Probability and Statistics for Data Science

Solutions to Exercises

Carlos Fernandez-Granda

This document contains all solutions to the exercises in the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Contents

1

Probability

Exercises

1.1 (Conditional probability space) We check that $\mathcal{C}_A$ satisfies the conditions:

- If $B \in \mathcal{C}_A$, then $B^c \in \mathcal{C}_A$. If the sample space is A then $B^c = A - B$. If $B \in \mathcal{C}_A$, there is some set $S \in \mathcal{C}$ such that $B = A \cap S$. This implies that $S^c \in \mathcal{C}$ because $\mathcal{C}$ is a valid collection. As a result, $S^c \cap A \in \mathcal{C}_A$. We end the proof proving $A - B = S^c \cap A$ by showing that they contain each other. (1) If $\omega \in A - B$, then ω belongs to A and not to B. This means that it cannot belong to S because otherwise it would belong to $B = A \cap S$. This implies $A - B \subseteq S^c \cap A$. (2) If $\omega \in S^c \cap A$, ω belongs to A and not to S. It cannot belong to B because then it would be in S. This implies $S^c \cap A \subseteq A - B$.
- If $B_1, B_2 \in \mathcal{C}_A$, then $B_1 \cup B_2 \in \mathcal{C}_A$. If $B_1, B_2 \in \mathcal{C}_A$, then there exist $S_1, S_2 \in \mathcal{C}$ such that $B_1 = A \cap S_1$ and $B_2 = A \cap S_2$. $S_1 \cup S_2$ is in $\mathcal{C}$, so $A \cap (S_1 \cup S_2)$ is in $\mathcal{C}_A$. This completes the proof because $A \cap (S_1 \cup S_2) = (A \cap S_1) \cup (A \cap S_2) = B_1 \cup B_2$.
- If $B_1, B_2, \ldots \in \mathcal{C}_A$ then $\cup_{i=1}^{\infty} B_i \in \mathcal{C}_A$. This holds by the same argument as the finite case.
- $\mathcal{C}_A$ contains the sample space. $A = A \cap A$, so $A \in \mathcal{C}_A$.

Note that by the definition for any $B \in \mathcal{C}_A$

$$\mathrm{P}_A(B) := \frac{\mathrm{P}(B)}{\mathrm{P}(A)}. \tag{1.1}$$

We check that $\mathcal{P}_A$ satisfies the conditions of a probability measure:

- $\mathrm{P}_A(B) \geq 0$ for any event $B \in \mathcal{C}_A$. This just follows from $\mathrm{P}(B) \geq 0$, and $\mathrm{P}(A) > 0$.
- If $B_1, B_2, \ldots, B_n \in \mathcal{C}_A$ are disjoint then $\mathrm{P}_A(\cup_{i=1}^n B_i) = \sum_{i=1}^n \mathrm{P}_A(B_i)$. Since $B_1, B_2, \ldots, B_n$ are also in $\mathcal{C}$ we have

$$\mathrm{P}_A\left(\cup_{i=1}^n B_i\right) := \frac{\mathrm{P}(\cup_{i=1}^n B_i)}{\mathrm{P}(A)} \tag{1.2}$$

$$= \frac{\sum_{i=1}^n \mathrm{P}(B_i)}{\mathrm{P}(A)} \tag{1.3}$$

$$= \sum_{i=1}^n \mathrm{P}_A(B_i). \tag{1.4}$$

- For a countably infinite sequence of disjoint sets $B_1, B_2, \ldots \in \mathcal{C}_A$ $\mathrm{P}(\lim_{n\to\infty} \cup_{i=1}^n B_i) = \lim_{n\to\infty} \sum_{i=1}^n \mathrm{P}(B_i)$. This holds by the same argument as the finite case.
- The probability of the sample space equals 1. By the definition of P_A

$$\mathrm{P}_A(A) := \frac{\mathrm{P}(A)}{\mathrm{P}(A)} = 1. \tag{1.5}$$

1.2 (Empirical probability measure) We check that $\mathcal{P}$ satisfies the conditions of a probability measure:

- $\mathrm{P}(B) \geq 0$ for any event $B \in \mathcal{C}$. The numerator and denominator are both nonnegative by definition.
- If $S_1, S_2, \ldots, S_n \in \mathcal{C}$ are disjoint then $\mathrm{P}\left(\cup_{i=1}^n S_i\right) = \sum_{i=1}^n \mathrm{P}(S_i)$.

$$\begin{aligned} \mathrm{P}\left(\cup_{i=1}^n S_i\right) &= \frac{\text{number of data points with value in } \cup_{i=1}^n S_i}{N} \qquad (1.6) \\ &= \frac{\text{number in } S_1 + \text{number in } S_2 + \ldots + \text{number in } S_n}{N} \\ &= \sum_{i=1}^n \mathrm{P}(S_i). \qquad (1.7) \end{aligned}$$

- For a countably infinite sequence of disjoint sets $S_1, S_2, \ldots \in \mathcal{C}$ $\mathrm{P}\left(\lim_{n\to\infty} \cup_{i=1}^n S_i\right) = \lim_{n\to\infty} \sum_{i=1}^n \mathrm{P}(S_i)$. This follows by the same argument as in the finite case.
- The probability of the sample space equals 1.

$$\mathrm{P}(\Omega) := \frac{\text{number of data points with value in } \Omega}{N} = 1. \qquad (1.8)$$

1.3 (Independence and complements) Yes. If A and B are independent then

$$\mathrm{P}(B) = \mathrm{P}(A^c \cap B) + \mathrm{P}(A \cap B) = \mathrm{P}(A^c \cap B) + \mathrm{P}(A)\mathrm{P}(B), \qquad (1.9)$$

which implies

$$\mathrm{P}(A^c \cap B) = \mathrm{P}(B)(1 - \mathrm{P}(A)) = \mathrm{P}(A^c)\mathrm{P}(B). \qquad (1.10)$$

1.4 (Conditional independence and complements) Not necessarily. For example, imagine that we flip two independent fair coins. A is the event that heads occurs on the first coin flip, C is the event that heads occurs on the second coin flip, and $B = C \cup A^c$ is the event that heads occurs in the second coin flip, but not in the first. Then $\mathrm{P}(C) = 0.5$, $\mathrm{P}(A \,|\, C) = 0.5$, $\mathrm{P}(A \,|\, C^c) = 0.5$, $\mathrm{P}(B \,|\, A, C) = \mathrm{P}(B \,|\, A^c, C) = 1$, $\mathrm{P}(B \,|\, A, C^c) = 0$, $\mathrm{P}(B \,|\, A^c, C^c) = 1$. We have

$$\begin{aligned} \mathrm{P}(A, B \,|\, C) &= \mathrm{P}(A \,|\, C)\mathrm{P}(B \,|\, A, C) = 0.5, \qquad (1.11) \\ \mathrm{P}(B \,|\, C) &= \mathrm{P}(A \,|\, C)\mathrm{P}(B \,|\, A, C) + \mathrm{P}(A^c \,|\, C)\mathrm{P}(B \,|\, A^c, C) = 1 \qquad (1.12) \end{aligned}$$

so $\mathrm{P}(A, B \,|\, C) = \mathrm{P}(A \,|\, C)\mathrm{P}(B \,|\, C) = 0.5$, which means that A and B are conditionally independent given C. However,

$$\begin{aligned} \mathrm{P}(A, B \,|\, C^c) &= \mathrm{P}(A \,|\, C^c)\mathrm{P}(B \,|\, A, C^c) = 0, \qquad (1.13) \\ \mathrm{P}(B \,|\, C^c) &= \mathrm{P}(A \,|\, C^c)\mathrm{P}(B \,|\, A, C^c) + \mathrm{P}(A^c \,|\, C^c)\mathrm{P}(B \,|\, A^c, C^c) = 0.5 \qquad (1.14) \end{aligned}$$

so $\mathrm{P}(A, B \,|\, C^c) \neq \mathrm{P}(A \,|\, C^c)\mathrm{P}(B \,|\, C^c) = 0.25$, which means that A and B are not conditionally independent given C^c.

1.5 (Partition and independence) Any two events A and B in a partition are disjoint, which means that $\mathrm{P}(A \cap B) = 0$. If they are independent then $\mathrm{P}(A)\,\mathrm{P}(B) = \mathrm{P}(A \cap B) = 0$ so that either $\mathrm{P}(A) = 0$ or $\mathrm{P}(B) = 0$.

1.6 (Conditional probability and complements) Yes. $\mathrm{P}(A|B) = 1$ implies that $\mathrm{P}(A \cap B) = P(B)$, so that

$$\mathrm{P}\left(B^c|A^c\right) = \frac{\mathrm{P}\left(A^c \cap B^c\right)}{\mathrm{P}\left(A^c\right)} \tag{1.15}$$

$$= \frac{\mathrm{P}\left((A \cup B)^c\right)}{\mathrm{P}\left(A^c\right)} \quad \text{by DeMorgan's law} \tag{1.16}$$

$$= \frac{1 - \mathrm{P}(A \cup B)}{1 - \mathrm{P}(A)} \tag{1.17}$$

$$= \frac{1 - \mathrm{P}(A) - \mathrm{P}(B) + \mathrm{P}(A \cap B)}{1 - \mathrm{P}(A)} \tag{1.18}$$

$$= \frac{1 - \mathrm{P}(A)}{1 - \mathrm{P}(A)} \quad \text{because } \mathrm{P}(A \cap B) = P(B) \tag{1.19}$$

$$= 1. \tag{1.20}$$

1.7 (The Linda problem) Let D be the event that the provided description applies to a person, B the event that someone is a bank teller and F the event that someone is active in the feminist movement. Since $D \cap B \ \cap F$ and $D \cap B \ \cap F^c$ form a partition of $D \cap B$, by the definition of conditional probability and the law of total probability,

$$\text{Conditional probability that Linda is a bank teller given the description} \tag{1.21}$$

$$= \mathrm{P}(B \,|\, D) \tag{1.22}$$

$$= \frac{\mathrm{P}(B, D)}{\mathrm{P}(D)} \tag{1.23}$$

$$= \frac{\mathrm{P}(B, D, F) + \mathrm{P}(B, D, F^c)}{\mathrm{P}(D)} \tag{1.24}$$

$$\geq \frac{\mathrm{P}(B, D, F)}{\mathrm{P}(D)} \tag{1.25}$$

$$= \mathrm{P}(B, F \,|\, D) \tag{1.26}$$

$$= \text{Conditional probability that Linda is a bank teller and active in the feminist movement given the description.} \tag{1.27}$$

1.8 (Quiz)

$$\mathrm{P}(\text{2nd question right} \,|\, \text{1st question right}) \tag{1.28}$$

$$\approx \frac{\text{Number of times 1st and 2nd question are both right}}{\text{Number of times 1st question is right}} \tag{1.29}$$

$$= \frac{4}{8} = \frac{1}{2}. \tag{1.30}$$

$$\mathrm{P}(\text{2nd question right} \,|\, \text{1st question wrong}) \tag{1.31}$$

$$\approx \frac{\text{Number of times 1st question is wrong and 2nd question is right}}{\text{Number of times 1st question is wrong}} \tag{1.32}$$

$$= \frac{1}{2}. \tag{1.33}$$

The conditional probabilities are the same, suggesting the events could be independent (we can't be sure because of limited data).

1.9 (Baby name)

a) We define the events G, B and M to represent that the baby is a girl, that Bob thinks

the baby is a girl, and that Margaret thinks that the baby is a girl. By the law of total probability and the chain rule

$$\begin{aligned} \mathrm{P}(B \cap M) &= \mathrm{P}(G \cap B \cap M) + \mathrm{P}\left(G^c \cap B \cap M\right) && (1.34)\\ &= \mathrm{P}(G)\,\mathrm{P}(M \mid G)\,\mathrm{P}(B \mid M) + \mathrm{P}\left(G^c\right)\mathrm{P}\left(M \mid G^c\right)\mathrm{P}(B \mid M) \\ &= 0.5 \cdot 0.8 \cdot 0.9 + 0.5 \cdot 0.2 \cdot 0.9 && (1.35)\\ &= 0.45. && (1.36) \end{aligned}$$

b) By Bayes' rule, the law of total probability and the chain rule

$$\begin{aligned} &\mathrm{P}(G \mid B) && (1.37)\\ &= \frac{\mathrm{P}(G \cap B)}{\mathrm{P}(B)} && (1.38)\\ &= \frac{\mathrm{P}(G \cap M^c \cap B) + \mathrm{P}(G \cap M \cap B)}{\mathrm{P}(G^c \cap M^c \cap B)) + \mathrm{P}(G^c \cap M \cap B) + \mathrm{P}(G \cap M^c \cap B) + \mathrm{P}(G \cap M \cap B)} \\ &= \frac{0.5 \cdot 0.2 \cdot 0.1 + 0.5 \cdot 0.8 \cdot 0.9}{0.5 \cdot 0.8 \cdot 0.1 + 0.5 \cdot 0.2 \cdot 0.9 + 0.5 \cdot 0.2 \cdot 0.1 + 0.5 \cdot 0.8 \cdot 0.9} && (1.39)\\ &= 0.74. && (1.40) \end{aligned}$$

1.10 (Cake)

a) By the definition of conditional probability,

$$\mathrm{P}(\text{Scott helps} \mid \text{On time}) = \frac{\mathrm{P}(\text{Scott helps, On time})}{\mathrm{P}(\text{On time})}. \quad (1.41)$$

Under the independence assumptions, by the law of total probability and the chain rule

$$\begin{aligned} &\mathrm{P}(\text{Scott helps, On time}) && (1.42)\\ &= \mathrm{P}(\text{Scott helps, Antonis helps, On time}) + \mathrm{P}(\text{Scott helps, Antonis doesn't, On time}) \\ &= \mathrm{P}(\text{Scott helps})\,\mathrm{P}(\text{Antonis helps})\,\mathrm{P}(\text{On time} \mid \text{Scott helps, Antonis helps}) \\ &\quad + \mathrm{P}(\text{Scott helps})\,\mathrm{P}(\text{Antonis doesn't})\,\mathrm{P}(\text{On time} \mid \text{Scott helps, Antonis doesn't}) \\ &= 0.4 \cdot 0.8 \cdot 1 + 0.4 \cdot 0.2 \cdot 0.5 && (1.43)\\ &= 0.36. && (1.44) \end{aligned}$$

Similarly,

$$\begin{aligned} &\mathrm{P}(\text{Scott doesn't, On time}) && (1.45)\\ &= \mathrm{P}(\text{Scott doesn't, Antonis helps, On time}) + \mathrm{P}(\text{Scott doesn't, Antonis doesn't, On time}) \\ &= \mathrm{P}(\text{Scott doesn't})\,\mathrm{P}(\text{Antonis helps})\,\mathrm{P}(\text{On time} \mid \text{Scott doesn't, Antonis helps}) \\ &\quad + \mathrm{P}(\text{Scott doesn't})\,\mathrm{P}(\text{Antonis doesn't})\,\mathrm{P}(\text{On time} \mid \text{Scott doesn't, Antonis doesn't}) \\ &= 0.6 \cdot 0.8 \cdot 0.5 + 0.6 \cdot 0.2 \cdot 0 && (1.46)\\ &= 0.24. && (1.47) \end{aligned}$$

Since $\mathrm{P}(\text{On time}) = \mathrm{P}(\text{Scott helps, On time}) + \mathrm{P}(\text{Scott doesn't, On time}) = 0.6$,

$$\mathrm{P}(\text{Scott helps} \mid \text{On time}) = \frac{0.36}{0.6} = 0.6. \quad (1.48)$$

b) By the definition of conditional probability,

$$\mathrm{P}(\text{Scott helps} \mid \text{On time, Antonis helps}) = \frac{\mathrm{P}(\text{Scott helps, Antonis helps, On time})}{\mathrm{P}(\text{On time, Antonis helps})}.$$

Under the independence assumptions, by the chain rule

$$\begin{aligned}
&\mathrm{P}\left(\text{Scott helps}, \text{Antonis helps}, \text{On time}\right) && (1.49)\\
&= \mathrm{P}\left(\text{Scott helps}\right)\mathrm{P}\left(\text{Antonis helps}\right)\mathrm{P}\left(\text{On time} \,|\, \text{Scott helps}, \text{Antonis helps}\right) && (1.50)\\
&= 0.4 \cdot 0.8 \cdot 1 && (1.51)\\
&= 0.32. && (1.52)
\end{aligned}$$

