(a) Let $s = \sqrt{V(\hat{Y}_*)}$, we have

$$P\left(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s\right) = P\left(-2 < \frac{Y_* - \hat{Y}}{s} < 2\right).$$

The variance is

$$V\left(\frac{Y_* - \hat{Y}}{s}\right) = \frac{1}{s^2}V\left(\hat{\theta} - \theta\right) + \frac{1}{s^2}V(\epsilon) = \frac{1}{s^2}V(\hat{\theta}) + \frac{1}{s^2}V(\epsilon) = 1 + \frac{\sigma^2}{s^2}.$$

So $(Y_* - \hat{Y})/s \sim \text{Normal}(0, 1 + \sigma^2/s^2)$, and

$$P\left(\hat{Y}_* - 2s < Y_* < \hat{Y}_* + 2s\right) = \Phi\left(2\left(1 + \frac{\sigma^2}{s^2}\right)\right) - \Phi\left(-2\left(1 + \frac{\sigma^2}{s^2}\right)\right) \neq 0.95,$$

if not $\sigma^2 \approx 0$.

(b) Introduce the correction factor

$$\hat{\xi}_n^2 = V(\hat{Y}_*) + \hat{\sigma}^2 = \left(\frac{1}{n} \frac{\sum_{i=1}^n (X_i - X_*)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + 1 \right) \hat{\sigma}^2.$$

We have

$$P\left(\hat{Y}_* - 2\hat{\xi}_n < Y_* < \hat{Y}_* + 2\hat{\xi}_n\right) = P\left(-2 < \frac{Y_* - \hat{Y}}{\hat{\xi}_n} < 2\right),$$

and

$$V\left(\frac{Y_* - \hat{Y}}{\hat{\xi}_n}\right) = \frac{1}{\hat{\xi}_n^2} V(Y_* - \hat{Y}_*) = \frac{V(\hat{Y}_*) + \sigma^2}{V(\hat{Y}_*) + \hat{\sigma}^2} \approx 1.$$

So $(Y_* - \hat{Y})/\hat{\xi}_n \sim \text{Normal}(0,1)$ (approximately), and

$$P\left(\hat{Y}_* - 2\hat{\xi}_n < Y_* < \hat{Y}_* + 2\hat{\xi}_n\right) \approx P(-2 < \text{Normal}(0,1) < 2) \approx 0.95.$$

Solution 13.11

See code.

The proposed reweighted least squares algorithm in the book contains many mistakes that make it not possible to implement.

1. The starting value $\boldsymbol{\beta}$ should not be random, as the initial value is of much importance. It's generally accepted that $\boldsymbol{\beta} = \mathbf{0}$ is a good initial value.

2. You get an floating point overflow when you calculate (13.32), instead use

$$p_i = \frac{1}{1 + \exp(-\beta_0 - \sum_{j=1}^k \beta_j x_{ij})}.$$

3. The formula for the updated $\hat{\boldsymbol{\beta}}^s$ in step 3 contains a typo an is missing a Z. The correct formula is

$$\hat{\boldsymbol{\beta}}^s = (X^t W X)^{-1} X^t W Z.$$

4. Before step 4 you should check for convergence. Calculate the log likelihood

$$\ell_n(\hat{\boldsymbol{\beta}}^s) = \sum_{i=1}^n \left(Y_i \log(p_i(\hat{\boldsymbol{\beta}}^s)) + (1 - Y_i) \log(1 - p_i(\hat{\boldsymbol{\beta}}^s)) \right).$$

If $|\ell_n(\hat{\boldsymbol{\beta}}^{s-1} - \ell_n(\hat{\boldsymbol{\beta}}^s)| < \epsilon$ (I use $\epsilon = 10^{-7}$) stop the iteration.

Chapter 14 - Multivariate Models

Solution 14.1

See solution 3.10.

Solution 14.2

We have $\log(f(X;\boldsymbol{p})) = \sum_{j=1}^k X_j \log(p_j)$, so $s(X;p_i) = X_i/p_i$, $s'(X;p_i) = -X_i/p_i^2$, and $-E(s'(X;p_i)) = n/p_i$. Therefore, the Fisher information matrix is given by $I(\theta) = \frac{1}{n} I_n(\theta) = \text{diag}(p_1^{-1}, p_2^{-1}, ..., p_k^{-1})$.

Solution 14.3

See code.

Solution 14.4

See code.

Solution 14.5

See code.

Solution 14.6

See solution 14.5.

Chapter 15 - Inference About Independence

Solution 15.1

1 → 2) If $Y \perp Z$, then $P(Y, Z) = P(Y)P(Z)$, so

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}} = \frac{P(Y=0, Z=0)P(Y=1, Z=1)}{P(Y=0, Z=1)P(Y=1, Z=0)} = 1.$$

2 ↔ 3) $\gamma = \log(\psi) = \log(1) = 0$.

2 → 4) If $\psi = 1$, then $p_{01}p_{10} = p_{00}p_{11}$. We have two cases. Suppose $i \neq j$. Without loss of generality, let $i = 0$ and $j = 1$. We have

$$\begin{aligned} p_{i\_}p_{\_j} &= (p_{i0} + p_{i1})(p_{0j} + p_{1j}) \\ &= p_{00}p_{01} + p_{00}p_{11} + p_{01}p_{01} + p_{01}p_{11} \\ &= p_{00}p_{01}\frac{p_{10}p_{01}}{p_{11}}p_{11} + p_{01}p_{01} + p_{01}p_{11} \\ &= p_{01}(p_{00} + p_{10} + p_{01} + p_{01}) \\ &= p_{01}. \end{aligned}$$

Suppose $i = j$. Without loss of generality, let $i = 0$ and $j = 0$. We have

$$\begin{aligned} p_{i\_}p_{\_j} &= (p_{i0} + p_{i1})(p_{0j} + p_{1j}) \\ &= p_{00}p_{00} + p_{00}p_{10} + p_{01}p_{00} + p_{01}p_{10} \\ &= p_{00}(p_{00} + p_{10} + p_{01} + \frac{p_{11}}{p_{10}}p_{10}) \\ &= p_{00}. \end{aligned}$$

Which shows that in all cases $p_{ij} = p_i p_j$.

1 ↔ 4) $Y \perp Z$ iff $p_{ij} = P(Y = i, Z = j) = P(Y = i)P(Z = j) = p_i p_j$.

Solution 15.2

Under the assumption of H_0, the maximum likelihood estimator for p is

$$\hat{p}_0 = \left(\frac{E_{00}}{n}, \frac{E_{01}}{n}, \frac{E_{10}}{n}, \frac{E_{11}}{n} \right).$$

Moreover,

$$\hat{p}_0 = \left(\frac{X_{0\_}X_{\_0}}{n}, \frac{X_{0\_}X_{\_1}}{n}, \frac{X_{1\_}X_{\_0}}{n}, \frac{X_{1\_}X_{\_1}}{n} \right),$$

because

$$E_{ij} = E(Y = i, Z = j) = nP(Y = i, Z = j) = nP(Y = i)P(Z = j) = \frac{X_{i\_}X_{\_j}}{n},$$

were we have used that $H_0 : Y \perp Z$. We have

$$\mathcal{L}_n(\hat{p}_0) = \prod_{i=0}^{1}\prod_{j=0}^{1} \left(\frac{X_{i\_}X_{\_j}}{n^2} \right)^{X_{ij}}, \quad \mathcal{L}_n(p_0) = \prod_{i=0}^{1}\prod_{j=0}^{1} \left(\frac{X_{ij}}{n} \right)^{X_{ij}}.$$

Using Theorem 10.22 we have that the ratio test is

$$T = 2\log\left(\frac{\mathcal{L}_n(\hat{p}_0)}{\mathcal{L}_n(p_0)} \right) = 2\sum_{i=0}^{1}\sum_{j=0}^{1} X_{ij} \log\left(\frac{X_{ij}n}{X_{i\_}X_{\_j}} \right),$$

and under $H_0 : Y \perp Z$, $T \xrightarrow{D} \chi_1^2$. So we reject H_0 when $T > \chi_1^2(\alpha)$.