By the law of total probability

$$\begin{aligned}
&\mathrm{P}\left(\text{Antonis helps}, \text{On time}\right) && (1.53)\\
&= \mathrm{P}\left(\text{Antonis helps}, \text{Scott helps}, \text{On time}\right) + \mathrm{P}\left(\text{Antonis helps}, \text{Scott doesn't}, \text{On time}\right)\\
&= \mathrm{P}\left(\text{Antonis helps}\right)\mathrm{P}\left(\text{Scott helps}\right)\mathrm{P}\left(\text{On time} \,|\, \text{Antonis helps}, \text{Scott helps}\right)\\
&\quad + \mathrm{P}\left(\text{Antonis helps}\right)\mathrm{P}\left(\text{Scott doesn't}\right)\mathrm{P}\left(\text{On time} \,|\, \text{Antonis helps}, \text{Scott doesn't}\right)\\
&= 0.8 \cdot 0.4 \cdot 1 + 0.8 \cdot 0.6 \cdot 0.5 && (1.54)\\
&= 0.56. && (1.55)
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathrm{P}\left(\text{Scott helps} \,|\, \text{On time}, \text{Antonis helps}\right) &= \frac{0.32}{0.56} && (1.56)\\
&= 0.57. && (1.57)
\end{aligned}$$

The event *Scott helps* is not conditionally independent from the event *Antonis helps* given the event *Milena finishes on time*, because

$$\mathrm{P}\left(\text{Scott helps} \,|\, \text{On time}, \text{Antonis helps}\right) \neq \mathrm{P}\left(\text{Scott helps} \,|\, \text{On time}\right). \quad (1.58)$$

1.11 (Baby sleep) We define the events W, G and B to represent that the baby wakes up, is a good sleeper and eats bad food respectively.

a) By the chain rule and the law of total probability

$$\begin{aligned}
\mathrm{P}\left(W\right) &= \mathrm{P}\left(B^c, W\right) + \mathrm{P}\left(B, W\right) && (1.59)\\
&= \mathrm{P}\left(B^c, G^c, W\right) + \mathrm{P}\left(B^c, G, W\right) + \mathrm{P}\left(B\right)\mathrm{P}\left(W \,|\, B\right) && (1.60)\\
&= \mathrm{P}\left(B^c\right)\mathrm{P}\left(G^c \,|\, B^c\right)\mathrm{P}\left(W \,|\, B^c, G^c\right) && (1.61)\\
&\quad + \mathrm{P}\left(B^c\right)\mathrm{P}\left(G \,|\, B^c\right)\mathrm{P}\left(W \,|\, B^c, G\right) + \mathrm{P}\left(B\right)\mathrm{P}\left(W \,|\, B\right). && (1.62)
\end{aligned}$$

Here we need independence assumptions to make progress. It seems reasonable that the food being good or not (events B and B^c) be independent from the baby being a good sleeper or not (events G and G^c). Under that assumption, and noticing that $\mathrm{P}\left(W \,|\, B\right) = 1$,

$$\begin{aligned}
\mathrm{P}\left(W\right) &= \mathrm{P}\left(B^c\right)\left(\mathrm{P}\left(G^c\right)\mathrm{P}\left(W \,|\, B^c, G^c\right) + \mathrm{P}\left(G\right)\mathrm{P}\left(W \,|\, B^c, G\right)\right) + \mathrm{P}\left(B\right)\\
&= 0.9 \cdot (0.6 \cdot 0.1 + 0.4 \cdot 0.8) + 0.1 && (1.63)\\
&= 0.442. && (1.64)
\end{aligned}$$

b) By Bayes' rule

$$\begin{aligned}
\mathrm{P}\left(B \,|\, W\right) &= \frac{\mathrm{P}\left(B\right)\mathrm{P}\left(W \,|\, B\right)}{\mathrm{P}\left(W\right)} && (1.65)\\
&= \frac{0.1 \cdot 1}{0.442} && (1.66)\\
&= 0.226. && (1.67)
\end{aligned}$$

c) By the definition of conditional probability, the law of total probability and the chain rule

$$\begin{aligned} \mathrm{P}\left(B \mid W, G\right) &= \frac{\mathrm{P}\left(B, G, W\right)}{\mathrm{P}\left(G, W\right)} && (1.68)\\ &= \frac{\mathrm{P}\left(B\right)\mathrm{P}\left(G\right)\mathrm{P}\left(W \mid B, G\right)}{\mathrm{P}\left(B\right)\mathrm{P}\left(G\right)\mathrm{P}\left(W \mid B, G\right) + \mathrm{P}\left(B^c\right)\mathrm{P}\left(G\right)\mathrm{P}\left(W \mid B^c, G^c\right)}. \end{aligned}$$

If the food is bad, the baby wakes up no matter what, so $\mathrm{P}\left(W \mid B, G\right) = 1$. Consequently

$$\begin{aligned} \mathrm{P}\left(B \mid W, G\right) &= \frac{\mathrm{P}\left(B\right)}{\mathrm{P}\left(B\right) + \mathrm{P}\left(B^c\right)\mathrm{P}\left(W \mid B^c, G\right)} && (1.69)\\ &= \frac{0.1}{0.1 + 0.9 \cdot 0.1} && (1.70)\\ &= 0.526 && (1.71) \end{aligned}$$

d) The events B and G are not conditionally independent given W because if they were $\mathrm{P}\left(B \mid W, G\right)$ would equal $\mathrm{P}\left(B \mid W\right)$. This makes sense because if we know that the baby has woken up, whether the food is bad or not provides information about whether the baby is a good sleeper (and vice versa). In particular, conditioned on W, if the baby is good, then the food is more likely to be bad.

1.12 (COVID-19 tests)

a) Yes, it is reasonable to assume that the test only depends on whether that particular employee is ill, and not the others, and the events *Employee i is ill*, for $1 \leq i \leq 10$, are all independent.

b) We define the events $I_1, \ldots, I_{10}$ to represent each employee being ill, and $T_1, \ldots, T_{10}$ to represent that the corresponding test is positive. The event that at least one test is positive is $\cup_{i=1}^{10} T_i$. By De Morgan's laws,

$$\begin{aligned} \mathrm{P}(\cup_{i=1}^{10} T_i) &= 1 - \mathrm{P}((\cup_{i=1}^{10} T_i)^c) && (1.72)\\ &= 1 - \mathrm{P}(\cap_{i=1}^{10} T_i^c). && (1.73) \end{aligned}$$

By the independence assumption

$$\begin{aligned} \mathrm{P}(\cap_{i=1}^{10} T_i^c) &= \Pi_{i=1}^{10} \mathrm{P}(T_i^c) && (1.74)\\ &= \Pi_{i=1}^{10} \mathrm{P}(I_i)\mathrm{P}(T_i^c \mid I_i) + \mathrm{P}(I_i^c)\mathrm{P}(T_i^c \mid I_i^c) && (1.75)\\ &= \left(0.01 \cdot 0.02 + 0.99 \cdot 0.95\right)^{10} && (1.76)\\ &= 0.543. && (1.77) \end{aligned}$$

We conclude that $\mathrm{P}(\cup_{i=1}^{10} T_i) = 0.457$.

c) By the definition of conditional probability,

$$\mathrm{P}(\cap_{i=1}^{10} I_i^c \mid \cup_{j=1}^{10} T_j) = \frac{\mathrm{P}(\cap_{i=1}^{10} I_i^c, \cup_{j=1}^{10} T_j)}{\mathrm{P}(\cup_{j=1}^{10} T_j)}. \quad (1.78)$$

The denominator was computed above. By the chain rule, the independence assumptions and DeMorgan's laws, the numerator equals

$$\begin{aligned} \mathrm{P}(\cap_{i=1}^{10} I_i^c, \cup_{j=1}^{10} T_j) &= \mathrm{P}(\cap_{i=1}^{10} I_i^c)\mathrm{P}(\cup_{j=1}^{10} T_j \mid \cap_{k=1}^{10} I_k^c) && (1.79)\\ &= \Pi_{i=1}^{10} \mathrm{P}(I_i^c)\left(1 - \mathrm{P}(\cap_{j=1}^{10} T_j^c \mid \cap_{k=1}^{10} I_k^c)\right) && (1.80) \end{aligned}$$

We assume that T_j^c is conditionally independent of T_l^c, $j \neq l$, conditioned on $\cap_k I_k^c$: if we

know that the other people do not have COVID-19, it is plausible that the outcome of the other tests should not provide information about the jth test. Under this assumption

$$\mathrm{P}(\cap_{j=1}^{10} T_j^c \mid \cap_{k=1}^{10} I_k^c) = \Pi_{j=1}^{10} \mathrm{P}(T_j^c \mid \cap_{k=1}^{10} I_k^c). \tag{1.81}$$

It is reasonable to assume that T_i is conditionally independent of $\cap_{j \neq i} I_j^c$ given I_i because one would expect that T_i only depends on I_i. Under this assumption

$$\Pi_{j=1}^{10} \mathrm{P}(T_j^c \mid \cap_{k=1}^{10} I_k^c) = \Pi_{j=1}^{10} \mathrm{P}(T_j^c \mid I_j^c). \tag{1.82}$$

Putting everything together

$$\mathrm{P}(\cap_{i=1}^{10} I_i^c \mid \cup_{j=1}^{10} T_j) = \frac{\Pi_{i=1}^{10} \mathrm{P}(I_i^c)(1 - \Pi_{j=1}^{10} \mathrm{P}(T_j^c \mid I_j^c))}{\mathrm{P}(\cup_{j=1}^{10} T_j)} \tag{1.83}$$

$$= \frac{0.99^{10}(1 - 0.95^{10})}{0.457} \tag{1.84}$$

$$= 0.793. \tag{1.85}$$

1.13 (Boxing championship)

a) From the assumptions,

$$\mathrm{P}\left(\text{Manny is champion}\right) = \mathrm{P}\left(\text{Manny beats Saul}\right) \cdot \mathrm{P}\left(\text{Manny beats Floyd}\right) \tag{1.86}$$

$$= 0.4 \cdot 0.25 = 0.1, \tag{1.87}$$

$$\mathrm{P}\left(\text{Saul is champion}\right) = \mathrm{P}\left(\text{Saul beats Manny}\right) \cdot \mathrm{P}\left(\text{Saul beats Floyd}\right) \tag{1.88}$$

$$= 0.6 \cdot 0.1 = 0.06, \tag{1.89}$$

$$\mathrm{P}\left(\text{Floyd loses}\right) = \mathrm{P}\left(\text{Manny is champion} \cup \text{Saul is champion}\right) \tag{1.90}$$

$$= \mathrm{P}\left(\text{Manny is champion}\right) + \mathrm{P}\left(\text{Saul is champion}\right) \tag{1.91}$$

$$= 0.16. \tag{1.92}$$

The conditional probability of interest is $\mathrm{P}\left(\text{Manny is champion} \,|\, \text{Floyd loses}\right)$. The event *Manny is champion* is included in the event *Floyd loses*, so their intersection is *Manny is champion*. Consequently,

$$\mathrm{P}\left(\text{Manny is champion} \,|\, \text{Floyd loses}\right) = \frac{\mathrm{P}\left(\text{Manny is champion}\right)}{\mathrm{P}\left(\text{Floyd loses}\right)} \tag{1.93}$$

$$= \frac{0.1}{0.16} = \frac{5}{8}. \tag{1.94}$$

b) There are 4 simulations in which Floyd loses, and Manny wins in 3 of them, so the estimated probability is 3/4. It is not surprising that the answer is different because we are approximating the probability using only 4 simulations.

1.14 (Videogame)

a) By the independence assumption,

$$\mathrm{P}\left(\text{Wins game}\right) \tag{1.95}$$

$$= \mathrm{P}\left(\text{Beats Honda}\right) \mathrm{P}\left(\text{Beats Zangief}\right) \mathrm{P}\left(\text{Beats Blanka}\right) \tag{1.96}$$

$$= \left(1 - \mathrm{P}\left(\text{Loses to Honda}\right)\right) \left(1 - \mathrm{P}\left(\text{Loses to Zangief}\right)\right) \left(1 - \mathrm{P}\left(\text{Loses to Blanka}\right)\right)$$

$$= (1 - (1 - 0.8)^2)(1 - (1 - 0.5)^2)(1 - (1 - 0.4)^2) \tag{1.97}$$

$$= 0.4608. \tag{1.98}$$

b) Game 1 (win): Beats Honda in 2 fights, beats Zangief, beats Blanka in 2 fights.
Game 2 (loss): Beats Honda in 2 fights, loses to Zangief.
Game 3 (loss): Beats Honda, beats Zangief in 2 fights, loses to Blanka.
Game 4 (loss): Beats Honda, beats Zangief in 2 fights, loses to Blanka.
Game 5 (win): Beats Honda, beats Zangief in 2 fights, beats Blanka.
Game 6 (win): Beats Honda, beats Zangief, beats Blanka in 2 fights.
Game 7 (win): Beats Honda in two fights, beats Zangief in 2 fights, beats Blanka.
Game 8 (loss): Beats Honda, loses to Zangief.
Game 9 (loss): Beats Honda, beats Zangief, loses to Blanka.
Game 10 (loss): Loses to Honda.
Game 11 (win): Beats Honda, beats Zangief in 2 fights, beats Blanka in 2 fights.
Game 12 (loss): Beats Honda, beats Zangief in 2 fights, loses to Blanka.

The estimated probability of winning the game is $5/12 = 0.42$, which is quite close to the true value. We can improve the accuracy by performing more simulations.

1.15 (Rare event) The probability of not observing the event in n independent simulations is $(1 - \mathrm{P}(A))^n$, consequently we require

$$1 - (1 - \mathrm{P}(A))^n \geq 0.99, \tag{1.99}$$

which is equivalent to

$$0.01 \geq (1 - \mathrm{P}(A))^n . \tag{1.100}$$

Taking logarithms on both sides,

$$\log 0.01 \geq n \log (1 - \mathrm{P}(A)) , \tag{1.101}$$

which implies

$$n \geq \frac{\log 0.01}{\log (1 - \mathrm{P}(A))}, \tag{1.102}$$

because $\log (1 - \mathrm{P}(A))$ is negative. For $\mathrm{P}(A) := 0.01$, $n \geq 458.2$, so we need at least 459 simulations.

1.16 (Streak of heads)

a) We represent heads with 1 and tails with 0. To compute the probabilities, we consider the $2^5 = 32$ possible sequences:
00000 00001 00010 00011 00100 00101 00110 00111 01000 01001 01010 01011 01100 01101 01110 01111 10000 10001 10010 10011 10100 10101 10110 10111 11000 11001 11010 11011 11100 11101 11110 11111
We have

$$\begin{aligned}
&\mathrm{P}(\text{sequence equals 00000}) && (1.103)\\
&= \mathrm{P}(\text{1st roll equals 0}, \text{2nd roll equals 0}, \ldots, \text{5th roll equals 0}) && (1.104)\\
&= \mathrm{P}(\text{1st roll equals 0})\mathrm{P}(\text{2nd roll equals 0}) \cdots \mathrm{P}(\text{5th roll equals 0}) && (1.105)\\
&= \frac{1}{32} && (1.106)
\end{aligned}$$

and by the same argument, all the other sequences also have probability 1/32.

Since the probability of the union of disjoint events is the sum of the individual probabilities,

$$\mathrm{P}(\text{at most 0 heads in a row}) = \frac{1}{32}, \tag{1.107}$$

$$\mathrm{P}(\text{at most 1 heads in a row}) = \frac{12}{32}, \tag{1.108}$$

$$\mathrm{P}(\text{at most 2 heads in a row}) = \frac{11}{32}, \tag{1.109}$$

$$\mathrm{P}(\text{at most 3 heads in a row}) = \frac{5}{32}, \tag{1.110}$$

$$\mathrm{P}(\text{at most 4 heads in a row}) = \frac{2}{32}, \tag{1.111}$$

$$\mathrm{P}(\text{at most 5 heads in a row}) = \frac{1}{32}. \tag{1.112}$$

b) A possible implementation in Python is:

```
def p_longest_streak(n, tries):
    p_longest = np.zeros(n+1)
    for i in range(tries):
        current_streak = 0
        longest_streak = 0
        for j in range(n):
            if np.random.rand() > 0.5:
                current_streak = current_streak + 1
            else:
                if current_streak > longest_streak:
                    longest_streak = current_streak
                current_streak = 0
        if current_streak > longest_streak:
            longest_streak = current_streak
        p_longest[longest_streak] = p_longest[longest_streak] + 1./tries
    return p_longest
```

The images are shown in Figure 1.1.

c) The probability is 0.319. It is therefore not unlikely to find a streak of 8 or more heads in a sequence of 200 fair coin flips, so it is very possible that the random generator is fine.

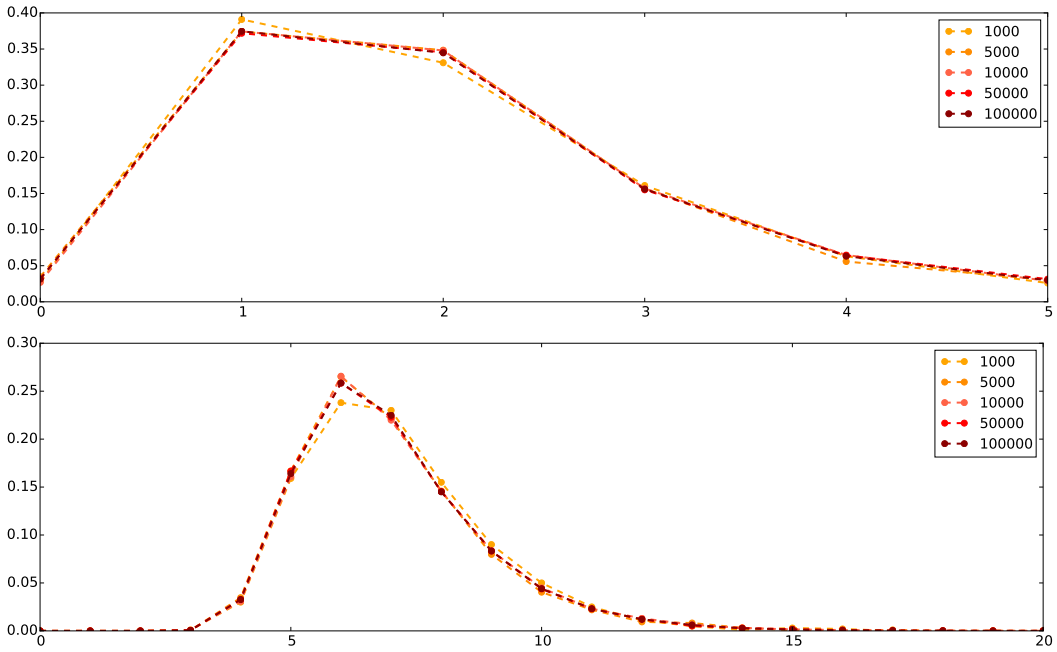


Figure 1.1 Probability of streaks of heads for sequences of length 5 (above) and 200 (below) estimated using different number of Monte Carlo runs (indicated in the legend).

2

Discrete Variables

Exercises

2.1 (Flipping until heads) By Definition 1.28, if the coin flips are independent, the probability of first obtaining $a-1$ tails and then heads equals

$$\begin{aligned} p_{\tilde{a}}(a) &:= \mathrm{P}(\tilde{a} = a) && (2.1)\\ &= \mathrm{P}(\text{flip } 1 = t, \ldots, \text{flip } a-1 = t, \text{flip } a = h) && (2.2)\\ &= \mathrm{P}(\text{flip } 1 = t) \cdots \mathrm{P}(\text{flip } a-1 = t)\, \mathrm{P}(\text{flip } a = h) && (2.3)\\ &= (1-\alpha)^{a-1}\, \alpha, && (2.4) \end{aligned}$$

where h represents heads and t tails. Here a is positive integer.

2.2 (Geometric pmf) For any $\alpha \in (0,1)$ the pmf is positive, so we only need to prove that it sums to one. By the geometric series identity $\sum_{k=1}^{\infty} r^k = r(1-r)^{-1}$, setting $r := 1-\alpha$,

$$\begin{aligned} \sum_{a=1}^{\infty} p_{\tilde{a}}(a) &= \frac{\alpha}{1-\alpha} \sum_{a=1}^{\infty} (1-\alpha)^a && (2.5)\\ &= \frac{\alpha}{1-\alpha} \frac{1-\alpha}{1-(1-\alpha)} = 1. && (2.6) \end{aligned}$$

2.3 (Poisson pmf) For any $\lambda > 0$ the pmf is positive, so we only need to prove that it sums to one. This follows directly from the Taylor series expansion of the exponential function:

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}. \tag{2.7}$$

Indeed,

$$\begin{aligned} \sum_{a=0}^{\infty} p_{\tilde{a}}(a) &= e^{-\lambda} \sum_{a=0}^{\infty} \frac{\lambda^a}{(a)!} && (2.8)\\ &= e^{-\lambda} e^{\lambda} = 1. && (2.9) \end{aligned}$$

2.4 (Memoryless property of the geometric distribution) By definition of conditional probability, we have

$$\mathrm{P}(\tilde{a} = a \,|\, \tilde{a} > 5) = \frac{\mathrm{P}(\tilde{a} = a, \tilde{a} > a)}{\mathrm{P}(\tilde{a} > a)}. \tag{2.10}$$

The numerator is zero unless $a > 5$, in that case

$$\mathrm{P}(\tilde{a} = a \,|\, \tilde{a} > 5) = \frac{(1-\alpha)^{a-1}\alpha}{\sum_{b=6}^{\infty}(1-\alpha)^{b-1}\alpha}. \tag{2.11}$$

The denominator contains the geometric sum

$$\sum_{b=6}^{\infty}(1-\alpha)^{b-1} = \frac{(1-\alpha)^6}{\alpha}. \tag{2.12}$$

We conclude

$$\mathrm{P}(\tilde{a} = a \,|\, \tilde{a} > 5) = (1-\alpha)^{a-6}\alpha. \tag{2.13}$$

This is the same as the geometric pmf if we plug in $a - 5$. In terms of the coin example, if we have obtained 5 tails, now the probability that we have to flip $b = a - 5$ more times is distributed like a geometric random variable. This makes sense because the flips are independent, so there is no difference between this and just considering another sequence of flips starting from the beginning.

2.5 (Old car)

a) A binomial distribution with parameters $n := 50$ and $\theta := 1/10$. They need to assume that the car breaks down independently each time.

b) We denote the binomial by $\tilde{x}$

$$\begin{aligned}
\mathrm{P}\left(\text{At most once}\right) &= p_{\tilde{x}}\left(0\right) + p_{\tilde{x}}\left(1\right) && (2.14)\\
&= \binom{50}{0}\left(1-\frac{1}{10}\right)^{50} + \binom{50}{1}\left(1-\frac{1}{10}\right)^{49}\frac{1}{10} && (2.15)\\
&= 0.9^{50} + 50 \cdot 0.9^{49} \cdot 0.1 && (2.16)\\
&= 3.38 \cdot 10^{-2}. && (2.17)
\end{aligned}$$

As explained in the section on the Poisson distribution, the pmf of a binomial with parameters n and λ/n converges to the pmf of a Poisson with parameter λ (in distribution), we set $\lambda = n/\theta = 5$ for a Poisson random variable $\tilde{y}$ and compute

$$\begin{aligned}
\mathrm{P}\left(\text{At most once}\right) &\approx p_{\tilde{y}}\left(0\right) + p_{\tilde{y}}\left(1\right) && (2.18)\\
&= e^{-\lambda} + \lambda e^{-\lambda} && (2.19)\\
&= 4.04 \cdot 10^{-2}. && (2.20)
\end{aligned}$$

c) The probability that the car breaks down $k-1$ times in $n-1$ drives is equal to the probability that a binomial random variable with parameters $n-1$ and 0.1 equals $k-1$. Consequently, under the assumption that the car breaks down independently each time,

$$\begin{aligned}
&\mathrm{P}\left(\text{breaks down } n\text{th time in } k\text{th drive}\right) && (2.21)\\
&= \mathrm{P}\left(\text{breaks down } k-1 \text{ times in } n-1 \text{ drives} \cap \{\text{breaks down in } k\text{th drive}\}\right) && (2.22)\\
&= \mathrm{P}\left(\text{breaks down } k-1 \text{ times in } n-1 \text{ drives}\right)\mathrm{P}\left(\{\text{breaks down in } k\text{th drive}\}\right) && (2.23)\\
&= \binom{n-1}{k-1}\left(\frac{9}{10}\right)^{n-k}\left(\frac{1}{10}\right)^{k}. && (2.24)
\end{aligned}$$

2.6 (Oil prospector) Let $\tilde{x}$ denote the random variable representing the number of times the prospector has to drill to find oil. By the law of total probability and the conditional

independence assumptions:

$$p_{\tilde{x}}(x) \tag{2.25}$$
$$= \mathrm{P}(\text{no oil in } x-1 \text{ attempts}, \text{oil in } x\text{th}) \tag{2.26}$$
$$= \sum_{r \in \{\text{poor, standard, rich}\}} \mathrm{P}(\text{region is } r)\, \mathrm{P}(\text{no oil in } x-1 \text{ attempts}, \text{oil in } x\text{th} \,|\, \text{region is } r)$$
$$= \frac{1}{10}\frac{1}{2^x} + \frac{8}{10}\frac{9^{x-1}}{10^x} + \frac{1}{10}\frac{99^{x-1}}{100^x} \tag{2.27}$$
$$= \frac{1}{10}\left(\frac{1}{2^x} + \frac{8 \cdot 9^{x-1}}{10^x} + \frac{99^{x-1}}{100^x}\right). \tag{2.28}$$

2.7 (Darts) Let $\tilde{a}$ denote the random variable. Note that the last attempt must always be a success. We can therefore decompose the event *a attempts required* into the intersection of $k-1$ *successes over first* $a-1$ *attempts* and a*th attempt is a success*. Since the attempts are all independent,

$$p_{\tilde{a}}(a) = \mathrm{P}(k-1 \text{ successes over first } a-1 \text{ attempts})\mathrm{P}(a\text{th attempt is a success}). \tag{2.29}$$

By exactly the same reasoning we use to derive the binomial distribution,

$$\mathrm{P}(k-1 \text{ successes over first } a-1 \text{ attempts}) = \binom{a-1}{k-1}\theta^{k-1}(1-\theta)^{a-1-(k-1)}, \tag{2.30}$$

as long as $a \geq k$. We conclude that

$$p_{\tilde{a}}(a) = \binom{a-1}{k-1}\theta^{k}(1-\theta)^{a-k}, \quad \text{for } a \geq k, \tag{2.31}$$

and zero otherwise.

2.8 (Binomial random variable)

$$\mathrm{P}(\tilde{a} = 1 \,|\, \tilde{a} \leq 1) = \frac{\mathrm{P}(\tilde{a} = 1, \tilde{a} \leq 1)}{\mathrm{P}(\tilde{a} \leq 1)} \tag{2.32}$$
$$= \frac{\mathrm{P}(\tilde{a} = 1)}{\mathrm{P}(\tilde{a} = 0) + \mathrm{P}(\tilde{a} = 1)} \tag{2.33}$$
$$= \frac{n\theta(1-\theta)^{n-1}}{(1-\theta)^n + n\theta(1-\theta)^{n-1}} \tag{2.34}$$
$$= \frac{n\theta}{1 + (n-1)\theta}. \tag{2.35}$$

2.9 (Intersection) The parametric model is $p_{\tilde{c}}(-1) = p_{\tilde{c}}(1) = \theta$ and $p_{\tilde{c}}(0) = 1 - 2\theta$. The log likelihood is

$$\log \mathcal{L}_X(\theta) = \sum_{i=1}^{6} \log p_{\tilde{c}}(x_i) \tag{2.36}$$
$$= 4 \log \theta + 2 \log(1 - 2\theta). \tag{2.37}$$

The derivatives of the log likelihood are

$$\frac{\mathrm{d} \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta} = \frac{4}{\theta} - \frac{4}{1-2\theta}, \tag{2.38}$$
$$\frac{\mathrm{d}^2 \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta^2} = -\frac{4}{\theta^2} - \frac{8}{(1-2\theta)^2} < 0 \quad \text{for all } \theta \in [0, 1]. \tag{2.39}$$

The function is concave so we can set the first derivative to zero, obtaining $\theta_{\mathrm{ML}} = 1/3$.

2.10 (Bad apples)

a) The probability of picking a bad apple is θ, so

$$\begin{aligned} p_{\tilde{t}}(1) &= (\text{Test is positive}) && (2.40)\\ &= \mathrm{P}\left(\text{Test is positive}, \text{Bad apple}\right) + \mathrm{P}\left(\text{Test is positive}, \text{Good apple}\right) && (2.41)\\ &= \mathrm{P}\left(\text{Bad apple}\right)\mathrm{P}\left(\text{Test is positive} \,|\, \text{Bad apple}\right) && (2.42)\\ &\quad + \mathrm{P}\left(\text{Good apple}\right)\mathrm{P}\left(\text{Test is positive} \,|\, \text{Good apple}\right) && (2.43)\\ &= 0.9\theta + 0.4(1-\theta) && (2.44)\\ &= 0.5\theta + 0.4. && (2.45) \end{aligned}$$

Since the pmf must add to one,

$$\begin{aligned} p_{\tilde{t}}(0) &= 1 - p_{\tilde{t}}(1) && (2.46)\\ &= 0.6 - 0.5\theta. && (2.47) \end{aligned}$$

b) The log-likelihood equals

$$\begin{aligned} \log \mathcal{L}\left(\theta\right) &= \sum_{i=1}^{n} \log p_{\tilde{t}}(x_i) && (2.48)\\ &= 5\log(0.5\theta + 0.4) + 5\log(0.6 - 0.5\theta). && (2.49) \end{aligned}$$

The first derivative is

$$\begin{aligned} \frac{\mathrm{d}\log \mathcal{L}\left(\theta\right)}{\mathrm{d}\theta} &= \frac{5 \cdot 0.5}{0.5\theta + 0.4} + \frac{5 \cdot (-0.5)}{0.6 - 0.5\theta} && (2.50)\\ &= \frac{2.5}{0.5\theta + 0.2} - \frac{2.5}{0.6 - 0.5\theta} && (2.51) \end{aligned}$$

and the second

$$\frac{\mathrm{d}^2 \log \mathcal{L}\left(\theta\right)}{\mathrm{d}\theta^2} = -\frac{2.5 \cdot 0.5}{(0.5\theta + 0.4)^2} - \frac{2.5 \cdot 0.5}{(0.6 - 0.5\theta)^2} < 0, \tag{2.52}$$

so the function is concave. Setting the first derivative to zero yields,

$$\begin{aligned} 0.5\theta_{\mathrm{ML}} + 0.4 &= 0.6 - 0.5\theta_{\mathrm{ML}}, && (2.53)\\ \theta_{\mathrm{ML}} &= 0.2. && (2.54) \end{aligned}$$

2.11 (False positives)

a) A test is negative only if the person does not have the disease and there is no false positive so

$$\begin{aligned} p_{\tilde{t}}(0) &= (\text{Test is negative}) && (2.55)\\ &= \mathrm{P}\left(\text{No disease}, \text{No false positive}\right) && (2.56)\\ &= \mathrm{P}\left(\text{No disease}\right)\mathrm{P}\left(\text{No false positive} \,|\, \text{No disease}\right) && (2.57)\\ &= 0.9\left(1-\theta\right). && (2.58) \end{aligned}$$

Since the pmf must add to one,

$$\begin{aligned} p_{\tilde{t}}(1) &= 1 - p_{\tilde{t}}(0) && (2.59)\\ &= 0.1 + 0.9\theta. && (2.60) \end{aligned}$$

b) Let $n_{\text{neg}} := n - n_{\text{pos}}$. The log likelihood is

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p_{\tilde{t}}(x_i) \tag{2.61}$$

$$= n_{\text{neg}} \log 0.9\,(1-\theta) + n_{\text{pos}} \log(0.1 + 0.9\theta) \tag{2.62}$$

$$= n_{\text{neg}} \log 0.9 + n_{\text{neg}} \log(1-\theta) + n_{\text{pos}} \log(0.1 + 0.9\theta)\,. \tag{2.63}$$