Solution 15.3

We have $\gamma = \log(\psi) = \log(p_{00}) + \log(p_{11}) - \log(p_{01}) - \log(p_{10})$. Such that, for $i \neq j$,

$$\frac{\partial \gamma}{\partial p_{ii}} = \frac{1}{p_{ii}}, \quad \frac{\partial \gamma}{\partial p_{ij}} = -\frac{1}{p_{ij}},$$

and the gradient is

$$\nabla \gamma = \left(\frac{1}{p_{00}}, \frac{1}{p_{01}}, \frac{1}{p_{10}}, \frac{1}{p_{11}} \right)^t.$$

Fisher's information matrix can be calculated as follows,

$$\mathcal{L}_n(p) = \prod_{i=0}^{1}\prod_{j=0}^{1} p_{ij}^{X_{ij}},$$

such that

$$\ell_n(p) = \sum_{i=0}^{1}\sum_{j=0}^{1} X_{ij}\log(p_{ij}).$$

We have

$$\frac{\partial \ell_n(p)}{\partial p_{ij}} = \frac{X_{ij}}{p_{ij}}, \quad \frac{\partial^2 \ell_n(p)}{\partial p_{ij}^2} = -\frac{X_{ij}}{p_{ij}^2}.$$

Note that $E(X_{ij}) = np_{ij}$, which gives information matrix

$$I(p) = \frac{1}{n} I_n(p) = \operatorname{diag}\left(\frac{1}{p_{00}}, \frac{1}{p_{01}}, \frac{1}{p_{10}}, \frac{1}{p_{11}} \right),$$

and $J(p) = I^{-1}(p) = \text{diag}(p_{00}, p_{01}, p_{10}, p_{11})$. Therefore, we have

$$\text{se}(\gamma)^2 = \nabla g^t J_n(p) \nabla g = \frac{p_{00}}{np_{00}^2} + \frac{np_{01}}{p_{01}^2} + \frac{p_{10}}{np_{10}^2} + \frac{p_{11}}{np_{11}^2} = \frac{1}{np_{00}} + \frac{1}{np_{01}} + \frac{1}{np_{10}} + \frac{1}{np_{11}}.$$

Finally, putting everything together, we get the results

$$\hat{\text{se}}(\hat{\gamma}) = \frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}},$$

and

$$\hat{\text{se}}(\hat{\psi}) = \hat{\text{se}}(e^{\hat{\gamma}}) = |e^{\hat{\gamma}}|\, \hat{\text{se}}(\hat{\gamma}) = \hat{\psi}\, \hat{\text{se}}(\hat{\gamma}).$$

Solution 15.4

We use the likelihood ratio test

$$\begin{aligned} T = 2 \cdot 14 \log\left(1311 \cdot \frac{14}{655 \cdot 76}\right) + 2 \cdot 641 \log\left(1311 \cdot \frac{641}{655 \cdot 1235}\right) \\ + 2 \cdot 62 \log\left(1311 \cdot \frac{62}{656 \cdot 76}\right) + 2 \cdot 594 \log\left(1311 \cdot \frac{594}{656 \cdot 1235}\right) \\ \approx 34.53. \end{aligned}$$

Under H_0: $T \xrightarrow{D} \chi_1^2$, so the probability that this can happen is $P(T > 34.53) = 4.2.10^{-9}$. So we reject the null hypothesis H_0 that Y (color of victim) and Z (death sentence) are independent. However, note that correlation ($Y \propto Z$) doesn't imply causality.

Solution 15.5

See code.

Solution 15.6

See code.

Solution 15.7

See code.

Chapter 16 - Causal Inference

I didn't really understand this chapter. I'm not 100% confident the solutions are correct.

Solution 16.1

Create the following population.

X	Y	C_0	C_1
0	1	1	0
0	0	0	0
1	1	1	1

With this population we have $\theta = E(C_1) - E(C_0) = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3} < 0$ and $\alpha = E(Y|X = 1) - E(Y|X = 0) = 1 - \frac{1}{2} = \frac{1}{2} > 0$.

Solution 16.2

Let X be randomly assigned. We have

$$r(x) = E(Y|X=x) = \int_X C_z(x) f(z|x) dz = \int_X C_z(x) f(z) dz = E(C(X)).$$

As a counter example, let $C(x) = 0$ if $x < 0$ and $C(x) = 1$ otherwise on $x \in [-1, 1]$. Assign X such that $P(X < 0) = 1$ and $P(X \geq 0) = 0$. Then $E(C(X)) = \frac{1}{2}$, but $r(x) = E(Y|X = x) = 0$, so $\theta(x) \neq r(x)$.

Solution 16.3

Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be binary measurements from an observational study. Following the hint,

$$\begin{aligned}
\theta &= E(C_1) - E(C_0) \\
&= E(C_1|X=1)P(X=1) + E(C_1|X=0)P(X=0) - (E(C_0|X=1)P(X=1) + E(C_0|X=0)P(X=0)) \\
&= (E(C_1|X_1)P(X=1) - E(C_0|X=1)P(X=1)) + (E(C_1|X=0)P(X=0) - E(C_0|X=0)P(X=0)).
\end{aligned}$$

Note that

$$-P(X=0) \leq (E(C_1|X=0)P(X=0) - E(C_0|X=0)P(X=0)) \leq P(X=1).$$

Hence we can estimate $L \leq \theta \leq U$, where

$$\begin{aligned}
L &= \frac{1}{n_1} \sum X_i(Y_i - 1) - \frac{1}{n_0} \sum (X_i - 1) Y_i - P(X=0), \\
U &= \frac{1}{n_1} \sum X_i(Y_i - 1) - \frac{1}{n_0} \sum (X_i - 1) Y_i + P(X=1),
\end{aligned}$$

where n_0 is the number of samples with $X_i = 0$ and $n_1 = n - n_0$. Furthermore, note that $U - L = P(X = 0) + P(X = 1) = 1$. So this estimation is bounded by 1.

Solution 16.4

I don't get this exercise. I think the idea behind the exercise is to create something similar as Firgure 16.2, but for more variables. But I fell like I'm missing some of the tools or definitions to do so.

Solution 16.5

Again, I don't understand well what is being asked. This is my take: Note that $m_0 = \lim_{y\to\infty} F^{-1}(\frac{1}{2}, y)$ and $m_1 = \lim_{x\to\infty} F^{-1}(x, \frac{1}{2})$. So, $\theta = m_1 - m_0 = \lim_{x\to\infty} \lim_{y\to\infty} \left(F^{-1}(x, \frac{1}{2}) - F^{-1}(\frac{1}{2}, y)\right)$.

Chapter 17 - Directed Graphs and Conditional Independence

Solution 17.1

If $f(x, y|z) = f(x|z)f(y|z)$, then $f(x|y, z) = \frac{f(x,y|z)}{f(y|z)} = f(x|z)$. In the opposite direction, if $f(x|y, z) = f(x|z)$, then $f(x, y|z) = f(x|y, z)f(y|z) = f(x|z)f(y|z)$.

Solution 17.2

(a) Let $X \perp Y|Z$, then $f_{Y,X|Z}(y, x|z) = f_{X,Y|Z}(x, y|z) = f(x|z)f(y|z) = f(y|z)f(x|z)$, so $Y \perp X|Z$.

(b) Let $X \perp Y | Z$, $h : \mathbb{R} \to \mathbb{R}$, and $U = h(X)$. Define $A_{h,x} = \{x' | h(x') \leq x\}$. We have

$$f(h(x)|y,z) = \left(\int_{A_{h,x}} f(x|y,z)dx \right)' = \left(\int_{A_{h,x}} f(x|z)dx \right)' = f(h(x)|z).$$

So, $U \perp Y | Z$.

(c) Let $X \perp Y | Z$, $U = h(X)$. Note from (b) that $Y \perp Y | Z$, so $f(h(x)|y,z) = f(h(x)|z)$. We have

$$f(x|y,z,h(x)) = \frac{f(x,h(x)|y,z)}{f(h(x)|y,z)} = \frac{f(x,h(x)|z)}{f(h(x)|z)} = f(x|z,h(x)).$$

So, $X \perp Y | (Z,U)$.

(d) Let $X \perp Y | Z$ and $X \perp W | (Y,Z)$. We have

$$f(w,y|x,z) = f(w|x,y,z)f(y|x,z) = f(w|y,z)f(y|z) = f(w,y|z).$$

(e) Let $X \perp Y | Z$ and $X \perp Z | Y$. Note that $f(x|y,z) = f(x|z)$ and $f(x|y,z) = f(x|y)$, so $f(x|z) = f(x|y)$. We have

$$f(x) = \int f(x|z)f(z)dz = \int f(x|y)f(z)dz = f(x|y) \int f(z)dz = f(x|y) = f(x|y,z).$$

So $X \perp (Y,Z)$.