The first derivative is

$$\frac{\mathrm{d} \log \mathcal{L}(\theta)}{\mathrm{d}\theta} = -\frac{n_{\text{neg}}}{1-\theta} + \frac{0.9 n_{\text{pos}}}{0.1 + 0.9\theta} \tag{2.64}$$

and the second

$$\frac{\mathrm{d}^2 \log \mathcal{L}(\theta)}{\mathrm{d}\theta^2} = -\frac{n_{\text{neg}}}{(1-\theta)^2} - \frac{0.9^2 n_{\text{pos}}}{(0.1 + 0.9\theta)^2} < 0, \tag{2.65}$$

so the function is concave. Setting the first derivative to zero yields,

$$0.1 n_{\text{neg}} + 0.9\theta_{\max} n_{\text{neg}} = 0.9 n_{\text{pos}} - 0.9 n_{\text{pos}} \theta_{\max}, \tag{2.66}$$

which implies

$$\theta_{\max} = \frac{0.9 n_{\text{pos}} - 0.1 n_{\text{neg}}}{0.9(n_{\text{pos}} + n_{\text{neg}})} \tag{2.67}$$

$$= \frac{n_{\text{pos}}}{n} - \frac{0.1 n_{\text{neg}}}{0.9 n} \tag{2.68}$$

$$= \phi - \frac{1}{9}(1-\phi) \tag{2.69}$$

$$= \frac{10\phi - 1}{9}. \tag{2.70}$$

Notice that for $\phi < 0.1$ the maximum is negative ($\theta_{\max} < 0$), and consequently out of the interval $[0, 1]$. Since we know that the log-likelihood is monotone decreasing for values of θ larger than the maximum (as it is concave), in such cases the maximum within $[0, 1]$ is at zero. We conclude that

$$\theta_{\text{ML}} = \begin{cases} 0 & \text{if } \phi < 0.1, \\ \frac{10\phi - 1}{9} & \text{if } \phi \geq 0.1. \end{cases} \tag{2.71}$$

2.12 (Chess games)

a) Under the independence assumption, the likelihood and log-likelihood equal

$$\mathcal{L}_X(\theta) = \theta^4 \alpha^2 (1 - \theta - \alpha)^4\,, \tag{2.72}$$

$$\log \mathcal{L}_X(\theta) = 4 \log \theta + 2 \log \alpha + 4 \log(1 - \theta - \alpha)\,. \tag{2.73}$$

The plot is shown in Figure 2.1.

b) From the plot we can see that the function has a single maximum. To find it, we set the partial derivatives to zero:

$$\frac{\mathrm{d} \log \mathcal{L}_X(\theta)}{\mathrm{d}\theta} = \frac{4}{\theta} - \frac{4}{1 - \theta - \alpha}, \tag{2.74}$$

$$\frac{\mathrm{d} \log \mathcal{L}_X(\alpha)}{\mathrm{d}\theta} = \frac{2}{\alpha} - \frac{4}{1 - \theta - \alpha}. \tag{2.75}$$

$$\tag{2.76}$$

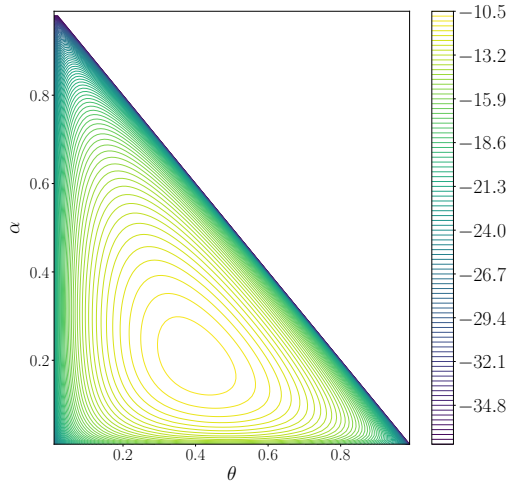


Figure 2.1 Log-likelihood of the parametric model for the *Chess games* problem.

Setting the first expression equal to zero yields $\alpha_{\mathrm{ML}} = 1 - 2\theta_{\mathrm{ML}}$. Plugging into the second and solving the equation, we conclude $\theta_{\mathrm{ML}} = 0.4$ and $\alpha_{\mathrm{ML}} = 0.2$.

c) The empirical pmf would assign $4/10 = 0.4$ to the probability of Garry winning, $2/10 = 0.2$ to the probability of Anish winning, and $4/10 = 0.4$ to the probability of a draw. This is exactly equivalent to the parametric model.

2.13 (Maximum-likelihood estimator for the geometric distribution) The log-likelihood equals

$$\log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\alpha) = \sum_{i=1}^{n} \log p_\alpha(x_i) \tag{2.77}$$

$$= \sum_{i=1}^{n} \log\left((1-\alpha)^{x_i-1}\alpha\right) \tag{2.78}$$

$$= \sum_{i=1}^{n} \left(\log\alpha + (x_i - 1)\log(1-\alpha)\right) \tag{2.79}$$

$$= n\log\alpha + \left(\sum_{i=1}^{n} x_i - n\right)\log(1-\alpha). \tag{2.80}$$

The derivative and second derivative of the log-likelihood are

$$\frac{\mathrm{d}\log\mathcal{L}_{\{x_1,\ldots,x_n\}}(\alpha)}{\mathrm{d}\alpha} = \frac{n}{\alpha} - \frac{\sum_{i=1}^{n} x_i - n}{1-\alpha}, \tag{2.81}$$

$$\frac{\mathrm{d}^2\log\mathcal{L}_{\{x_1,\ldots,x_n\}}(\alpha)}{\mathrm{d}\alpha^2} = -\frac{n}{\alpha^2} - \frac{\sum_{i=1}^{n} x_i - n}{(1-\alpha)^2} \tag{2.82}$$

$$= -n\cdot\frac{1 - 2\alpha + \alpha^2\frac{1}{n}\sum_{i=1}^{n} x_i}{\alpha^2(1-\alpha)^2} \tag{2.83}$$

$$< -n\cdot\frac{1-2\alpha+\alpha^2}{\alpha^2(1-\alpha)^2} = -\frac{n}{\alpha^2} < 0. \tag{2.84}$$

Where we have used the assumption that $x_i \geq 1$ for all i, so that $\frac{1}{n}\sum_{i=1}^{n} x_i \geq 1$. The function is concave, as the second derivative is negative. The maximum-likelihood estimator of the parameter (see Definition 2.25) is consequently at the point where the first derivative equals zero, namely

$$\alpha_{\mathrm{ML}} := \arg\max_{\alpha} \log \mathcal{L}_{\{x_1,\ldots,x_n\}}(\alpha) \tag{2.85}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1}. \tag{2.86}$$

2.14 (Multinoulli distribution)

a) The parameters θ_1, ..., θ_m must be between 0 and 1, as they represent entries of a pmf. They also need to sum to one, so that the pmf is valid.

b) We are interested in deriving the maximum likelihood estimate of θ_k, for an arbitrary $1 \leq k \leq m$. The likelihood and log-likelihood equal

$$\mathcal{L}_X(\theta) = \prod_{j=1}^{m} \theta_j^{n_j} \tag{2.87}$$

$$= \left(1 - \sum_{l\neq k} \theta_l\right)^{n_k} \prod_{j\neq k} \theta_j^{n_j}, \tag{2.88}$$

$$\log \mathcal{L}_X(\theta) = n_k \log\left(1 - \sum_{l\neq k} \theta_l\right) + \sum_{j\neq k} n_j \log \theta_j, \tag{2.89}$$

where $n_j := \left|\left\{i : x_i = v_j\right\}\right|$ is the number of data points equal to v_j for $1 \leq j \leq m$. We have used the fact that $\theta_k = 1 - \sum_{l\neq k} \theta_l$ because the parameters need to sum up to one. Assuming the log-likelihood is concave, we can obtain the maximum-likelihood estimator setting the partial derivatives (and hence the gradient) to zero:

$$\frac{\mathrm{d}\log \mathcal{L}_X(\theta)}{\mathrm{d}\theta_j} = -\frac{n_k}{1 - \sum_{l\neq k}\theta_l} + \frac{n_j}{\theta_j} = 0, \qquad 1 \leq j \leq m. \tag{2.90}$$

This yields

$$n_k \theta_j = n_j \theta_k, \qquad 1 \leq j \leq m, \tag{2.91}$$

and consequently

$$n_k = \sum_{j=1}^{m} n_k \theta_j = \sum_{j=1}^{m} n_j \theta_k \tag{2.92}$$

$$= n\theta_k, \tag{2.93}$$

because $\sum_{j=1}^{m} n_j = n$. We conclude that the maximum-likelihood estimator of θ_k is $\frac{n_k}{n}$.

c) The parametric model is exactly equivalent to the nonparametric empirical-pmf estimator.

2.15 (Adaptive test)

a) Let $\tilde{c}$ be the number of correct answers.

$$p_{\tilde{c}}(0) = \mathrm{P}\left(\text{1st wrong}\right)\mathrm{P}\left(\text{2nd wrong}\,|\,\text{1st wrong}\right) \tag{2.94}$$
$$= (1-\theta^2) \tag{2.95}$$
$$= 1 - 2\theta + \theta^2, \tag{2.96}$$
$$p_{\tilde{c}}(1) = \mathrm{P}\left(\text{1st correct}\right)\mathrm{P}\left(\text{2nd wrong}\,|\,\text{1st correct}\right) \tag{2.97}$$
$$+ \mathrm{P}\left(\text{1st wrong}\right)\mathrm{P}\left(\text{2nd correct}\,|\,\text{1st wrong}\right) \tag{2.98}$$
$$= \theta\left(1 - \frac{\theta}{2}\right) + (1-\theta)\theta \tag{2.99}$$
$$= 2\theta - \frac{3\theta^2}{2}, \tag{2.100}$$
$$p_{\tilde{c}}(2) = \mathrm{P}\left(\text{1st correct}\right)\mathrm{P}\left(\text{2nd correct}\,|\,\text{1st correct}\right) \tag{2.101}$$
$$= \frac{\theta^2}{2}. \tag{2.102}$$

b)

$$\log \mathcal{L}\left(\theta\right) = \sum_{i=1}^{n} \log p_{\tilde{c}}(x_i) \tag{2.103}$$
$$= \log p_{\tilde{c}}(2) + 2\log p_{\tilde{c}}(1) \tag{2.104}$$
$$= \log\left(\frac{\theta^2}{2}\right) + 2\log\left(2\theta - \frac{3\theta^2}{2}\right) \tag{2.105}$$
$$= 2\log\theta - \log 2 + 2\log\left(2\theta - \frac{3\theta^2}{2}\right). \tag{2.106}$$

$$\frac{\mathrm{d}\log\mathcal{L}\left(\theta\right)}{\mathrm{d}\theta} = \frac{2}{\theta} + \frac{2\left(2-3\theta\right)}{2\theta - \frac{3\theta^2}{2}}. \tag{2.107}$$

Setting the derivative to zero yields

$$\frac{1}{\theta_{\mathrm{ML}}} = \frac{-2 + 3\theta_{\mathrm{ML}}}{2\theta_{\mathrm{ML}} - \frac{3\theta_{\mathrm{ML}}^2}{2}}, \tag{2.108}$$
$$2\theta_{\mathrm{ML}} - \frac{3\theta_{\mathrm{ML}}^2}{2} = -2\theta_{\mathrm{ML}} + 3\theta_{\mathrm{ML}}^2, \tag{2.109}$$
$$4\theta_{\mathrm{ML}} = \frac{9\theta_{\mathrm{ML}}^2}{2}. \tag{2.110}$$

This expression is zero if we set θ_{ML} equal to zero, but then the derivative would be infinite! The only value for which the derivative is zero is

$$\theta_{\mathrm{ML}} = \frac{8}{9}. \tag{2.111}$$

By the assumption that the log-likelihood is concave, this is the maximum-likelihood estimate.

2.16 (Air traffic)

a) To design the parametric model, we assume that:

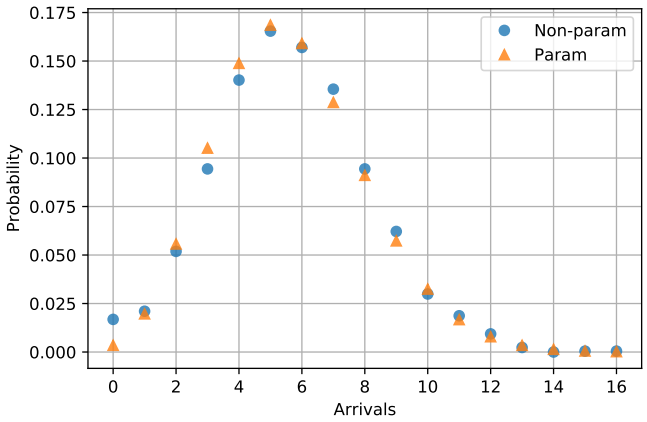


Figure 2.2 Parametric and nonparametric pmfs for the *Air traffic* problem.

a) For any period of time of length t, if t is small enough, the probability of a flight arriving in that period is equal to λt and the probability of more than one flight is negligible. λ is a fixed parameter representing the total flights that we expect to arrive every 10 minutes.
b) Each flight arrives independently from other flights..

These are the assumptions we made to derive the Poisson distribution, so we model the number of arrivals using a Poisson parametric model. The code to fit the nonparametric and parametric models is the following: The parametric and nonparametric pmfs are shown in Figure 2.2.

b) The test RMSD is 0.016 for the nonparametric model and 0.021 for the parametric model. In this case, the nonparametric model is more accurate.

3

Continuous Variables

Exercises

3.1 (Probability of individual points) Let $S := \{s \in \mathbb{R} : \mathrm{P}(\tilde{a} = s) \neq 0\}$ be the set of values that are assigned nonzero probability. Since the events $\tilde{a} = s$ are all disjoint (the function cannot map the same outcome to two values), we must have

$$\sum_{s \in S} \mathrm{P}(\tilde{a} = s) = 1. \tag{3.1}$$

A technical detail is that if S is uncountable the sum may not be well defined. We define it as the supremum of the sums of any countable subset of S. Now consider the following partition of S

$$A_n := \left\{ a \in S : \frac{1}{n-1} \geq \mathrm{P}(\tilde{a} = a) > \frac{1}{n} \right\}, \quad n = 2, 3, \ldots \tag{3.2}$$

None of these sets can contain more than n elements (otherwise the sum of the probabilities would be larger than one, contradicting (3.1)). The union of these subsets is equal to S, and there are a countable number of them, each containing a finite number of elements. We therefore conclude that S is countable.

3.2 (Continuous cdf) For any fixed a, leveraging the properties of probability measures, we have

$$\begin{aligned}
F_{\tilde{a}}(a) &= \mathrm{P}(\tilde{a} \leq a) && (3.3)\\
&= \mathrm{P}\left(\lim_{n \to \infty} \{\tilde{a} \leq a - 1\} \cup \cup_{i=2}^{n} \left\{ a - \frac{1}{n-1} < \tilde{a} \leq a - \frac{1}{n} \right\} \cup \{\tilde{a} = a\} \right) && (3.4)\\
&= \mathrm{P}(\tilde{a} \leq a - 1) + \mathrm{P}\left(\lim_{n \to \infty} \cup_{i=2}^{n} \left\{ a - \frac{1}{n-1} < \tilde{a} \leq a - \frac{1}{n} \right\} \right) + \mathrm{P}(\tilde{a} = a) && (3.5)\\
&= \mathrm{P}(\tilde{a} \leq a - 1) + \lim_{n \to \infty} \sum_{i=2}^{n} \mathrm{P}\left(a - \frac{1}{n-1} < \tilde{a} \leq a - \frac{1}{n} \right) + \mathrm{P}(\tilde{a} = a) && (3.6)\\
&= \lim_{n \to \infty} \mathrm{P}\left(\tilde{a} \leq a - \frac{1}{n} \right) + \mathrm{P}(\tilde{a} = a) && (3.7)\\
&= \lim_{n \to \infty} F_{\tilde{a}}\left(a - \frac{1}{n} \right) + \mathrm{P}(\tilde{a} = a). && (3.8)
\end{aligned}$$

Consequently, $\mathrm{P}(\tilde{a} = a)$ is zero, if and only if

$$\lim_{n \to \infty} F_{\tilde{a}}\left(a - \frac{1}{n} \right) = F_{\tilde{a}}(a), \tag{3.9}$$

i.e. if and only if $F_{\tilde{a}}$ is continuous at a.