Solution 17.3

(a) We have

X	Y	$P(X,Y\|Z=0)$	$P(X,Y\|Z=1)$
0	0	0.81	0.25
0	1	0.09	0.25
1	0	0.09	0.25
1	1	0.01	0.25

(b) We just have to check that $P(X,Y|Z) = P(X|Z)P(Y|Z)$, so

X	Y	Z	$P(X\|Z)$	$P(Y\|Z)$	$P(X\|Z)P(Y\|Z)$	$P(X,Y\|Z)$
0	0	0	0.9	0.9	0.81	0.81
1	0	0	0.1	0.9	0.09	0.09
0	1	0	0.9	0.1	0.09	0.09
1	1	0	0.1	0.1	0.01	0.01
0	0	1	0.5	0.5	0.25	0.25
1	0	1	0.5	0.5	0.25	0.25
0	1	1	0.5	0.5	0.25	0.25
1	1	1	0.5	0.5	0.25	0.25

So $P(X,Y|Z) = P(X|Z)P(Y|Z)$.

(c) The marginal distributions of X and Y are $P(X=0) = 0.7$, $P(X=1) = 0.3$), and $P(Y=0) = 0.7$ and $P(Y=1) = 0.3$.

(d) Take $X = 0$ and $Y = 0$, then $P(X=0)P(Y=0) = 0.49$, but $P(X=0, Y=0) = 0.53$). So $P(X,Y) \neq P(X)P(Y)$.

Solution 17.4

Apply Theorem 17.6.

Solution 17.5

Using Theorem 17.6, $X \perp \tilde{X} | \pi_X = X \perp Z$.

To show that X and Z are dependent of Y, we only have to find one example of f satisfying $X \to Y \leftarrow Z$ such that not $X \perp Z | Y$. Let $P(X=0) = \frac{1}{2} = P(X=1)$ and $P(Z=0) = \frac{1}{2} = P(Z=1)$. Define Y by $Y=1$ if $X=1$ and $Z=1$, otherwise $Y=0$. Note that for this case we have $f(x,y,z) = f(x)f(y|x,z)f(z)$, because

X	Y	Z	$P(X)$	$P(Y)$	$P(Z)$	$P(Y\|X,Z)$	$P(X)P(Y\|X,Z)P(Z)$	$P(X,Y,Z)$
0	0	0	0.5	0.75	0.5	1	0.25	0.25
1	0	0	0.5	0.75	0.5	1	0.25	0.25
0	1	0	0.5	0.25	0.5	0	0	0
1	1	0	0.5	0.25	0.5	0	0	0
0	0	1	0.5	0.75	0.5	1	0.25	0.25
1	0	1	0.5	0.75	0.5	0	0	0
0	1	1	0.5	0.25	0.5	0	0	0
1	1	1	0.5	0.25	0.5	1	0.25	0.25

Note that $f_{X,Z|Y}(1,1|1) = 1$, but $f_{X|Y}(1|1) = \frac{1}{2} = f_{Z|Y}(1|1)$, so $f_{X,Z|Y}(1,1|1) \neq f_{X|Y}(1|1) f_{Z|Y}(1|1)$.

Solution 17.6

See code.

Solution 17.7

(a)

$$f(x, \boldsymbol{y}, \boldsymbol{z}) = f(x) \left(\prod_{i=1}^{4} f(y_i | x, z_i) f(z_i) \right).$$

(b) We have

$$\begin{aligned} f(x, z_i) &= \int_{\boldsymbol{y}} \int_{\boldsymbol{z}_{(j)}} f(x, \boldsymbol{y}, \boldsymbol{z}) d\boldsymbol{y} d\boldsymbol{z}_{(j)} \\ &= \int_{\boldsymbol{y}} \int_{\boldsymbol{z}_{(j)}} f(x) f(z_i) \left(\prod_{j=1}^{4} f(y_j | x, z_i) \right) \left(\prod_{j \neq i} f(z_j) \right) d\boldsymbol{y} d\boldsymbol{z}_{(j)} \\ &= f(x) f(z_i) \int_{\boldsymbol{z}_{(j)}} \int_{\boldsymbol{y}} \left(\prod_{j=1}^{4} f(y_j | x, z_i) \right) d\boldsymbol{y} \left(\prod_{j \neq i} f(z_j) \right) d\boldsymbol{z}_{(j)} \\ &= f(x) f(z_i). \end{aligned}$$

Therefore, $X \perp Z_i$, for all i.

Solution 17.8

(a) The joint distribution function is given by $f(x, y, z) = f(x)f(y|x)f(z|x, y)$. So

$$f(x,y,z) = \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{1-x} \left(\frac{e^{4x-2}}{1+e^{4x-2}}\right)^y \left(1-\frac{e^{4x-2}}{1+e^{4x-2}}\right)^{1-y} \left(\frac{e^{2x+2y-2}}{1+e^{2x+2y-2}}\right)^z \left(1-\frac{e^{2x+2y-2}}{1+e^{2x+2y-2}}\right)^{1-z}.$$

We have $f(z|y) = \sum_x f(z|x, y)f(x|y) = f(z|0, y)f(0|y) + f(z|1, y)f(1|y)$. Write out $f(Y|X)$ explicitly, we see that

$$X|Y \sim \text{bernoulli}\left(\frac{e^{2y}}{1+e^2}\right).$$

These two facts together gives a rather large expression for $f(z|y) = f(z|0, y)f(0|y) + f(z|1, y)f(1|y)$. (I could not find a simple expressions) In particular, if $Z = 1$ and $Y = 1$, we have

$$f(1|1) = f(1|0,1)f(0|1) + f(1|1,1)f(1|1) = \frac{1}{2}\left(\frac{1}{1+e^2}\right) + \left(\frac{e^2}{1+e^2}\right)^2 \approx 0.8354.$$

(b) See code.

(c) To calculate $P(Z = z|Y := y)$, we define $f^*(x, z) = f(x)f(z|x, y)$. Therefore we have the expression

$$P(Z = z|Y := y) = \sum_{x\in\{0,1\}} f^*(x,z) = f(0)f(z|0,y) + f(1)f(z|1,y).$$

In particular,

$$P(Z = 1|Y := 1) = \frac{1}{4} + \frac{1}{2}\frac{e^2}{1+e^2} \approx 0.6904.$$

(d) See code.

Solution 17.9

Let $X \sim \text{Normal}(0, 1)$, $Y|X \sim \text{Normal}(\alpha x, 1)$ and $Z|X, Y \sim \text{Normal}(\beta y+\gamma z, 1)$. As this is a tedious exercise with a lot of bookkeeping I skip most of the calculations (which I did on paper).

(a) Note that $f(x, y, z) = f(x)f(y|x)f(z|x, y)$. Expanding the equation, we get $(X, Y, Z) \sim \text{Normal}(0, \Sigma_{XYZ})$, where

$$\Sigma_{XYZ} = \begin{pmatrix} 1 & \alpha & \alpha\beta+\gamma \\ \alpha & 1+\alpha^2 & \beta+\alpha^2\beta^2+\alpha\gamma \\ \alpha\beta+\gamma & \beta+\alpha^2\beta^2+\alpha\gamma & 1+(1+\alpha^2)\beta^2+\gamma^2+2\alpha\beta\gamma \end{pmatrix}.$$

According to Theorem 2.44, $(Y, Z) \sim \text{Normal}(0, \Sigma_{YZ})$, and $Z \sim \text{Normal}(\Sigma_{zy}\Sigma_{yy}^{-1}y, \Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz})$. Therefore,

$$E(Z|Y) = \Sigma_{zy}\Sigma_{yy}^{-1}y = \frac{\beta+\alpha^2\beta^2+\alpha\gamma}{1+\alpha^2}.$$

(b) Let $Y := y$, then $f^*(x, z) = f(x)f(z|x, y)$. Expanding $f^*(x, z)$ shows that $X, Z|Y := y \sim (\mu_{XZ}, \Sigma_{XZ})$, where

$$\mu_{XZ} = \begin{pmatrix} 0 \\ \beta y \end{pmatrix}, \quad \Sigma_{XZ} = \begin{pmatrix} 1 & \gamma \\ \gamma & 1+\gamma^2 \end{pmatrix}.$$

Again, by Theorem 2.44, $Z|Y := y \sim \text{Normal}(\mu_z, \Sigma_z) = \text{Normal}(\beta y, 1 + \gamma^2)$. So $E(Z|Y := y) = \beta y$.

(c) Note that $(Y, Z) \sim \text{Normal}(0, \Sigma_{YZ})$, so the correlation coefficient is

$$\rho_{YZ} = \text{Cov}(Y,Z) = \frac{\sigma_{yz}^2}{\sigma_y\sigma_z} = \frac{\beta+\alpha^2\beta^2+\alpha\gamma}{\sqrt{(1+\alpha^2)(1+(1+\alpha^2)\beta^2+\gamma^2+2\alpha\beta\gamma}}.$$

(d) Suppose Y and Z are dependent, then $\beta \neq 0$. But we can choose α and γ such that $\beta+\alpha^2\beta^2+\alpha\gamma=0$, which makes $\rho_{YZ}=0$. The opposite is also true. Suppose Y and Z are independent, then $\beta=0$. Choose $\alpha,\gamma\neq 0$, then $\beta+\alpha^2\beta^2+\alpha\gamma\neq 0$, and $\rho_{YZ}\neq 0$.

(e) Let $Y \sim \text{Normal}(\alpha,1)$, so Y is independent of X. We have pdf $f(x,y,z)=f(x)f(y)f(z|x,y)$. Expanding $f(x,y,z)$ we find $(X,Y,Z)\sim \text{Normal}(\mu,\Sigma)$, where

$$\mu=\begin{pmatrix} 0 \\ \frac{\alpha}{\sqrt{1+\beta^2}} \\ 0 \end{pmatrix}, \quad \Sigma=\begin{pmatrix} 1 & 0 & \gamma \\ 0 & 1 & \beta \\ \gamma & \beta & 1+\beta^2+\gamma^2 \end{pmatrix}.$$

The correlation coefficient is

$$\rho_{YZ}=\frac{\text{Cov}(Y,Z)}{\sigma_y\sigma_z}=\frac{\beta}{\sqrt{1+\beta^2+\gamma^2}}.$$

So $\rho=0$ iff $\beta=0$ iff Y and Z are independent. Set $Y:=y$. With exactly the same calculation as (b) we have $Z|Y:=y\sim\text{Normal}(\beta y,1+\gamma^2)$. Now $Z|Y:=y$ is independent of Y iff $\beta=0$ iff $\rho_{YZ}=0$.

Chapter 18 - Undirected Graphs

Solution 18.1

(a) Relation $X_1 \perp X_3 | X_2$ is translated to

$$X_1 - X_2 - X_3$$

(b) Relations $X_1 \perp X_2 | X_3$ and $X_1 \perp X_3 | X_2$ translate to

$$X_1 \qquad X_2 - X_3$$

(c) Relations $X_1 \perp X_2 | X_3$, $X_1 \perp X_3 | X_2$ and $X_2 \perp X_3 | X_1$ translate to

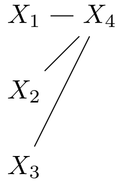
$$X_1 \qquad X_2 \qquad X_3$$

Solution 18.2

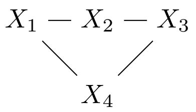
(a) Relations $X_1 \perp X_3 | X_2, X_4$, $X_1 \perp X_4 | X_2, X_3$, and $X_2 \perp X_4 | X_1, X_3$ translate to

$$X_1 - X_2 - X_3 - X_4$$

(b) Relations $X_1 \perp X_2 | X_3, X_4$, $X_1 \perp X_3 | X_2, X_4$, and $X_2 \perp X_3 | X_1, X_4$ translate to



(c) Relations $X_1 \perp X_3 | X_2, X_4$ and $X_2 \perp X_4 | X_1, X_3$ translate to



Solution 18.3

(a) $\{X_2\}$.

(b) $\{X_2, X_3\}$.

(c) $\{X_1, X_2, X_3, X_4\}$.

(d) $\{X_2, X_3, X_5\}$

Solution 18.4

Let X_1, X_2 and X_3 be discrete random variables. We want to test null-hypothesis $H_0 : X_1 \perp X_2 | X_3$ against $H_1 : X_1 \propto X_2 | X_3$. Note that if we fix $X_3 = k$, we can test $H_0^k : X_1 \perp X_2 | X_3 = k$ against $H_1^k : X_1 \propto X_2 | X_3 = k$ using the log likelihood ratio test,

$$T_k = 2 \sum_{i,j} X_{ijk} \log \left(\frac{X_{ijk} X_{\_\_k}}{X_{i\_k} X_{j\_k}} \right).$$

For each k we get a p-value $p_k = P(\chi^2_{(I-1)(J-1)+1} = T_k)$ for which we can reject H_0^k. Now use Bonferroni method to reject H_0 if and only if there is a k such that $p_k \leq \frac{\alpha}{2}$.

Solution 18.5

(a) The sum of all events is 471. Therefore, the maximum likelihood estimator is each X_{ijk} divided by 471. Results are roughly,

X_1	(M, D)	(M, S)	(B, D)	(B, S)
B	0.07	0.12	0.10	0.24
D	0.09	0.16	0.05	0.16

(b) We have $P(X_3 = D | X_1 = G, X_2 = B) = P(G, B, D)/P(G, B) \approx 0.24$ and $V(\hat{p}) = \frac{1}{n}\hat{p}(1-\hat{p}) = \frac{1}{26+76} 0.24(1-0.24) \approx 0.0018$, so $\hat{\text{se}}(\hat{p}) \approx 0.04$.

(c) We use solution 18.4 to test $H_0 : X_i \perp X_j | X_k$ against $H_1 : X_i \propto X_j | X_k$. See code. We can reject $X_1 \perp X_2 | X_3$ with a p-value of 0.0195, $X_1 \perp X_3 | X_2$ with a p-value of 0.7699, and $X_2 \perp X_3 | X_1$ with a p-value of 0.2095. Using Bonferroni correction we reject H_0 if the p-value is below $0.05/3 \approx 0.017$. We conclude that X_1, X_2 and X_3 cannot be shown to be dependent.

Chapter 19 - The log-linear model

Solution 19.1

$$\begin{aligned} p_{00} &= e^{\beta_1}, & p_{10} &= e^{\beta_1+\beta_2}, \\ p_{01} &= e^{\beta_1+\beta_3}, & p_{02} &= e^{\beta_2+\beta_4}, \\ p_{11} &= e^{\beta_1+\beta_2+\beta_3+\beta_5}, & p_{12} &= e^{\beta_1+\beta_2+\beta_4+\beta_6}. \end{aligned}$$

Solution 19.2

→) If $X_b \perp X_c | X_a$, then

$$\frac{f(x_a, x_b, x_c)}{f(x_a)} = f(x_b, x_c|x_a) = f(x_b|x_a)f(x_c|x_a) = \frac{f(x_a, x_b)}{f(x_a)}\frac{f(x_a, x_c)}{f(x_a)}.$$

Rearranging the equation we find

$$f(x_a, x_b, x_c) = f(x_a, x_b)\frac{f(x_a, x_c)}{f(x_a)} := g(x_a, x_b)h(x_a, x_c).$$

←) Note that

$$f(x_b, x_c|x_a) = \frac{f(x_a, x_b, x_c)}{f(x_a)} = \frac{g(x_a, x_b)h(x_a, x_c)}{f(x_a)} = g'(x_b)h'(x_c).$$

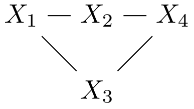
Using solution 2.12 we have $X_b \perp X_c | X_a$.

Solution 19.3

Let $\mathcal{G} = (E, V)$ be a graphical model. Take $\log(f(x)) = \sum_{A \subset S} \psi_A(x)$ with $\psi_A(x) = 0$ iff A is non-empty and for all $\{i, j\} \in A$ and $(i, j) \notin E$. Let $A \subset S$ such that $\psi_A = 0$ and $A \subset B \subset S$. Let $\{i, j\} \in A$, then $(i, j) \notin E$. Since $A \subset B$, $\{i, j\} \in B$, and hence $\psi_B = 0$. Therefore, $\mathcal{G}$ is hierarchical.