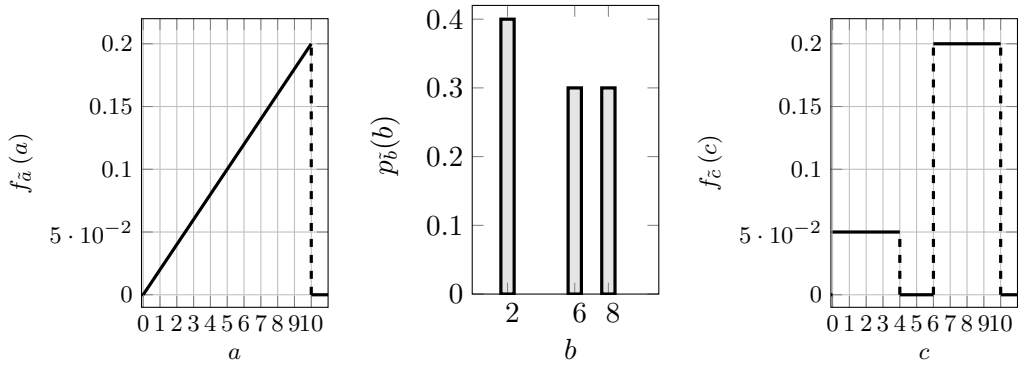


Figure 3.1 Pdfs of $\tilde{a}$ and $\tilde{c}$ and pmf of $\tilde{b}$ in Exercise 3.4.

3.3 (Median of affine transformation) If the median of $\tilde{a}$ is m, then

$$\mathrm{P}(\tilde{a} \leq m) = \frac{1}{2}, \tag{3.10}$$

which implies

$$\begin{aligned} \mathrm{P}(\tilde{b} \leq \alpha m + \beta) &= \mathrm{P}(\alpha\tilde{a} + \beta \leq \alpha m + \beta) && (3.11)\\ &= \mathrm{P}(\tilde{a} \leq m), && (3.12) \end{aligned}$$

so the median of $\tilde{b}$ is $\alpha m + \beta$.

3.4 (Three cdfs)

a) The median is the value for which the cdf equals 0.5. The median of $\tilde{a}$ is 7, the median of $\tilde{b}$ is 6 and the median of $\tilde{c}$ is 7.5.

b) Figure 3.1 shows the pdfs of $\tilde{a}$ and $\tilde{c}$, and the pmf of $\tilde{b}$.

3.5 (Cumulative distribution function)

a) The median is 0.2, because the probability that $\tilde{x} \leq 0.2$ is 0.5.

b) The slope of the cdf is clearly larger at 0.2, so the density at 0.2 is higher than at 0.8.

c)

$$\begin{aligned} \mathrm{P}(\tilde{x} \leq 0.4 \,|\, \tilde{x} < 0.2) &= \frac{\mathrm{P}(0.2 \leq \tilde{x} \leq 0.4 \,|\, \tilde{x} \geq 0.2)}{\mathrm{P}(\tilde{x} \geq 0.2)} && (3.13)\\ &= \frac{F_{\tilde{x}}(0.4) - F_{\tilde{x}}(0.2)}{1 - F_{\tilde{x}}(0.2)} && (3.14)\\ &= \frac{0.8 - 0.5}{0.5} && (3.15)\\ &= 0.6. && (3.16) \end{aligned}$$

d) We apply inverse-transform sampling. Since $F_{\tilde{x}}(0.08) = 0.2$, $F_{\tilde{x}}(0.15) = 0.4$, and $F_{\tilde{x}}(0.25) = 0.55$, the simulated samples are 0.08, 0.15, and 0.25.

3.6 (Step pdf)

a) The pdf must integrate to 1, the value of the integral is $0.8 \cdot b/2 + 0.2 \cdot b/2 = 0.5b$ so b must equal 2.

b) For $a < 0$, the cdf is equal to zero, because the probability that $\tilde{a} \leq a$ is zero. For $0 \leq a < 1$,

$$F_{\tilde{a}}(a) = \mathrm{P}\left(\tilde{a} \leq a\right) \tag{3.17}$$

$$= \int_{t=0}^{a} 0.8 \, \mathrm{d}t \tag{3.18}$$

$$= 0.8t. \tag{3.19}$$

For $1 \leq a < 2$,

$$F_{\tilde{a}}(a) = \mathrm{P}\left(\tilde{a} \leq a\right) \tag{3.20}$$

$$= \int_{t=0}^{1} 0.8 \, \mathrm{d}t + \int_{t=1}^{a} 0.2 \, \mathrm{d}t \tag{3.21}$$

$$= 0.8 + 0.2(a-1) \tag{3.22}$$

$$= 0.6 + 0.2a. \tag{3.23}$$

For $a \geq 2$, the cdf is equal to one, because the probability that $\tilde{a} \leq a$ is one.

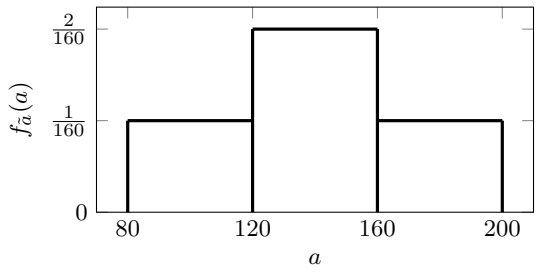
The median of $\tilde{a}$, which we denote by m satisfies $F_{\tilde{a}}(m) = 0.5$. We notice that the probability that $\tilde{a} < 1$ is more than 0.5 (it's 0.8), so the median must be below, which implies

$$F_{\tilde{a}}(m) = 0.8m = 0.5, \tag{3.24}$$

so $m = 5/8$.

3.7 (Pigs)

a) The histogram is shown below.



b) The maximum likelihood estimate for the μ parameter of the Gaussian is the average of the data, which equals 145 kg. This is also the median because the Gaussian pdf is centered at μ and is symmetric, so the probability that the corresponding random variable is smaller than μ is 0.5.

3.8 (Scaled exponential distribution) The cdf of $\tilde{a}$ equals $F_{\tilde{a}}(a) = 1 - \exp(-\lambda a)$ for $a \geq 0$ and

0 for $a < 0$. The cdf of $\tilde{b}$ therefore equals

$$F_{\tilde{b}}(b) := \mathrm{P}(\tilde{b} \leq b) \tag{3.25}$$
$$= \mathrm{P}(\alpha\tilde{a} \leq b) \tag{3.26}$$
$$= \mathrm{P}\left(\tilde{a} \leq \frac{b}{\alpha}\right) \tag{3.27}$$
$$= F_{\tilde{a}}\left(\frac{b}{\alpha}\right) \tag{3.28}$$
$$= 1 - \exp\left(-\frac{\lambda}{\alpha}b\right), \tag{3.29}$$

so $\tilde{b}$ is an exponential random variable with parameter λ/α.

3.9 (Gaussian pdf) Let $f_{\tilde{a}}$ denote the pdf of a Gaussian random variable with mean μ and standard deviation σ. By the change of variables $t = (a - \mu)/\sqrt{2}\sigma$,

$$\int_{-\infty}^{\infty} f_{\tilde{a}}(a)\ \mathrm{d}a = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-t^2\right)\ \mathrm{d}t. \tag{3.30}$$

Let us define

$$I = \int_{-\infty}^{\infty} \exp\left(-t^2\right)\ \mathrm{d}t. \tag{3.31}$$

We perform a change of variable to polar coordinates to the square of I:

$$I^2 = \int_{-\infty}^{\infty} \exp\left(-x^2\right) \mathrm{d}x \int_{-\infty}^{\infty} \exp\left(-y^2\right) \mathrm{d}y \tag{3.32}$$
$$= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \exp\left(-\left(x^2 + y^2\right)\right) \mathrm{d}x\mathrm{d}y \tag{3.33}$$
$$= \int_{\theta=0}^{2\pi} \int_{r=-\infty}^{\infty} r \exp\left(-r^2\right) \mathrm{d}\theta\mathrm{d}r \tag{3.34}$$
$$= \pi \exp(-r^2)]_0^{\infty} = \pi. \tag{3.35}$$

Consequently,

$$\int_{-\infty}^{\infty} f_{\tilde{a}}(a)\ \mathrm{d}a = \frac{I}{\sqrt{\pi}} = 1. \tag{3.36}$$

3.10 (Uniform distribution) Recall that $F_{\tilde{u}}(u) = u$ for $0 \leq u \leq 1$. For $0 \leq w \leq 1$, the cdf of $\tilde{w}$ equals

$$F_{\tilde{w}}(w) = \mathrm{P}\left(1 - \tilde{u} \leq w\right) \tag{3.37}$$
$$= \mathrm{P}\left(\tilde{u} \geq 1 - w\right) \tag{3.38}$$
$$= 1 - \mathrm{P}\left(\tilde{u} \leq 1 - w\right) \tag{3.39}$$
$$= 1 - F_{\tilde{u}}(1 - w) = 1 - (1 - w) = w. \tag{3.40}$$

For $w > 1$, $1 - \tilde{u}$ is always smaller than w, so $F_{\tilde{w}}(w) = 1$. For $w < 0$, $1 - \tilde{u}$ is always greater than w, so $F_{\tilde{w}}(w) = 0$. We conclude that $\tilde{w}$ indeed follows a uniform distribution in $[0, 1]$.

3.11 (Nuclear power plant)

a) The pdf should integrate to one. We have

$$\int_{-\infty}^{\infty} f_{\tilde{t}}(t)\ \mathrm{d}t = \int_{-1}^{0} \alpha\ \mathrm{d}t + \int_{0}^{\infty} \alpha \exp(t)\ \mathrm{d}t \tag{3.41}$$

$$= \alpha(0 - (-1)) + \alpha(\exp(0) - \exp -\infty) \tag{3.42}$$

$$= 2\alpha, \tag{3.43}$$

so $\alpha = 1/2$.

b) We compute the cdf of $\tilde{t}$ conditioned on the event $\left\{\tilde{t} < 0\right\}$:

$$F_{\tilde{t}\,|\,\tilde{t}<0}(t) := \mathrm{P}(\tilde{t} \leq t \,|\, \tilde{t} < 0) \tag{3.44}$$

$$= \frac{\mathrm{P}(\tilde{t} \leq \min\{0, t\})}{\mathrm{P}(\tilde{t} < 0)} \tag{3.45}$$

$$= \begin{cases} 0 & \text{if } t < -1, \\ \frac{\int_{a=-1}^{t} f_{\tilde{t}}(a)\ \mathrm{d}a}{\int_{a=-1}^{0} f_{\tilde{t}}(a)\ \mathrm{d}a} & \text{if } -1 \leq t < 0, \\ 1 & \text{if } t \geq 0. \end{cases} \tag{3.46}$$

For $-1 \leq t < 0$,

$$\int_{a=-1}^{t} f_{\tilde{t}}(a)\ \mathrm{d}a = \frac{t+1}{2}, \tag{3.47}$$

so

$$F_{\tilde{t}\,|\,\tilde{t}<0}(t) = \begin{cases} 0 & \text{if } t < -1, \\ 1 + t & \text{if } -1 \leq t < 0, \\ 1 & \text{if } t \geq 0. \end{cases} \tag{3.48}$$

Differentiating, we obtain

$$f_{\tilde{t}\,|\,\tilde{t}<0}(t) = \begin{cases} 0 & \text{if } t < -1, \\ 1 & \text{if } -1 \leq t < 0, \\ 0 & \text{if } t \geq 0. \end{cases} \tag{3.49}$$

The conditional distribution is uniform between -1 and 0.

3.12 (Evaluating survival analysis)

a) Since $S_{\text{true}}(t) := \mathrm{P}\left(\tilde{t} > t\right) = 1 - F_{\tilde{t}}(t)$,

$$\tilde{w} := S_{\text{true}}(\tilde{t}) \tag{3.50}$$

$$= 1 - F_{\tilde{t}}\left(\tilde{t}\right). \tag{3.51}$$

By the probability integral transform $F_{\tilde{t}}\left(\tilde{t}\right)$ is a uniform random variable in $[0, 1]$. Consequently, the distribution of $\tilde{w}$ is also uniform in $[0, 1]$ by Exercise 3.10.

b) If S_{est} is a good approximation to S_{true}, then $S_{\text{est}}(\tilde{t})$ should be approximately uniformly distributed. We can evaluate whether this is the case by computing the transformed samples

$$w_i := S_{\text{est}}(t_i), \qquad 1 \leq i \leq n, \tag{3.52}$$

and checking whether their empirical cdf resembles the linear cdf of a uniform random variable in $[0, 1]$.

3.13 (Feeding data into its own empirical cdf) Since $x_1, x_2, \cdots, x_n$ are sorted,

$$y_i = F_X(x_i) \tag{3.53}$$

$$= \frac{1}{n}\sum_{j=1}^{n} 1\left(x_j \leq x_i\right) = \frac{i}{n}. \tag{3.54}$$

The answer does not depend on the dataset.

3.14 (Uniform temperature)

a) We denote the temperature by $\tilde{t}$. Its pdf is uniform between 60 and 80, so it equals 1/20. For $t \leq 60$ the cdf equals zero, because the probability that $\tilde{t} \leq 60$ is zero. For $t \geq 80$ the cdf equals one, because the probability that $\tilde{t} \leq 80$ is one. For $60 < t < 80$,

$$F_{\tilde{t}}(t) = \mathrm{P}\left(\tilde{t} \leq t\right) \tag{3.55}$$

$$= \int_{t=60}^{t} \frac{1}{20}\, \mathrm{d}t \tag{3.56}$$

$$= \frac{t-60}{20}. \tag{3.57}$$

The first quartile q_1 satisfies

$$F_{\tilde{t}}(q_1) = \frac{q_1 - 60}{20} = \frac{1}{4}, \tag{3.58}$$

so $q_1 = 65$.

b) We apply the inverse transform method. The inverse of the cdf satisfies

$$F_{\tilde{t}}(F_{\tilde{t}}^{-1}(u)) = \frac{F_{\tilde{t}}^{-1}(u) - 60}{20} \tag{3.59}$$

$$= u, \tag{3.60}$$

so $F_{\tilde{t}}^{-1}(u) = 60 + 20u$. We apply this inverse cdf to generate the samples. For u=0.1, the sample is $t = 62$. For $u = 0.8$ the sample is $t = 76$.

3.15 (Rounded-up measurements)

a) Let $\tilde{d}$ be the time the particle takes to decay, so $\tilde{r} = \lceil \tilde{d} \rceil$. Its pmf equals

$$\mathrm{P}(\tilde{r} = r) = \mathrm{P}\left(r - 1 \leq \tilde{d} < r\right) = \int_{r-1}^{r} \lambda\, e^{-\lambda x} \mathrm{d}x = e^{-\lambda(r-1)} - e^{-\lambda r} \tag{3.61}$$

$$= \left(e^{-\lambda}\right)^{r-1}\left(1 - e^{-\lambda}\right) \quad \text{for } r = 1, 2, 3, \ldots. \tag{3.62}$$

The reading is a geometric random variable with parameter $1 - e^{-\lambda}$.

b) Let $\tilde{x}$ be the error. For $x \leq 0$, $F_{\tilde{x}}(x) = 0$ since the error cannot be smaller than 0. It

also cannot be larger than 1, so for $x \geq 1$ $F_{\tilde{x}}(x) = 1$. For $0 \leq x \leq 1$,

$$\begin{aligned}
F_{\tilde{x}}(x) &= \mathrm{P}(\tilde{x} \leq x) && (3.63)\\
&= \mathrm{P}\left(\lceil \tilde{d} \rceil - \tilde{d} \leq x\right) && (3.64)\\
&= \mathrm{P}\left(\cup_{i=1}^{\infty} \left\{ i - x \leq \tilde{d} \leq i \right\}\right) \quad \text{union of disjoint events} && (3.65)\\
&= \sum_{i=1}^{\infty} \mathrm{P}\left(i - x \leq \tilde{d} \leq i\right) && (3.66)\\
&= \sum_{i=1}^{\infty} \lambda \int_{i-x}^{i} e^{-\lambda x} \mathrm{d}x && (3.67)\\
&= \sum_{i=1}^{\infty} e^{-\lambda(i-x)} - e^{-\lambda i} && (3.68)\\
&= \left(e^{\lambda x} - 1\right) \sum_{i=1}^{\infty} e^{-\lambda i} && (3.69)\\
&= \frac{e^{-\lambda}\left(e^{\lambda x} - 1\right)}{1 - e^{-\lambda}} = \frac{e^{\lambda x} - 1}{e^{\lambda} - 1}. && (3.70)
\end{aligned}$$

Differentiating we obtain

$$f_{\tilde{x}}(x) = \begin{cases} \frac{\lambda e^{\lambda x}}{e^{\lambda} - 1} & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.71}$$

3.16 (Half life)

a) We have

$$\begin{aligned}
\mathrm{P}(\tilde{t} > t_{1/2}) &= \int_{t_{1/2}}^{\infty} \lambda \exp(-\lambda x)\, \mathrm{d}x && (3.72)\\
&= \exp\left(-\lambda t_{1/2}\right). && (3.73)
\end{aligned}$$

Setting equal to 1/2 yields $t_{1/2} = \frac{\ln 2}{\lambda}$. This is the time that it takes for the particle to decay with probability 1/2. Consequently, from a large group of particles with this distribution, about half of them will decay after that time, so this is a reasonable definition of half life.

b) We have

$$\begin{aligned}
\mathrm{P}(t_{1/2} < \tilde{t} < t) &= \int_{t_{1/2}}^{t} \lambda \exp(-\lambda x)\, \mathrm{d}x && (3.74)\\
&= \exp\left(-\lambda t_{1/2}\right) - \exp(-\lambda t) && (3.75)\\
&= \frac{1}{2} - \exp(-\lambda t). && (3.76)
\end{aligned}$$

Setting equal to 1/4 yields $t_{1/4} = \frac{\ln 4}{\lambda} = \frac{2 \ln 2}{\lambda} = 2t_{1/2}$. Intuitively, this means that after two half-lifes, one fourth of the particles remain. This makes sense: the number of particles is halved after the first half-time, and then halved again after the second.