The opposite is not true as you can see in example 19.11.

Solution 19.4

(a) 

(b) $X_1 \perp X_4 | \{X_2, X_3\}$, $X_2 \perp X_3 | \{X_1, X_4\}$.

(c) The graph is not graphical, because $(1, 2) \in E$, but $\psi_{123} = 0$. The graph is hierarchical as $\emptyset$ is the only non-zero subset.

Solution 19.5

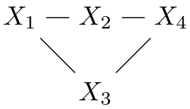
In this particular case it's easier to calculate ${}^2\log(f)$. Note that $\log(f) = \frac{{}^2\log(f)}{\log(2)}$. So we can recover $\log(f)$ from ${}^2\log(f)$. We have

$${}^2\log(f)(x_1, x_2, x_3) = 1 + 3x_1x_2 + 2x_3 + x_1x_3,$$

and

$$\log(f) = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{13}.$$

The graph is given by



In particular, $X_1 \perp X_2$ and $X_2 \perp X_3$.

Solution 19.6

I skip drawing the graphs, because you can see immediately from the formulas if the graphs are graphical or hierarchical.

(a) Graphical and hierarchical.

(b) Graphical and hierarchical.

(c) Graphical and hierarchical.

(d) Not hierarchical, hence not graphical. In particular, $\psi_{12} = 0$, but $\{1,2\} \subset \{1,2,3,4\}$ and $\psi_{1234} \neq 0$.

Chapter 20 - Non-Parametric Curve Estimation

Solution 20.1

(a) We have

$$E_{X_i}(\hat{f}(x)) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} f(y)dy = \frac{1}{h}\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} f(y)dy.$$

For calculating the variance note that $E(K(x^2)) = E(K(x))$, so

$$\begin{aligned}
V_{X_i}(\hat{f}(x)) &= V_{X_i}\left(\frac{1}{hn}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\right)\\
&= \frac{1}{h^2n^2}\sum_{i=1}^{n}V\left(K\left(\frac{x-X_i}{h}\right)\right)\\
&= \frac{1}{h^2n^2}\sum_{i=1}^{n}\left[\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} f(y)dy - \left(\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} f(y)dy\right)^2\right]\\
&= \frac{1}{h^2n}\left[\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} f(y)dy - \left(\int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} f(y)dy\right)^2\right].
\end{aligned}$$

(b) This exercise can only be true if some extra requirements are put on f, which are not. As a counter example take $f(x) = 1$ if $1 < x < 2$ or $x = 0$, and $f(x) = 0$ otherwise. Then $\hat{f}_n(0) = 0$, but $f(0) = 1$. So $P(|\hat{f}_n(0) - f(0)| < \epsilon) = 0$ for all $\epsilon < 1$, and not $\hat{f}_n(0) \xrightarrow{P} f(0)$!

Solution 20.2

See code.

Solution 20.3

See code.

Solution 20.4

This is a classical exercise.

$$\begin{aligned}
R(g,\hat{g}_n) &= E(L(g,\hat{g}_n)) \\
&= E\left(\int (g(u)-\hat{g}_n(u))^2 du\right) \\
&= \int E((g(u)-\hat{g}_n(u))^2)du \\
&= \int E(g(u)-\hat{g}_n(u))^2 + V(g(u)-\hat{g}_n(u))du \\
&= \int E(g(u)-\hat{g}_n(u))^2 du + \int V(\hat{g}_n(u))du = \int b^2(u)du + \int v(u)du.
\end{aligned}$$

Solution 20.5

As x is fixed and $x \in B_i$, $\hat{f}(x) = \frac{\hat{p}_i}{h}$. So $E(\hat{f}(x)) = E(\hat{p}_i/h) = p_i/h$, as $\hat{p}_i$ is the maximum likelihood estimator of p_i. For the variance, note that $\hat{p}_i = \nu_i/n$. We can write $\nu_i = \sum_{j=1}^n X_j$ where $X_j \sim \text{Bernoulli}(p_i)$. Hence $\nu_i \sim \text{Binomial}(n,p_i)$. So $V(\hat{f}(x)) = \frac{1}{n^2h^2}V(\nu_i) = \frac{p_i(1-p_i)}{n^2h^2}$.

Solution 20.6

We split the solution into several parts. Note that we have

$$\hat{J}(h) = \int \hat{f}^2(x)dx - \frac{2}{n}\sum_{i=1}^n \hat{f}_{(-i)}(x_i).$$

The first part can be written to

$$\begin{aligned}
\int \hat{f}^2(x)dx &= \int \sum_{i=1}^m \sum_{j=1}^m \frac{\hat{p}_i\hat{p}_j}{h^2} I(x \in B_i \cap B_j)dx \\
&= \sum_{i=1}^m \frac{\hat{p}_i^2}{h^2}\int I(x \in B_i)dx \\
&= \frac{1}{h}\sum_{i=1}^m \hat{p}_i^2.
\end{aligned}$$

This gives the first part of the solution. For the second part we first calculate

$$\begin{aligned}
\hat{f}_{(-i)}(x) &= \sum_{j=1}^m \frac{\hat{p}_{j,(-i)}}{h} I(x \in B_j) \\
&= \sum_{j=1}^m \sum_{k \neq i} \frac{1}{h}\frac{1}{n-1} I(x_k \in B_j)I(x \in B_j) = \sum_{j=1}^m \frac{\nu_j - 1}{h(n-1)} I(x \in B_j)
\end{aligned}$$

Now, we have

$$\sum_{i=1}^n \hat{f}_{(-i)}(x_i) = \sum_{i=1}^n \sum_{j=1}^m \frac{\nu_j - 1}{h(n-1)} I(x_j \in B_j) = \sum_{j=1}^m \frac{\nu_j(\nu_j-1)}{h(n-1)}.$$

Combining the two parts gives

$$\begin{aligned}
\hat{J}(h) &= \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{hn(n-1)}\sum_{j=1}^{m}\nu_j(\nu_j-1) \\
&= \frac{1}{h}\sum_{j=1}^{m}\left[\frac{\nu_j^2}{h^2} - \frac{2}{n(n-1)}\nu_j(\nu_j-1)\right] \\
&= \frac{1}{h}\sum_{j=1}^{m}\left[\frac{2n\nu_j-(n+1)\nu_j^2}{n^2(n-1)}\right] \\
&= \frac{2}{h}\sum_{j=1}^{m}\frac{\nu_j}{n(n-1)} - \frac{1}{h}\frac{n+1}{n-1}\sum_{j=1}^{m}\left(\frac{\nu_j}{n}\right)^2 \\
&= \frac{2}{h(n-1)}\sum_{j=1}^{m}\hat{p}_j - \frac{1}{h}\frac{n+1}{n-1}\sum_{j=1}^{m}\hat{p}_j^2 = \frac{2}{h(n-1)} - \frac{1}{h}\frac{n+1}{n-1}\sum_{j=1}^{m}\hat{p}_j^2
\end{aligned}$$

Solution 20.7

I'm going to skip this exercise. You can find the proof in Silverman, 1986, paragraph 3.4.3.

Solution 20.8

Let $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ be regression data with $0 \leq x_i \leq 1$. We define $\hat{r}_n(x) = \overline{Y}_j$, where $\overline{Y}_j$ is the mean of all $Y_i \in B_j$ and $x \in B_j$.

We want to apply Theorem 20.4. Let $A = \int_0^1 r(x)dx$, and define $f(x) = r(x)/A$ and $\hat{f}_n(x) = \hat{r}(x)/A$. Note that f is a probability density function, and

$$\hat{f}_n(x) = \frac{\frac{1}{k}\sum_{Y_i\in B_j} Y_j}{\frac{1}{n}\sum_i Y_i} \approx \frac{\nu_j}{nh} = \frac{\hat{p}_j}{h}.$$

Now we can apply Theorem 20.4. We have

$$R(\hat{r}_n, r) = E(L(\hat{r}_n, r)) = E(L(A\hat{f}_n, Af)) = A^2R(\hat{f}_n, f) \approx \frac{h^2A^2}{12}\int f'(u)^2du + \frac{A^2}{nh},$$

which minimizes at

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6A^2}{\int r'(u)^2du}\right)^{1/3}.$$

Solution 20.9

We need assumptions $r \in C^1$ and r' is bounded. Note that $Y_i = r(X_i) + \epsilon_i$ and $Y_{i+1} = r(X_{i+1}) + \epsilon_{i+1} \approx r(X_i) + hr'(X_i) + \epsilon_{i+1}$ where $X_{i+1} = X_i + h$ and h small. So $Y_{i+1} - Y_i \approx hr'(X_i) + \epsilon_{i+1} - \epsilon_i$. When $n \to \infty$

and $h \to 0$, we have

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{2(n-1)} \sum_{i=1}^{n} (Y_{i+1} - Y_i)^2 \\
&\approx \frac{1}{2(n-1)} \sum_{i=1}^{n} (hr'(X_i) + \epsilon_{i+1} - \epsilon_i)^2 \\
&\to \frac{1}{2(n-1)} \sum_{i=1}^{n} (\epsilon_{i+1}^2 - 2\epsilon_i\epsilon_{i+1} + \epsilon_i^2) \\
&= \frac{1}{n-1} \left(\sum_{i=2}^{n} \epsilon_i^2 + \epsilon_1 + \epsilon_{n+1} - \sum_{i=1}^{n} \epsilon_i\epsilon_{i+1} \right) \\
&\approx \frac{1}{n-1} \left(\sum_{i=2}^{n} \epsilon_i^2 - \sum_{i=1}^{n} \epsilon_i\epsilon_{i+1} \right) = \frac{\sigma^2}{n-1} \left(\sum_{i=2}^{n} Z_i^2 - \sum_{i=1}^{n} Z_i Z_{i+1} \right),
\end{aligned}
$$

where $\epsilon = \sigma Z_i$ and $Z_i \sim \text{Normal}(0,1)$. Note that $E(Z_i Z_{i+1}) = E(Z_i)E(Z_{i+1})$, and from Solution 3.12, $E(Z_i^2) = 1$. Therefore, $E(\hat{\sigma}^2) = \sigma$.

Solution 20.10

Start with the right hand side

$$
\begin{aligned}
\sum_{i=1}^{n} \frac{(Y_i - r(x_i))^2}{(1 - w_i(x_i))^2} &= \sum_{i=1}^{n} \left(\frac{Y_i - \sum_{k=1}^{n} w_k(x) Y_k}{1 - w_i(x_i)} \right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i \frac{(1 - w_i(x_i))}{(1 - w_i(x_i))} - \sum_{j=1, j\neq i}^{n} \frac{w_j(x_i)}{1 - w_i(x_i)} Y_j \right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \sum_{j=1, j\neq i}^{n} w_{(-i)j}(x_i) Y_j \right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{r}_{(-i)}(x_i) \right)^2,
\end{aligned}
$$

where we used that

$$
\frac{w_j(x_i)}{1 - w_j(x_i)} = \frac{K((x_i - x_j)/h)}{\sum_{k=1}^{n} K((x_i - x_k)/h) - K((x_i - x_j)/h)}.
$$

Chapter 21 - Smoothing Using Orthogonal Functions

Solution 21.1

From Lemma 20.1 we know that

$$
R(J) = R(f, \hat{f}_J) = \int b^2(x)dx + \int V(x)dx.
$$

As $b(x) = E(\hat{f}(x)) - f(x) = -\sum_{j=J+1}^{\infty} \beta_j \phi_j(x)$, we have

$$
\int b^2(x)dx = \int \sum_{i,j=J}^{\infty} \beta_i \beta_j \phi_i(x) \phi_j(x) dx = \sum_{j=J+1}^{\infty} \beta_j^2.
$$

For the second part

$$\begin{aligned}\int V(x)dx &= \int E((\hat{f}_J(x) - E(\hat{f}_J(x)))^2)dx \\ &= \int E(\hat{f}(x)^2) - 2E(\hat{f}(x)E(\hat{f}(x))) + E(\hat{f}(x))^2dx \\ &= \sum_{j=1}^{J} E(\hat{b}_j^2) - 2\sum_{j=1}^{J}\beta_j^2 + \sum_{j=1}^{J}\beta_j^2 \\ &= \sum_{j=1}^{J}\frac{\sigma_j^2}{n},\end{aligned}$$

as $E(\hat{\beta}_j^2) = V(\hat{\beta}_j) + E(\hat{\beta}_j)^2 = \frac{\sigma_j^2}{n} + \beta_j^2$. Combining all above yields the result.

Solution 21.2

Note that

$$E(\hat{\beta}_j^2) = V(\hat{\beta}_j) + E(\hat{\beta}_j)^2 = \frac{\sigma^2}{n} + \beta_j^2.$$

Therefore,

$$\begin{aligned}R(r,\hat{r}) &= E\left(\int (r(x) - \hat{r}(x))^2dx\right) \\ &= E\left(\int r^2(x) - 2r(x)\hat{r}(x) + \hat{r}^2(x)dx\right) \\ &= E\left(\int \sum_{i,j=1}^{\infty}\beta_i\beta_j\phi_i(x)\phi_j(x) - 2\sum_{i=1}^{\infty}\sum_{j=1}^{J}\beta_i\hat{\beta}_j\phi_i(x)\phi_j(x) + \sum_{i,j=1}^{J}\hat{\beta}_i\hat{\beta}_j\phi_i(x)\phi_j(x)dx\right) \\ &= \sum_{j=1}^{\infty}\beta_j^2 - 2\sum_{j=1}^{J}\beta_j^2 + \sum_{j=1}^{J}E(\hat{beta}^2) \\ &= \sum_{j=1}^{\infty}\beta_j^2 - 2\sum_{j=1}^{J}\beta_j^2 + \sum_{j=1}^{J}(\frac{\sigma^2}{n} + \beta_j^2) \\ &= \sum_{j=J+1}^{\infty}\beta_j^2 + J\frac{\sigma^2}{n}.\end{aligned}$$

Solution 21.3

Straightforward calculation.

Solution 21.4

Parseval's lemma. We have

$$||f||^2 = \int f^2(x)dx = \int\sum_{i,j=1}^{\infty}\beta_i\beta_j\phi_i(x)\phi_j(x)dx = \sum_{j=1}^{\infty}\beta_j^2 = ||\beta||^2.$$

Solution 21.5

See code.

Solution 21.6

A tedious exercise. I only give the solutions.

(a) $f(x) = \sqrt{2}\cos(3\pi x)$.

(b) $a_j = (\pi(1 - j^2))^{-1}$ if j is even, zero otherwise.

The rest of the exercise is in the code.

Solution 21.7

See code.

Solution 21.8

Let $D_{jk} = \{x : \psi_{jk}(x) \neq 0\}$, then $D_{jk} = \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]$. Suppose $(j_1, k_j) \neq (j_2, k_2)$, w.l.o.g., $j_1 \leq j_2$. If $j_1 = j_2$, then $D_{j_1k_1} \cap D_{j_2k_2} =$, so assume $j_1 < j_2$. Suppose $D_{j_1k_1} \cap D_{j_2k_2} \neq$. Split $D_{j_1k_1} = D^-_{j_1k_1} \cup D^+_{j_1k_1}$, where $D^-_{j_1k_1} = \left[\frac{k}{2^k}, \frac{k+1/2}{2^k}\right]$ and $D^+_{j_1k_1} = \left[\frac{k+1/2}{2^k}, \frac{k+1}{2^k}\right]$. Suppose $D_{j_2k_2} \subset D_{j_1k_1}$, and $D_{j_2k_2} \cap D^+_{j_1k_1} \neq$ and $D_{j_2k_2} \cap D^-_{j_1k_1} \neq$. Then

$$\frac{k_2}{2^{j_2}} < \frac{k_1 + \frac{1}{2}}{2^{j_1}} < \frac{k_2 + 1}{2^{j_2}}.$$

Rearanging the symbols gives $0 < 2^{j_2 - j_1}k_1 + 2^{j_2 - j_1 - 1} - 2^{j_2}k_2 < 1$. But, because $j_1 < j_2$, the value in the middle is an whole integer, which is a contradiction. So we have $D_{j_2k_2} \subset D^-_{j_1k_1}$ or $D_{j_2k_2} \subset D^+_{j_1k_1}$. In both cases we see that

$$\langle \psi_{j_1,k_1}, \psi_{j_2,k_2} \rangle = 0.$$

Lastly, note that $\langle \psi, \psi \rangle = 1$, $\langle \phi, \phi \rangle = 1$, and $\langle \psi, \phi \rangle = 0$.