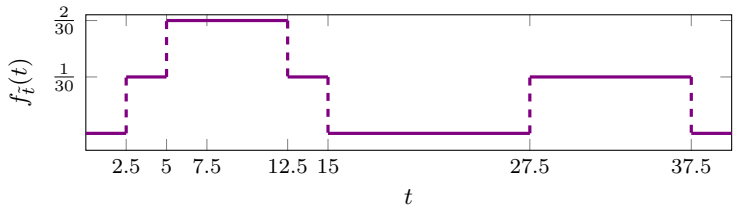
c)

$$\begin{aligned} \mathrm{P}(\tilde{t} > kt_{1/2}) &= \int_{kt_{1/2}}^{\infty} \lambda \exp\left(-\lambda x\right) \, \mathrm{d}x && (3.77)\\ &= \exp\left(-\lambda k t_{1/2}\right) && (3.78)\\ &= \exp\left(-\ln 2^k\right) && (3.79)\\ &= \frac{1}{2^k}. && (3.80) \end{aligned}$$

This says that if we wait for k half times, the remaining particles are halved k times, which is again consistent with the intuitive definition of half time.

3.17 (Earthquake)

a) The KDE estimate looks like this:



The probability is equal to

$$\begin{aligned} \mathrm{P}(\tilde{t} > 10) &= \frac{2 \cdot 2.5}{30} + \frac{2.5}{30} + \frac{10}{30} && (3.81)\\ &= \frac{17.5}{30} && (3.82)\\ &= 0.583, && (3.83) \end{aligned}$$

which is the area under the pdf to the right of 10.

b) From the notes, the ML estimate for the parameter of the exponential is

$$\begin{aligned} \lambda_{\mathrm{ML}} &= \frac{3}{7.5 + 10 + 32.5} && (3.84)\\ &= \frac{3}{50}. && (3.85) \end{aligned}$$

The probability equals

$$\begin{aligned} \mathrm{P}(\tilde{t} > 10) &= \int_{10}^{\infty} f_{\tilde{t}}(t) \, \mathrm{d}t && (3.86)\\ &= \int_{10}^{\infty} \frac{3\exp(-3t/50)}{50} \, \mathrm{d}t && (3.87)\\ &= -\exp(-3t/50)]_{10}^{\infty} && (3.88)\\ &= 0.549. && (3.89) \end{aligned}$$

c) The parametric method requires less data but makes a stronger assumption about the distribution. The nonparametric method is more flexible but requires more data.

3.18 (Uniform distribution with a bump)

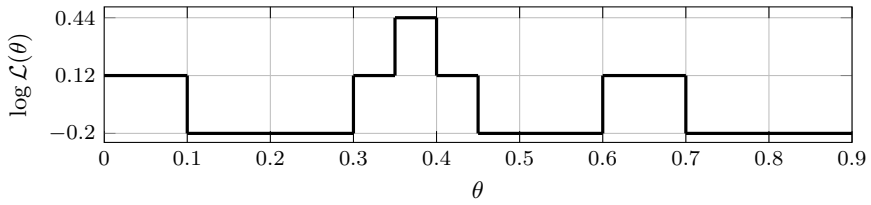
a)

$$F_{\tilde{a}}(a) = \begin{cases} 0 & \text{if } a \leq 0, \\ 0.9a & \text{if } 0 \leq a \leq \theta, \\ 0.9\theta + 1.9(a-\theta) = 1.9a - \theta & \text{if } \theta \leq a \leq \theta + 0.1, \\ 0.9\theta + 0.19 + 0.9(a-\theta-0.1) = 0.9a + 0.1 & \text{if } \theta \leq a \leq \theta + 0.1, \\ 1 & \text{if } a \geq 1. \end{cases} \tag{3.90}$$

b)

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{4} \log f_{\tilde{a}}(x_i) = \begin{cases} 3\log 0.9 + \log 1.9 = 0.12 & \text{if } 0 \leq \theta \leq 0.1, \\ 4\log 0.9 = -0.2 & \text{if } 0.1 < \theta < 0.3, \\ 0.12 & \text{if } 0.3 \leq \theta < 0.35, \\ 2\log 0.9 + 2\log 1.9 = 0.44 & \text{if } 0.35 \leq \theta \leq 0.4, \\ 0.12 & \text{if } 0.4 < \theta \leq 0.45, \\ -0.2 & \text{if } 0.45 < \theta < 0.6, \\ 0.12 & \text{if } 0.6 \leq \theta \leq 0.7, \\ -0.2 & \text{if } 0.7 < \theta \leq 0.9. \end{cases} \tag{3.91}$$

The log-likelihood function looks like this:



3.19 (Planet)

a) Since

$$F_{\tilde{t}}(t) = \int_{-\infty}^{t} f_{\tilde{t}}(u)\ \mathrm{d}u, \tag{3.92}$$

for $t < 0$,

$$F_{\tilde{t}}(t) = \int_{-\infty}^{t} \frac{\lambda \exp(\lambda u)}{2}\ \mathrm{d}u \tag{3.93}$$

$$= \frac{\exp(\lambda t)}{2}, \tag{3.94}$$

and for $t \geq 0$,

$$F_{\tilde{t}}(t) = \int_{-\infty}^{0} \frac{\lambda \exp(\lambda u)}{2}\ \mathrm{d}u + \int_{0}^{t} \frac{\lambda \exp(-\lambda u)}{2}\ \mathrm{d}u \tag{3.95}$$

$$= \frac{1}{2} + \frac{1 - \exp(-\lambda t)}{2} \tag{3.96}$$

$$= 1 - \frac{\exp(-\lambda t)}{2}. \tag{3.97}$$

b) The log-likelihood is

$$\log \mathcal{L}_X(\lambda) = \sum_{i=1}^{n} \log f_\lambda(x_i) \tag{3.98}$$

$$= \sum_{i=1}^{n} \log \frac{\lambda \exp(-\lambda |x_i|)}{2} \tag{3.99}$$

$$= n \log \lambda - n \log 2 - \lambda \sum_{i=1}^{n} |x_i| \,. \tag{3.100}$$

The derivative and second derivative of the log-likelihood function are

$$\frac{\mathrm{d} \log \mathcal{L}_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} |x_i| \,, \tag{3.101}$$

$$\frac{\mathrm{d}^2 \log \mathcal{L}_{x_1,\ldots,x_n}(\lambda)}{\mathrm{d}\lambda^2} = -\frac{n}{\lambda^2} < 0 \quad \text{for all } \lambda > 0. \tag{3.102}$$

The function is concave, as the second derivative is negative, so there cannot be different local maxima. The maximum is obtained by setting the first derivative to zero, which yields

$$\lambda_{\mathrm{ML}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} |x_i|} \tag{3.103}$$

$$= \frac{1}{39}. \tag{3.104}$$

c) The conditional cdf of $\tilde{t}$ given $\tilde{t} \geq 0$ evaluated at $t > 0$ is

$$F_{\tilde{t}\,|\,\tilde{t}>0}(t) = \mathrm{P}\left(\tilde{t} \leq t \,|\, \tilde{t} > 0\right) \tag{3.105}$$

$$= \frac{\mathrm{P}\left(0 < \tilde{t} \leq t\right)}{\mathrm{P}\left(\tilde{t} > 0\right)} \tag{3.106}$$

$$= \frac{F_{\tilde{t}}(t) - F_{\tilde{0}}}{1 - F_{\tilde{t}}(0)} \tag{3.107}$$

$$= \frac{1 - \frac{\exp(-\lambda t)}{2} - \frac{1}{2}}{\frac{1}{2}} \tag{3.108}$$

$$= 1 - \exp(-\lambda t). \tag{3.109}$$

Differentiating with respect to t yields an exponential pdf $f_{\tilde{t}|\tilde{t}>t_0}(t) = \lambda e^{-\lambda t}$ with parameter λ.

3.20 (Triangular pdf)

a) The possible values of w are $w \geq \max(x_1, \ldots, x_n) = 1.5$.

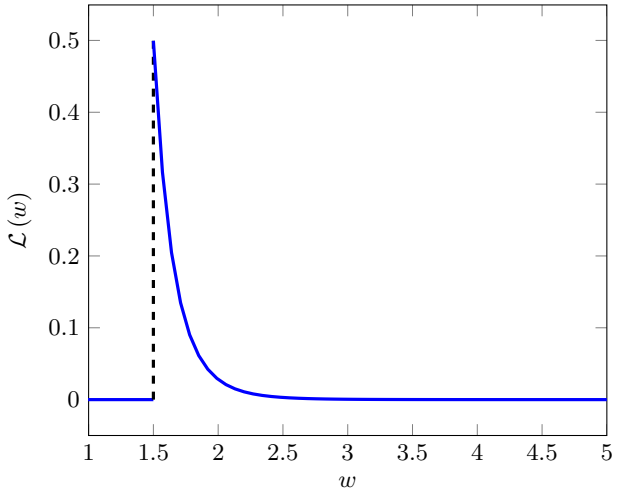
b) The likelihood equals

$$\mathcal{L}(w) = \prod_{i}^{n} f_w(x_i) \tag{3.110}$$

$$= \frac{2^5}{w^{10}} \prod_{i}^{n} x_i \tag{3.111}$$

$$= \frac{28.8}{w^{10}} \tag{3.112}$$

if $w \geq \max(x_1, \ldots, x_n)$ and 0 otherwise because in that case $f_w(\max(x_1, \ldots, x_n)) = 0$.



c) The maximum likelihood estimate is 1.5 because the likelihood is decreasing over $[1.5, \infty)$.

d) The ML estimate systematically underestimates the true parameter. It is equal to the maximum value in the data, which has to be smaller than w_{true}.

e) We apply inverse-transform sampling. The cdf equals

$$F_w(x) = \begin{cases} 0, & x \leq 0, \\ \frac{x^2}{w^2}, & \text{for } 0 \leq x \leq w, \\ 1, & \text{for } x \geq w. \end{cases} \tag{3.113}$$

The inverse of the cdf in the interval of interest is $F_w^{-1}(y) = w\sqrt{y}$. The sample is therefore $2\sqrt{0.64} = 1.6$.

3.21 (Rat)

a) The pdf needs to be nonnegative, which requires $0 \leq \alpha \leq 1$. We also need the pdf to integrate to one, which it does if α is in that range:

$$\int_{a=0}^{1} f_{\tilde{a}}(a)\, \mathrm{d}a = \int_{a=0}^{0.5} 2\alpha \, \mathrm{d}a + \int_{a=0.5}^{1} 2(1-\alpha) \, \mathrm{d}a \tag{3.114}$$

$$= 2\alpha \cdot 0.5 + 2(1-\alpha) \cdot 0.5 \tag{3.115}$$

$$= 1. \tag{3.116}$$

b) Expressing the pdf as a function of α the likelihood of each data point is equal to 2α if the point is between 0 and 0.5, and to $2(1-\alpha)$ if it is between 0.5 and 1. Let n be the number of data, $n_{[0,0.5]}$ the number of data between 0 and 0.5, and $n_{(0.5,1]}$ the number of points between 0.5 and 1. We have,

$$\log \mathcal{L}(X) = \sum_{i=1}^{n} \log f_\alpha(x_i) \tag{3.117}$$

$$= n_{[0,0.5]} \log 2\alpha + n_{(0.5,1]} \log(2(1-\alpha)). \tag{3.118}$$

The first and second derivatives of the log likelihood equal

$$(\log \mathcal{L}(X))' = \frac{n_{[0,0.5]}}{\alpha} - \frac{n_{(0.5,1]}}{1-\alpha}, \tag{3.119}$$

$$(\log \mathcal{L}(X))'' = -\frac{n_{[0,0.5]}}{\alpha^2} - \frac{n_{(0.5,1]}}{(1-\alpha)^2}. \tag{3.120}$$

The function is concave, so we can set the first derivative to zero to find the ML estimate, it equals

$$\alpha_{\mathrm{ML}} = \frac{n_{[0,0.5]}}{n} \tag{3.121}$$

$$= \frac{2}{5}. \tag{3.122}$$

c) The probability of the rat being in the first half is

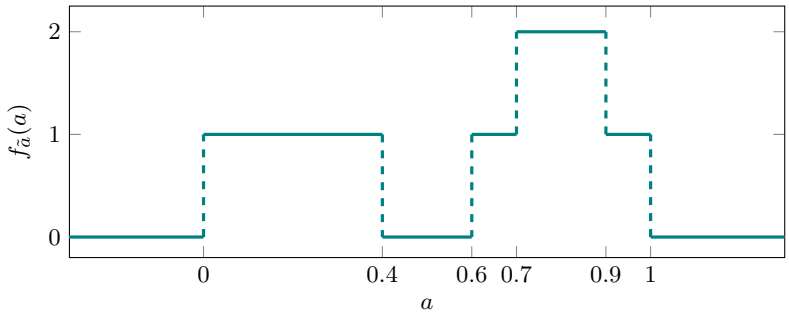
$$\mathrm{P}(0 \leq \tilde{a} \leq 0.5) = \int_{a=0}^{0.5} 2\alpha \, \mathrm{d}a \tag{3.123}$$

$$= \alpha. \tag{3.124}$$

Estimating this probability using empirical probabilities yields exactly the same estimate for α as the ML estimate,

$$\hat{\alpha_{\mathrm{ML}}} = \frac{n_{[0,0.5]}}{n}. \tag{3.125}$$

d) The estimated pdf is:



What is problematic is that the density is zero between 0.4 and 0.6, which seems an artifact of the limited number of data. This can be alleviated by increasing the width of the rectangular kernel.

4

Multiple Discrete Variables

Exercises

4.1 (Halloween)

a) Lou grabs $\tilde{l}$ chocolate bars, where $\tilde{l}$ is a random variable with values between 0 and 2. Since all three values have the same probability, $p_{\tilde{l}}(l) = \frac{1}{3}$, for $l \in \{0, 1, 2\}$ and zero otherwise. Ellie grabs $\tilde{e}$ chocolate bars, where $\tilde{e}$ is a random variable with values between 0 and $2 - \tilde{l}$. Since they have the same probability, the conditional pmf equals

$$p_{\tilde{e}\,|\,\tilde{l}}(a\,|\,l) = \begin{cases} \frac{1}{3-l} & \text{if } 0 \leq a \leq 2-l, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

By the chain rule, the joint pmf is:

$$p_{\tilde{e},\tilde{l}}(a, l) = \frac{1}{3(3-l)} \text{ for } l = 0, .., 2 \text{ and } a = 0, .., 2-l. \tag{4.2}$$

In the (a, l) plane, the points on which the joint pmf is defined form a triangle.

b) To find the marginal pmf, we sum over l. For each value of a the sum has a different number of terms:

$$p_{\tilde{e}}(0) = \frac{1}{3}\left(\frac{1}{3} + \frac{1}{2} + \frac{1}{1}\right) = \frac{11}{18}, \tag{4.3}$$

$$p_{\tilde{e}}(1) = \frac{1}{3}\left(\frac{1}{3} + \frac{1}{2}\right) = \frac{5}{18}, \tag{4.4}$$

$$p_{\tilde{e}}(2) = \frac{1}{3}\frac{1}{3} = \frac{1}{9}. \tag{4.5}$$

c) When $\tilde{e} = 1$, $\tilde{l}$ must be either 0 or 1. By the definition of conditional pmf,

$$p_{\tilde{l}\,|\,\tilde{e}}(0\,|\,1) = \frac{p_{\tilde{e},\tilde{l}}(1, 0)}{p_{\tilde{e}}(1)} \tag{4.6}$$

$$= \frac{\frac{1}{3}\frac{1}{3}}{\frac{5}{18}} = \frac{2}{5}, \tag{4.7}$$

$$p_{\tilde{l}\,|\,\tilde{e}}(1\,|\,1) = \frac{p_{\tilde{e},\tilde{l}}(1, 1)}{p_{\tilde{e}}(0)} \tag{4.8}$$

$$= \frac{\frac{1}{3}\frac{1}{2}}{\frac{5}{18}} = \frac{3}{5}. \tag{4.9}$$

4.2 (Empirical conditional pmf) The number of data points in X equal to a is n_a, so

$$p_X(a) := \frac{1}{n}\sum_{i=1}^{n} 1\left(x_i = a\right) \tag{4.10}$$

$$= \frac{n_a}{n}. \tag{4.11}$$

We can write the number of data point pairs such that the first entry is a and the second entry is b in the two following ways,

$$\sum_{\{y \in Y_a\}} 1\left(y = b\right) = \sum_{i=1}^{n} 1\left(x_i = a, y_i = b\right). \tag{4.12}$$

Consequently,

$$p_{Y\,|\,X}(b\,|\,a) := \frac{1}{n_a}\sum_{\{y \in Y_a\}} 1\left(y = b\right) \tag{4.13}$$

$$= \frac{1}{np_X(a)}\sum_{i=1}^{n} 1\left(x_i = a, y_i = b\right) \tag{4.14}$$

$$= \frac{p_{X,Y}(a,b)}{p_X(a)}. \tag{4.15}$$

4.3 (Noisy data)

a)

$$\mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{x} = 1\right) \tag{4.16}$$

$$= \frac{\mathrm{P}\left(\tilde{z} = 1, \tilde{x} = 1\right)}{\mathrm{P}\left(\tilde{x} = 1\right)} \tag{4.17}$$

$$= \frac{\mathrm{P}\left(\tilde{z} = 1, \tilde{x} = 1, \tilde{y} = 0\right) + \mathrm{P}\left(\tilde{z} = 1, \tilde{x} = 1, \tilde{y} = 1\right)}{\mathrm{P}\left(\tilde{x} = 1\right)} \tag{4.18}$$

$$= \frac{\mathrm{P}\left(\tilde{x} = 1\right)\mathrm{P}\left(\tilde{y} = 0\,|\,\tilde{x} = 1\right)\mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{x} = 1, \tilde{y} = 0\right) + \mathrm{P}\left(\tilde{x} = 1\right)\mathrm{P}\left(\tilde{y} = 1\,|\,\tilde{x} = 1\right)\mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{x} = 1, \tilde{y} = 1\right)}{\mathrm{P}\left(\tilde{x} = 1\right)}$$

$$= \mathrm{P}\left(\tilde{y} = 0\,|\,\tilde{x} = 1\right)\mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{y} = 0\right) + \mathrm{P}\left(\tilde{y} = 1\,|\,\tilde{x} = 1\right)\mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{x} = 1, \tilde{y} = 1\right) \tag{4.19}$$

$$= 0.1 \cdot 0.1 + 0.9 \cdot 0.9 \tag{4.20}$$

$$= 0.82. \tag{4.21}$$

b) $\mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{x} = 1, \tilde{y} = 1\right) = \mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{y} = 1\right) = 0.9 \neq \mathrm{P}\left(\tilde{z} = 1\,|\,\tilde{x} = 1\right)$, so $\tilde{y}$ and $\tilde{z}$ are not conditionally independent given $\tilde{x}$.

c) By definition of the problem $\mathrm{P}\left(\tilde{z} = z\,|\,\tilde{x} = x, \tilde{y} = y\right) = \mathrm{P}\left(\tilde{z} = z\,|\,\tilde{y} = y\right)$ for any values of x, y and z, so $\tilde{z}$ and $\tilde{x}$ are conditionally independent given $\tilde{y}$.

4.4 (Fire alarm)

a)

$$p_{\tilde{a}}(1) = \sum_{x\in\{0,1\}} \sum_{b\in\{0,1\}} p_{\tilde{a},\tilde{x},\tilde{b}}(1,x,b) \tag{4.22}$$

$$= p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,0) + p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,1) + p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,0) + p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,1) \tag{4.23}$$

$$= p_{\tilde{x}}(0)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,0) + p_{\tilde{x}}(0)p_{\tilde{b}}(1)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,1) \tag{4.24}$$

$$+ p_{\tilde{x}}(1)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,1,0) + p_{\tilde{x}}(1)p_{\tilde{b}}(1)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,1,1) \tag{4.25}$$

$$= 0 + 0.8 \cdot 0.5 \cdot 0.2 + 0.2 \cdot 0.5 \cdot 0.8 + 0.2 \cdot 0.5 \cdot 0.4 \tag{4.26}$$

$$= 0.2. \tag{4.27}$$

b)

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(0,0\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,0)}{p_{\tilde{a}}(1)} \tag{4.28}$$

$$= \frac{p_{\tilde{x}}(0)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,0)}{p_{\tilde{a}}(1)} \tag{4.29}$$

$$= 0, \tag{4.30}$$

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(0,1\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,0,1)}{p_{\tilde{a}}(1)} \tag{4.31}$$

$$= \frac{p_{\tilde{x}}(0)p_{\tilde{b}}(1)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,0,1)}{p_{\tilde{a}}(1)} \tag{4.32}$$

$$= \frac{0.8 \cdot 0.5 \cdot 0.2}{0.2} \tag{4.33}$$

$$= 0.4, \tag{4.34}$$

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(1,0\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,0)}{p_{\tilde{a}}(1)} \tag{4.35}$$

$$= \frac{p_{\tilde{x}}(1)p_{\tilde{b}}(0)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,1,0)}{p_{\tilde{a}}(1)} \tag{4.36}$$

$$= \frac{0.2 \cdot 0.5 \cdot 0.8}{0.2} \tag{4.37}$$

$$= 0.4, \tag{4.38}$$

$$p_{\tilde{x},\tilde{b}\,|\,\tilde{a}}(1,1\,|\,1) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1,1,1)}{p_{\tilde{a}}(1)} \tag{4.39}$$

$$= \frac{p_{\tilde{x}}(1)p_{\tilde{b}}(1)p_{\tilde{a}\,|\,\tilde{x},\tilde{b}}(1\,|\,1,1)}{p_{\tilde{a}}(1)} \tag{4.40}$$

$$= \frac{0.2 \cdot 0.5 \cdot 0.4}{0.2} \tag{4.41}$$

$$= 0.2. \tag{4.42}$$

c) We compare $p_{\tilde{x} \mid \tilde{a},\tilde{b}}(0 \mid 1, 0)$ and $p_{\tilde{x} \mid \tilde{a}}(0 \mid 1)$,

$$p_{\tilde{x} \mid \tilde{a},\tilde{b}}(0 \mid 1, 0) = \frac{p_{\tilde{a},\tilde{x},\tilde{b}}(1, 0, 0)}{p_{\tilde{a},\tilde{b}}(1, 0)} \tag{4.43}$$

$$= \frac{p_{\tilde{x}}(0) p_{\tilde{b}}(0) p_{\tilde{a} \mid \tilde{x},\tilde{b}}(1 \mid 0, 0)}{\sum_{x=0}^{1} p_{\tilde{x}}(x) p_{\tilde{b}}(0) p_{\tilde{a} \mid \tilde{x},\tilde{b}}(1 \mid x, 0)} \tag{4.44}$$

$$= 0, \tag{4.45}$$

$$p_{\tilde{x} \mid \tilde{a}}(0 \mid 1) = \frac{p_{\tilde{a},\tilde{x}}(1, 0)}{p_{\tilde{a}}(1)} \tag{4.46}$$

$$= \frac{\sum_{b=0}^{1} p_{\tilde{a},\tilde{x},\tilde{b}}(1, 0, b)}{p_{\tilde{a}}(1)} \tag{4.47}$$

$$= \frac{\sum_{b=0}^{1} p_{\tilde{x}}(0) p_{\tilde{b}}(b) p_{\tilde{a} \mid \tilde{x},\tilde{b}}(1 \mid 0, b)}{p_{\tilde{a}}(1)} \tag{4.48}$$

$$= \frac{0.8 \cdot 0.5 \cdot 0.2}{0.2} = 0.4 \neq 0, \tag{4.49}$$

which implies that $\tilde{x}$ and $\tilde{b}$ are not conditionally independent given $\tilde{a}$. Intuitively, if we know that the alarm is ringing, then the state of the battery does provide information about fire. For example, if the battery is fine, then the alarm is not ringing for that reason, so it is more likely that there is a fire.

4.5 (Volcano eruption)

a)

$$p_{\tilde{v} \mid \tilde{s}_1,\tilde{s}_2,\tilde{s}_3}(1 \mid 1, 1, 1) \tag{4.50}$$

$$= \frac{p_{\tilde{v},\tilde{s}_1,\tilde{s}_2,\tilde{s}_3}(1, 1, 1, 1)}{p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3}(1, 1, 1, 1)} \tag{4.51}$$

$$= \frac{p_{\tilde{v},\tilde{s}_1,\tilde{s}_2,\tilde{s}_3}(1, 1, 1, 1)}{p_{\tilde{v},\tilde{s}_1,\tilde{s}_2,\tilde{s}_3}(1, 1, 1, 1) + p_{\tilde{v},\tilde{s}_1,\tilde{s}_2,\tilde{s}_3}(0, 1, 1, 1)} \tag{4.52}$$

$$= \frac{p_{\tilde{v}}(1)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 1) p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 1) p_{\tilde{s}_3 \mid \tilde{v}}(1 \mid 1)}{p_{\tilde{v}}(1)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 1) p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 1) p_{\tilde{s}_3 \mid \tilde{v}}(1 \mid 1) + p_{\tilde{v}}(0)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 0) p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 0) p_{\tilde{s}_3 \mid \tilde{v}}(1 \mid 0)}$$

$$= \frac{0.1 \cdot 0.5 \cdot 0.8 \cdot 0.9}{0.1 \cdot 0.5 \cdot 0.8 \cdot 0.9 + 0.9 \cdot 0.2 \cdot 0.2 \cdot 0.2} \tag{4.53}$$

$$= \frac{5}{5+1} = \frac{5}{6}. \tag{4.54}$$

b) We compare the marginal probability that the first sensor is activated,

$$p_{\tilde{s}_1}(1) = p_{\tilde{v},\tilde{s}_1}(1, 1) + p_{\tilde{v},\tilde{s}_1}(0, 1) \tag{4.55}$$

$$= p_{\tilde{v}}(1)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 1) + p_{\tilde{v}}(0)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 0) \tag{4.56}$$

$$= 0.9 \cdot 0.2 + 0.1 \cdot 0.5 = 0.23, \tag{4.57}$$

and the conditional probability that the first sensor is activated given that the second

is activated,

$$p_{\tilde{s}_1 \mid \tilde{s}_2}(1 \mid 1) = \frac{p_{\tilde{s}_1, \tilde{s}_2}(1,1)}{p_{\tilde{s}_2}(1)} \tag{4.58}$$

$$= \frac{p_{\tilde{v}}(1)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 1) p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 1) + p_{\tilde{v}}(0)\, p_{\tilde{s}_1 \mid \tilde{v}}(1 \mid 0) p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 0)}{p_{\tilde{v}}(1)\, p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 1) + p_{\tilde{v}}(0)\, p_{\tilde{s}_2 \mid \tilde{v}}(1 \mid 0)} \tag{4.59}$$

$$= \frac{0.9 \cdot 0.2 \cdot 0.2 + 0.1 \cdot 0.5 \cdot 0.8}{0.9 \cdot 0.2 + 0.1 \cdot 0.8} \tag{4.60}$$

$$= 0.29. \tag{4.61}$$

Since they are different, the two random variables are not independent. Intuitively, if the second sensor is activated then an eruption is more likely, which makes it more likely for the first sensor to be activated. Therefore the state of the second sensor provides information about the state of the first sensor.

4.6 (Flu and COVID-19)

a)

$$p_{\tilde{t}}(1) = \sum_{c \in \{0,1\}} \sum_{x \in \{0,1\}} \sum_{s \in \{0,1\}} p_{\tilde{s}, \tilde{t}, \tilde{c}, \tilde{x}}(s, 1, c, x) \tag{4.62}$$

$$= \sum_{c \in \{0,1\}} \sum_{x \in \{0,1\}} \sum_{s \in \{0,1\}} p_{\tilde{c}}(c) p_{\tilde{x}}(x) p_{\tilde{s}, \tilde{t} \mid \tilde{c}, \tilde{x}}(s, 1 \mid c, x) \tag{4.63}$$

$$= 0.2 \cdot 0.9(0.2 + 0.4) + 0.8 \cdot 0.1(0.4 + 0.4) + 0.2 \cdot 0.1 \tag{4.64}$$

$$= 0.192. \tag{4.65}$$

b)

$$p_{\tilde{c} \mid \tilde{t}}(1 \mid 1) = \frac{p_{\tilde{c}, \tilde{t}}(1,1)}{p_{\tilde{t}}(1)} \tag{4.66}$$

$$= \frac{\sum_{x \in \{0,1\}} \sum_{s \in \{0,1\}} p_{\tilde{s}, \tilde{t}, \tilde{c}, \tilde{x}}(s, 1, 1, x)}{p_{\tilde{t}}(1)} \tag{4.67}$$

$$= \frac{\sum_{x \in \{0,1\}} \sum_{s \in \{0,1\}} p_{\tilde{c}}(1) p_{\tilde{x}}(x) p_{\tilde{s}, \tilde{t} \mid \tilde{c}, \tilde{x}}(s, 1 \mid 1, x)}{p_{\tilde{t}}(1)} \tag{4.68}$$

$$= \frac{0.2 \cdot 0.9\,(0.2 + 0.4) + 0.2 \cdot 0.1}{0.192} \tag{4.69}$$

$$= 0.67. \tag{4.70}$$

4.7 (Vibrations)

a) By conditional independence of $\tilde{s}$ and $\tilde{e}$ given $\tilde{v}$

$$p_{\tilde{s}}(1) = \sum_{e=0}^{1} \sum_{v=0}^{2} p_{\tilde{e}, \tilde{v}}(e, v) p_{\tilde{s} \mid \tilde{e}, \tilde{v}}(1 \mid e, v) \tag{4.71}$$

$$= \sum_{e=0}^{1} \sum_{v=0}^{2} p_{\tilde{e}, \tilde{v}}(e, v) p_{\tilde{s} \mid \tilde{v}}(1 \mid v) \tag{4.72}$$

$$= p_{\tilde{s} \mid \tilde{v}}(1 \mid 1)(p_{\tilde{e}, \tilde{v}}(0,1) + p_{\tilde{e}, \tilde{v}}(1,1)) + p_{\tilde{s} \mid \tilde{v}}(1 \mid 2)(p_{\tilde{e}, \tilde{v}}(0,2) + p_{\tilde{e}, \tilde{v}}(1,2)) \tag{4.73}$$

$$= 0.5(0.05 + 0.05) + 0.1 \tag{4.74}$$

$$= 0.15 \tag{4.75}$$

and consequently $p_{\tilde{s}}(0) = 1 - p_{\tilde{s}}(1) = 0.85$.

b)

$$p_{\tilde{e}\,|\,\tilde{s}}(1\,|\,1) = \frac{p_{\tilde{e},\tilde{s}}(1,1)}{p_{\tilde{s}}(1)} \tag{4.76}$$

$$= \frac{\sum_{v=0}^{2} p_{\tilde{e},\tilde{v},\tilde{s}}(1,v,1)}{p_{\tilde{s}}(1)} \tag{4.77}$$

$$= \frac{\sum_{v=0}^{2} p_{\tilde{e},\tilde{v},\tilde{s}}(1,v)p_{\tilde{s}\,|\,\tilde{v}}(1\,|\,v)}{p_{\tilde{s}}(1)} \tag{4.78}$$

$$= \frac{p_{\tilde{s}\,|\,\tilde{v}}(1\,|\,1)p_{\tilde{e},\tilde{v}}(1,1) + p_{\tilde{s}\,|\,\tilde{v}}(1\,|\,2)p_{\tilde{e},\tilde{v}}(1,2)}{p_{\tilde{s}}(1)} \tag{4.79}$$

$$= \frac{0.5 \cdot 0.05 + 0.1}{0.15} \tag{4.80}$$

$$= 0.833. \tag{4.81}$$

c) We have

$$p_{\tilde{e}}(1) = \sum_{v=0}^{2} p_{\tilde{e},\tilde{v}}(1,v) \tag{4.82}$$

$$= 0.05 + 0.1 \tag{4.83}$$

$$= 0.15 \neq p_{\tilde{e}\,|\,\tilde{s}}(1\,|\,1), \tag{4.84}$$

so they are not independent. This makes sense, because the sensor reading is more likely to be one if there are vibrations, which is more likely if there is an earthquake.