Solution 21.9

See code.

Solution 21.10

(a) We have

$$E(\hat{\beta}_{jk}) = \frac{1}{n}\sum_{i=1}^{n} E(\psi_{jk}(X_i)) = \frac{1}{n}\sum_{i=1}^{n}\int_{D_{jk}} \psi_{jk}(x)f(x)dx = \frac{1}{n}\sum_{i=1}^{n}\beta_{jk} = \beta_{jk}$$

(b) Okay.

(c) Apply Theorem 21.5.

(d) See code.

Solution 21.11

I'm pretty sure this exercise is wrong, the numbers don't seem to be correct.

Solution 21.12

See code.

Chapter 22 - Classification

Solution 22.1

Suppose there is an h such that $L(h) < L(h^*)$, then

$$\int_0^1 P(h(x) = Y|X = x)P(x)dx > \int_0^1 P(h^*(x) = Y|X = x)P(x)dx.$$

With sufficient properties on the distribution there is an x such that $P(h(x) = Y|X = x) > P(h^*(x) = Y|X = x)$. But this contradicts the choice of h^*.

Solution 22.2

In general we have

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)f(y)}{P(x|0)f(0) + P(x|1)f(1)} = \frac{P(x|y)\pi_y}{P(x|0)\pi_0 + P(x|1)\pi_1},$$

where $\pi_y = P(Y = y)$. In particular

$$P(1|x) = \frac{P(x|1)\pi_1}{P(x|0)\pi_0 + P(x|1)\pi_1} = \frac{P(x|1)}{P(x|0)\frac{\pi_0}{\pi_1} + P(x|1)} > \frac{1}{2},$$

if and only if

$$\frac{\pi_1}{\pi_0} f(x|1) > f(x|0).$$

Expanding $f(x|0)$ and $f(x|1)$, and taking the log on both sides gives

$$r_1^2 < r_0^2 + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) + 2\log\left(\frac{\pi_1}{\pi_0}\right).$$

Solution 22.3

See code.

Solution 22.4

I have no idea how to solve this.

Solution 22.5

See code.

Solution 22.6

Similar to 22.4. I don't understand VC-theory.

Solution 22.7

Suppose there is a linear classifier $r(x) = \beta_1 x + \beta_0$ that perfectly classifies the data. Note that $Y_i = 1$ if $r(X_i) > \frac{1}{2}$. So we have $Y_i = 1$ if and only if $X_i > \frac{1}{2\beta_1}(1 - \beta_0)$. But in that case there will always be some X_i with $Y_i = 0$, but $X_i > \frac{1}{2\beta_1}(1 - \beta_0)$.

Note that the kernelized data $Z_i = (X_i, X_i^2)$ can be linearly seperated. Indeed, we can take $r(Z_i) = \frac{1}{2}X_i^2$.

Solution 22.8

See code.

Solution 22.9

See code.

Solution 22.10

The CDF is given by

$$F(t) = P(R < t) = 1 - P(R \geq t) = 1 - \prod_{i=1}^{n} P(X_i \geq t) = 1 - \prod_{i=1}^{n} (1 - v_d(t)) = 1 - (1 - t^2 v_d(1))^n.$$

Hence, we have

$$F^{-1}(q) = \left(\frac{1 - (1-q)^{\frac{1}{n}}}{v_d(1)} \right)^{\frac{1}{d}}.$$

The median is given by

$$F^{-1}\left(\frac{1}{2}\right) = \left(\frac{1 - (\frac{1}{2})^{\frac{1}{n}}}{v_d(1)} \right)^{\frac{1}{d}}.$$

See code for last part of the exercise. If $n = 100$, $d \geq 6$; $n = 1000$, $d \geq 8$; $n = 10000$, $d \geq 9$.

Solution 22.11

See code.

Solution 22.12

See code.

Solution 22.13

I have no idea.

Chapter 23 - Probability Redux: Stochastic Processes

Solution 23.1

$P(X_0 = 0, X_1 = 1, X_2 = 2) = 0.3 \cdot 0.2 \cdot 0.0 = 0.0$ and $P(X_0 = 0, X_1 = 1, X_2 = 1) = 0.3 \cdot 0.2 \cdot 0.1 = 0.006$.

Solution 23.2

Sequence $X_0, X_1, ...$ is a Markov chain because

$$P(X_n | X_{n-1}, ..., X_1) = P(\max(Y_n, X_{n-1}) | X_{n-1}, ..., X_1) = P(\max(Y_n, X_{n-1}) | X_{n-1}) = P(X_n | X_{n-1}).$$

The transition matrix is

$$P = \begin{pmatrix} 0.1 & 0.3 & 0.2 & 0.4 \\ 0.0 & 0.4 & 0.2 & 0.4 \\ 0.0 & 0.0 & 0.6 & 0.4 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{pmatrix}.$$

Solution 23.3

If we take $\pi = \frac{1}{a+b}(b, a)$, then $\pi P = \pi$. So the limiting distribution is

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi \\ \pi \end{pmatrix} = \frac{1}{a+b}\begin{pmatrix} b & a \\ b & a \end{pmatrix}.$$

Solution 23.4

See code.

Solution 23.5

(a) We have

$$M(n+1) = E(X_{n+1}) = E\left(\sum_{i=1}^{X_n} Y_i^{(n)}\right) = E\left(\sum_{i=1}^{X_n} Y\right) = E(X_n Y) = \mu E(X_n) = \mu M(n).$$

Variance is more tricky,

$$\begin{aligned}
V(n+1) &= V(X_{n+1}) \\
&= E(X_{n+1}^2) - E(X_{n+1})^2 \\
&= E\left(\sum_{i=1}^{X_n}\sum_{j=1}^{X_n} Y_i^{(n)} Y_j^{(n)}\right) - \mu^2 EM(n)^2 \\
&= E\left(\sum_{i=1}^{X_n} Y^2 + \sum_{i\neq j} Y_i^{(n)} Y_j^{(n)}\right) - \mu^2 EM(n)^2 \\
&= E(X_n)E(Y^2) + E(X_n(X_n-1))E(Y)^2 - \mu^2 M(n)^2 \\
&= M(n)E(Y^2) + (E(X_n^2) - E(X_n))E(Y)^2 - \mu^2 M(n)^2 \\
&= M(n)E(Y^2) + (V(X_n) + E(X_n)^2 - E(X_n))E(Y)^2 - \mu^2 M(n)^2 \\
&= M(n)(V(Y) + E(Y)^2) + (V(n) + M(n)^2 - M(n))\mu^2 \\
&= \sigma^2 M(n) + \mu^2 V(n).
\end{aligned}$$

(b) Follows from induction on n. We have $M(0) = 1$ and $V(0) = 0$. Then, $M(n) = \mu M(n-1) = \mu^n$, and $V(n) = \sigma^2 M(n-1) + \mu^2 V(n-1) = \sigma^2 \mu^{n-1}\frac{1-\mu^n}{1-\mu}$.

(c) We have 3 cases. If $\mu > 1$, then $V(n) \to \infty$ as $n \to \infty$. If $\mu = 1$, then $V(n) = n\sigma^2 \to \infty$ as $n \to \infty$. If $\mu < 1$, then $V(n) \to 0$ as $n \to \infty$.

(d) Let $N = \max n | X_n = 0$ be the extinction time. Let $F(n) = P(N \leq n)$. We introduce some notation. If $X_1 = k$, i.e., we have k arch animals, let $Z_i^{(n,k)}$ all offspring from arch animal i at time n. Note that

$$X_n = Z_1^{(n,k)} + Z_2^{(n,k)} + ... + Z_k^{(n,k)},$$

and $P(Z_i^{(n,k)} = 0) = F(n-1)$. We have

$$\begin{aligned}
F(n) &= P(N \leq n) \\
&= \sum_{k=0}^{\infty} P(X_n = 0 | X_1 = k) P(X_1 = k) \\
&= \sum_{k=0}^{\infty} P(X_1 = k) \prod_{i=1}^{k} P(Z_i^{(n,k)} = 0) = \sum_{k=0}^{\infty} p_k F(n-1)^k.
\end{aligned}$$

(e) We have $F(n) = \frac{1}{4} + \frac{1}{2}F(n-1) + \frac{1}{4}F(n-1)^2 = \frac{1}{4}(1+F(n-1))^2$. I have no idea how to find a closed expression for this recurrence relation.

Solution 23.6

Calculated with the computer, $\pi \approx (0.11, 0.90, 0.43)$.

Solution 23.7

Let $p_{ij}(m) = P(X_{m+1} = j | X_m = i)$ and $p_{ji}(n) = P(X_{n+1} = i | X_n = j)$. We have

$$\sum_n p_{jj}(n) \geq \sum_n p_{ji}(n) p_{ii}(n+1) p_{ij}(n+2) \geq p_{ji}(a) \left(\sum_n p_{ii}(n+1) \right) p_{ij}(b) \to \infty,$$

for some a, b, when $n \to \infty$. So j is recurrent.

Solution 23.8

Recurrent: 3, 5, 6. Transient: 1, 2, 4.

Solution 23.9

We have $\pi = (\frac{1}{2}, \frac{1}{2})$ as stationary distribution. The book isn't clear what they mean exactly with convergence. Probably, this exercise wants to show that this chain doesn't converge (it's flipping between position 1 and 2), but does have a stationary distribution.

Solution 23.10

Let $\pi = (a, b, c, d, e)$. Solving $\pi P = \pi$ we find $\pi = a(1, p, p^2, p^3, p^4)$. We need $\sum \pi_i = 1$, so $a = \frac{1-p}{1-p^5}$. We get

$$\pi = \left(\frac{1-p}{1-p^5}, p\frac{1-p}{1-p^5}, p^2\frac{1-p}{1-p^5}, p^3\frac{1-p}{1-p^5}, p^4\frac{1-p}{1-p^5} \right)$$

Solution 23.11

Note that in particular $X(t) \sim \text{Poisson}(\Lambda(t))$. So

$$Y(s) = X(\Lambda^{-1}(s)) \sim \text{Poisson}(\Lambda(\Lambda^{-1}(s))) = \text{Poisson}(s).$$

Solution 23.12

We have

$$\begin{aligned} P(X(t) = m | X(t+s) = n) &= \frac{P(X(t) = m, X(t+s) - X(t) = n - m)}{P(X(t+s) = n)} \\ &\propto P(X(t) = m) P(X(t+s) - X(t) = n - m) \\ &\propto \frac{n!}{m!(n-m)!} t^m s^{n-m} \frac{1}{(s+t)^n} \\ &= \binom{n}{m} \left(\frac{t}{s+t} \right)^m \left(\frac{s}{s+t} \right)^{n-m} \sim \text{Binomial}\left(n, \frac{t}{s+t} \right). \end{aligned}$$

Solution 23.13

Note that $X(t) \sim \text{Poisson}(\lambda t)$, so

$$\begin{aligned} P(X(t) = 1, 3, 5, ...) &= \sum_{x=1,3,5,...} P(X(t) = x) \\ &= \sum_{n=0}^{\infty} \frac{1}{(2n+1)!} e^{-\lambda t} (\lambda t)^{2n+1} \\ &= e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^{2n+1}}{(2n+1)!} \\ &= e^{-\lambda t} \sinh(\lambda t) = \frac{1}{2}(1 - e^{-2\lambda t}). \end{aligned}$$

Solution 23.14

We take initial condition $X(0) = 0$. The time that person P_i spends longer than s time on the server is $P(P_i > s) = 1 - G(s)$. If we want to find out how many persons are only at time t, we have to count all persons that stayed longer than $t - t_0^{(i)}$, where $t_0^{(i)}$ is the time person P_i logged in. Mathematically, with $N_i = X_i - X_{i-1}$.

$$Y(t) = \sum_{i=1}^{t} \sum_{j=1}^{N_i} I(P_{i,j} > t - i),$$

where

$$\begin{aligned} I(P_{i,j} > t - i) &\sim \text{Bernoulli}(1 - G(t - i)), \\ \sum_{j=1}^{N_i} I(P_{i,j} > t - i) &\sim \text{Binomial}(N_i, 1 - G(t - i)), \\ N_i &\sim \text{Poisson}(\lambda). \end{aligned}$$

Putting everything together, we find

$$Y(t) = \sum_{i=1}^{t} \sum_{j=1}^{N_i} I(P_{i,j} > t - i) \sim \text{Poisson}\left(\lambda \sum_{i=1}^{t} (1 - G(t - i))\right).$$

Solution 23.15

It seems like you need extra restrictions on f, but I have no idea how to solve this exercise.

Solution 23.16

For the cumulative distribution function we have

$$F(t) = 1 - P(X > t) = 1 - \frac{1}{0!}(\lambda \pi t^2)^0 e^{-\lambda \pi t^2} = 1 - e^{-\lambda \pi t^2}.$$

Therefore, the probability density function is $f(t) = F'(t) = 2\lambda \pi t e^{-\lambda \pi t^2}$. The expected value is

$$E(X) = \int_0^{\infty} t f(t) dt = \int_0^{\infty} 2\lambda \pi t e^{-\lambda \pi t^2} dt = \frac{1}{2\sqrt{\lambda}}.$$

The last integral needs some explaination. Note that

$$\int_0^\infty x^2 e^{-x^2} dx = \frac{1}{2}\int_0^\infty x \cdot 2xe^{-x^2} dx = -\frac{1}{2}xe^{-x^2}\Big|_0^\infty + \frac{1}{2}\int_0^\infty e^{-x^2} dx = \frac{1}{4}\sqrt{\pi}\mathrm{erf}(x)\Big|_0^\infty = \frac{1}{4}\sqrt{\pi}.$$

And more generally for $A \neq 0$,

$$\int_0^\infty 2Ax^2 e^{-Ax^2} dx = \frac{2}{\sqrt{A}}\int_0^\infty y^2 e^{-y^2} dy = \frac{1}{2}\sqrt{\frac{\pi}{A}}.$$

Chapter 24 - Simulation Methods

Solution 24.1

(a) See code.

(b) We have

$$\begin{aligned} E((Y_i - Y)^2) &= V(Y_i - Y) + E(Y_i - Y)^2, \\ V(Y_i - Y) &= V(Y_i) = E(Y_i^2) - E(Y_i)^2 = I - I^2, \\ E(Y_i - I) &= E(Y_i) - I = I - I = 0. \end{aligned}$$

Therefore,

$$E(\mathrm{se}^2) = \frac{1}{N}E(s^2) = \frac{1}{N(N-1)}\sum_{i=1}^N E((Y_i - Y)^2) = \frac{1}{N-1}(I - I^2).$$

(c) See code.

(d) The optimal g is

$$g^*(x) = \frac{h(x)f(x)}{\int_1^2 h(x)f(x)dx} = \begin{cases} \frac{\phi(x)}{I} & \text{if } 1 < x < 2, \\ 0 & \text{else} \end{cases}$$

We see that we need to know the exact value of I (which we try to find) for the optimal g. The variance if

$$\begin{aligned} V\left(\frac{fh}{g^*}\right) &= \int \left(\frac{fh}{g^*}\right)^2 (x)g^*(x)dx - \left(\int \frac{fh}{g^*}(x)g^*(x)dx\right)^2 \\ &= \int_1^2 \frac{f^2(x)}{g^*(x)}dx - I^2 \\ &= \int_1^2 I\phi(x)dx - I^2 = 0. \end{aligned}$$

Solution 24.2

(a) This follows from the law of large numbers.

(b) We use that $f(x,y) = f(x|y)f(y)$. With this, we have

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^N \frac{f(x,Y_i)}{f(X_i,Y_i)}w(X_i) = \frac{1}{N}\sum_{i=1}^N \frac{f(x|Y_i)f(Y_i)}{f(X_i|Y_i)f(Y_i)}w(X_i) = \frac{1}{N}\sum_{i=1}^N \frac{f(x|Y_i)}{f(X_i|Y_i)}w(X_i).$$

Solution 24.3

(a) The results follows from $f_Y(x) = g(x)\frac{f(x)}{Mg(x)} = \frac{1}{M}f(x) \propto f(x)$.

(b) See code.

Solution 24.4

See code.

Solution 24.5

See code. I think I implemented everything correctly, however, due to floating point errors the simulated distribution doesn't converge to the desired distribution. To make this exercise work, you have to figure out how to calculated $\mathcal{L}_n(\beta_{\text{new}})/\mathcal{L}_n(\beta_{\text{old}})$ numerically.