4.8 (Footprints)

a) The joint pmf $p_{\tilde{a},\tilde{s},\tilde{h}}(a,s,h)$ is obtained by dividing the counts by the total number of data (200):

		s		
h	a = Coyote	Small	Medium	Large
	Oval	0.4	0.15	0.025
	Circular	0.15	0.025	0

	s		
a = Wolf	Small	Medium	Large
Oval	0	0.025	0.05
Circular	0.025	0.075	0.075

b)

$$p_{\tilde{a}\,|\,\tilde{s},\tilde{h}}\,(\text{Wolf}\,|\,\text{Med, Circular}) = \frac{p_{\tilde{a},\tilde{s},\tilde{h}}\,(\text{Wolf, Med, Circular})}{p_{\tilde{s},\tilde{h}}\,(\text{Med, Circular})} \tag{4.85}$$

$$= \frac{p_{\tilde{a},\tilde{s},\tilde{h}}\,(\text{Wolf, Med, Circular})}{p_{\tilde{a},\tilde{s},\tilde{h}}\,(\text{Wolf, Med, Circular}) + p_{\tilde{a},\tilde{s},\tilde{h}}\,(\text{Coyote, Med, Circular})} \tag{4.86}$$

$$= \frac{0.075}{0.075 + 0.025} = 0.75. \tag{4.87}$$

4.9 (Election)

a) Candidate A wins if they win 2 or 3 states. Consequently, under the independence assumptions,

$$p_{\tilde{o}}(1) = \mathrm{P}(\text{A wins all states}) + \sum_{i=1}^{3} \mathrm{P}(\text{A wins all states except } i) \tag{4.88}$$

$$= 0.6^3 + 3 \cdot 0.4 \cdot 0.6^2 \tag{4.89}$$

$$= 0.648. \tag{4.90}$$

b) By the definition of conditional pmf,

$$p_{\tilde{s}_2 \,|\, \tilde{o}}\,(0 \,|\, 0) = \frac{p_{\tilde{s}_2,\tilde{o}}\,(0,0)}{p_{\tilde{o}}(0)} \tag{4.91}$$

$$= \frac{\sum_{s_1=0}^{1} \sum_{s_3=0}^{1} p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}\,(s_1,0,s_3,0)}{p_{\tilde{o}}(0)} \tag{4.92}$$

$$= \frac{p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}\,(0,0,0,0) + p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}\,(1,0,0,0) + p_{\tilde{s}_1,\tilde{s}_2,\tilde{s}_3,\tilde{o}}\,(0,0,1,0)}{p_{\tilde{o}}(0)} \tag{4.93}$$

$$= \frac{p_{\tilde{s}_1}(0)p_{\tilde{s}_2}(0)p_{\tilde{s}_3}(0) + p_{\tilde{s}_1}(1)p_{\tilde{s}_2}(0)p_{\tilde{s}_3}(0) + p_{\tilde{s}_1}(0)p_{\tilde{s}_2}(0)p_{\tilde{s}_3}(1)}{p_{\tilde{o}}(0)} \tag{4.94}$$

$$= \frac{0.4^3 + 2 \cdot 0.4^2 \cdot 0.6}{1 - 0.648} \tag{4.95}$$

$$= 0.727. \tag{4.96}$$

c) We have

$$p_{\tilde{s}_1 \,|\, \tilde{o}}(1 \,|\, 1) \tag{4.97}$$

$$= \frac{p_{\tilde{s}_1,\tilde{o}}(1,1)}{p_{\tilde{o}}(1)} \tag{4.98}$$

$$= \frac{\mathrm{P}(\text{A wins all states}) + \mathrm{P}(\text{A wins 1 and 2 but not 3}) + \mathrm{P}(\text{A wins 1 and 3 but not 2})}{p_{\tilde{o}}(1)}$$

$$= \frac{0.6^3 + 2 \cdot 0.4 \cdot 0.6^2}{0.648} \tag{4.99}$$

$$= 0.778. \tag{4.100}$$

However, $p_{\tilde{s}_1 \,|\, \tilde{o},\tilde{s}_2}(1 \,|\, 1, 0) = 1$ because if A wins the election then they cannot lose state 1 and state 2. Intuitively, even if the state results are independent, they both determine the result of the election, so revealing who wins the election *connects* them. For example, if candidate A has won the election but lost state 2, then this completely determines the result of state 1.

4.10 (Interview)

a) By the independence assumptions

$$p_{\tilde{d}_1,\tilde{d}_2\,|\,\tilde{q}}(1,1\,|\,1) = \frac{\mathrm{P}(\tilde{e}_1\tilde{q}=1,\tilde{e}_2\tilde{q}=1,\tilde{q}=1)}{\mathrm{P}(\tilde{q}=1)} \tag{4.101}$$

$$= \frac{\mathrm{P}(\tilde{q}=1)\mathrm{P}(\tilde{e}_1\tilde{q}=1,\tilde{e}_2\tilde{q}=1\,|\,\tilde{q}=1)}{\mathrm{P}(\tilde{q}=1)} \tag{4.102}$$

$$= \mathrm{P}(\tilde{e}_1\tilde{q}=1,\tilde{e}_2\tilde{q}=1\,|\,\tilde{q}=1) \tag{4.103}$$

$$= \mathrm{P}(\tilde{e}_1=1,\tilde{e}_2=1\,|\,\tilde{q}=1) \tag{4.104}$$

$$= \mathrm{P}(\tilde{e}_1=1,\tilde{e}_2=1) \tag{4.105}$$

$$= \mathrm{P}(\tilde{e}_1=1)\mathrm{P}(\tilde{e}_2=1) \tag{4.106}$$

$$= 0.64. \tag{4.107}$$

b)

$$p_{\tilde{q}\,|\,\tilde{d}_1,\tilde{d}_2}(1\,|\,1,1) = \frac{p_{\tilde{q},\tilde{d}_1,\tilde{d}_2}(1,1,1)}{p_{\tilde{d}_1,\tilde{d}_2}(1,1)}. \tag{4.108}$$

The numerator equals

$$p_{\tilde{q},\tilde{d}_1,\tilde{d}_2}(1,1,1) = \mathrm{P}(\tilde{e}_1\tilde{q}=1,\tilde{e}_2\tilde{q}=1\,|\,\tilde{q}=1) \tag{4.109}$$

$$= \mathrm{P}(\tilde{q}=1)\mathrm{P}(\tilde{e}_1=1,\tilde{e}_2=1\,|\,\tilde{q}=1) \tag{4.110}$$

$$= \mathrm{P}(\tilde{q}=1)\mathrm{P}(\tilde{e}_1=1,\tilde{e}_2=1) \tag{4.111}$$

$$= \mathrm{P}(\tilde{q}=1)\mathrm{P}(\tilde{e}_1=1)\mathrm{P}(\tilde{e}_2=1) \tag{4.112}$$

$$= 0.16. \tag{4.113}$$

The denominator equals

$$p_{\tilde{d}_1,\tilde{d}_2}(1,1) = \mathrm{P}(\tilde{e}_1\tilde{q}=1,\tilde{e}_2\tilde{q}=1) \tag{4.114}$$

$$= \mathrm{P}(\tilde{e}_1=1,\tilde{e}_2=1,\tilde{q}=1) + \mathrm{P}(\tilde{e}_1=-1,\tilde{e}_2=-1,\tilde{q}=-1) \tag{4.115}$$

$$= \mathrm{P}(\tilde{e}_1=1)\mathrm{P}(\tilde{e}_2=1)\mathrm{P}(\tilde{q}=1) + \mathrm{P}(\tilde{e}_1=-1)\mathrm{P}(\tilde{e}_2=-1)\mathrm{P}(\tilde{q}=-1)$$

$$= 0.19. \tag{4.116}$$

Consequently,

$$p_{\tilde{q}\,|\,\tilde{d}_1,\tilde{d}_2}(1\,|\,1,1) = 0.84. \tag{4.117}$$

c) As shown above $p_{\tilde{d}_1,\tilde{d}_2}(1,1) = 0.19$, but

$$p_{\tilde{d}_1}(1) = \mathrm{P}(\tilde{e}_1\tilde{q}=1) \tag{4.118}$$

$$= \mathrm{P}(\tilde{e}_1=1,\tilde{q}=1) + \mathrm{P}(\tilde{e}_1=-1,\tilde{q}=-1) \tag{4.119}$$

$$= \mathrm{P}(\tilde{e}_1=1)\mathrm{P}(\tilde{q}=1) + \mathrm{P}(\tilde{e}_1=-1)\mathrm{P}(\tilde{q}=-1) \tag{4.120}$$

$$= 0.35, \tag{4.121}$$

and by the same reasoning,

$$p_{\tilde{d}_2}(1) = 0.35. \tag{4.122}$$

Since $0.19 \neq 0.35^2$, the random variables are not independent.

d) For any x_1, x_2 and q in $\{-1, 1\}$,

$$\begin{aligned} p_{\tilde{d}_1,\tilde{d}_2 \,|\, \tilde{q}}(x_1, x_2 \,|\, q) &= \mathrm{P}(\tilde{e}_1\tilde{q} = x_1, \tilde{e}_2\tilde{q} = x_2 \,|\, \tilde{q} = q) && (4.123)\\ &= \mathrm{P}(\tilde{e}_1 = x_1/q, \tilde{e}_2 = x_2/q \,|\, \tilde{q} = q) && (4.124)\\ &= \mathrm{P}(\tilde{e}_1 = x_1/q)\mathrm{P}(\tilde{e}_2 = x_2/q) \quad \text{by independence} && (4.125)\\ &= \mathrm{P}(\tilde{e}_1 = x_1/q \,|\, \tilde{q} = q)\mathrm{P}(\tilde{e}_2 = x_2/q \,|\, \tilde{q} = q) \quad \text{by independence} \\ &= \mathrm{P}(\tilde{e}_1\tilde{q} = x_1 \,|\, \tilde{q} = q)\mathrm{P}(\tilde{e}_2\tilde{q} = x_2 \,|\, \tilde{q} = q) && (4.126)\\ &= p_{\tilde{d}_1 \,|\, \tilde{q}}(x_1 \,|\, q) p_{\tilde{d}_2 \,|\, \tilde{q}}(x_2 \,|\, q), && (4.127) \end{aligned}$$

so $\tilde{d}_1$ and $\tilde{d}_2$ are conditionally independent given $\tilde{q}$.

4.11 (Surgery)

a) Let us define the potential outcomes $\widetilde{\mathrm{po}}_A$ and $\widetilde{\mathrm{po}}_B$. $\widetilde{\mathrm{po}}_A = 1$ and $\widetilde{\mathrm{po}}_B = 1$ indicate recovery after procedure A and B respectively. $\widetilde{\mathrm{po}}_A = 0$ and $\widetilde{\mathrm{po}}_B = 0$ indicate non-recovery. The random variable $\tilde{y}$ represents the observed outcome and $\tilde{t}$ the treatment, so that

$$\tilde{y} := \begin{cases} \widetilde{\mathrm{po}}_A & \text{if } \tilde{t} = A, \\ \widetilde{\mathrm{po}}_B & \text{if } \tilde{t} = B. \end{cases} \tag{4.128}$$

We also define a random variable $\tilde{d}$ to indicate whether each case is mild ($\tilde{d} = m$) or serious ($\tilde{d} = s$). We denote the probability that a case is mild if the treatment is A or B by α_A and α_B respectively. The observed probability of recovery for patients undergoing procedure A is

$$\begin{aligned} \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = A\right) &= \mathrm{P}\left(\widetilde{\mathrm{po}}_A = 1 \,|\, \tilde{t} = A\right) && (4.129)\\ &= \mathrm{P}\left(\widetilde{\mathrm{po}}_A = 1, \tilde{d} = m \,|\, \tilde{t} = A\right) + \mathrm{P}\left(\widetilde{\mathrm{po}}_A = 1, \tilde{d} = s \,|\, \tilde{t} = A\right) && (4.130)\\ &= \mathrm{P}\left(\tilde{d} = m \,|\, \tilde{t} = A\right) \mathrm{P}\left(\widetilde{\mathrm{po}}_A = 1 \,|\, \tilde{t} = A, \tilde{d} = m\right) && (4.131)\\ &\quad + \mathrm{P}\left(\tilde{d} = s \,|\, \tilde{t} = A\right) \mathrm{P}\left(\widetilde{\mathrm{po}}_A = 1 \,|\, \tilde{t} = A, \tilde{d} = s\right) && (4.132)\\ &= 0.9\alpha_A + 0.5(1 - \alpha_A). && (4.133) \end{aligned}$$

Setting this equal to 0.58 yields $\alpha_A = 0.2$.

Similarly,

$$\begin{aligned} \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = B\right) &= \mathrm{P}\left(\widetilde{\mathrm{po}}_B = 1 \,|\, \tilde{t} = B\right) && (4.134)\\ &= \mathrm{P}\left(\widetilde{\mathrm{po}}_B = 1, \tilde{d} = m \,|\, \tilde{t} = B\right) + \mathrm{P}\left(\widetilde{\mathrm{po}}_B = 1, \tilde{d} = s \,|\, \tilde{t} = B\right) && (4.135)\\ &= \mathrm{P}\left(\tilde{d} = m \,|\, \tilde{t} = B\right) \mathrm{P}\left(\widetilde{\mathrm{po}}_B = 1 \,|\, \tilde{t} = B, \tilde{d} = m\right) && (4.136)\\ &\quad + \mathrm{P}\left(\tilde{d} = s \,|\, \tilde{t} = B\right) \mathrm{P}\left(\widetilde{\mathrm{po}}_B = 1 \,|\, \tilde{t} = B, \tilde{d} = s\right) && (4.137)\\ &= 0.8\alpha_B + 0.2(1 - \alpha_B). && (4.138) \end{aligned}$$

Setting this equal to 0.68 yields $\alpha_B = 0.8$.

What is happening is that the fraction of patients with mild cases is much lower for procedure A (20%) than for procedure B (80%). Since those patients are more likely to recover, irrespective of the procedure, this inflates the recovery rate for procedure B.

b) We can correct for the confounder by taking into account whether each case is mild or serious. Under the assumption that the treatment $\tilde{t}$ and the potential outcomes $\widetilde{\mathrm{po}}_A$

and $\widetilde{\text{po}}_B$ are conditionally independent given the degree of severity $\tilde{d}$, then

$$\begin{aligned} \mathrm{P}\left(\tilde{y} \,|\, \tilde{t} = A, \tilde{d} = m\right) &= \mathrm{P}\left(\widetilde{\text{po}}_A \,|\, \tilde{t} = A, \tilde{d} = m\right) && (4.139) \\ &= \mathrm{P}\left(\widetilde{\text{po}}_A \,|\, \tilde{d} = m\right). && (4.140) \end{aligned}$$

By the same argument,

$$\begin{aligned} \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = A, \tilde{d} = s\right) &= \mathrm{P}\left(\widetilde{\text{po}}_A = 1 \,|\, \tilde{d} = s\right), && (4.141) \\ \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = B, \tilde{d} = m\right) &= \mathrm{P}\left(\widetilde{\text{po}}_B = 1 \,|\, \tilde{d} = m\right), && (4.142) \\ \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = B, \tilde{d} = s\right) &= \mathrm{P}\left(\widetilde{\text{po}}_B = 1 \,|\, \tilde{d} = s\right). && (4.143) \end{aligned}$$

Consequently, we can compute the *true* efficacy as follows,

$$\begin{aligned} \mathrm{P}\left(\widetilde{\text{po}}_A\right) &= \sum_{d \in \{m,s\}} \mathrm{P}\left(\widetilde{\text{po}}_A = 1 \,|\, \tilde{d} = d\right) \mathrm{P}\left(\tilde{d} = d\right) && (4.144) \\ &= \sum_{d \in \{m,s\}} \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = A, \tilde{d} = d\right) \mathrm{P}\left(\tilde{d} = d\right), && (4.145) \\ \mathrm{P}\left(\widetilde{\text{po}}_B\right) &= \sum_{d \in \{m,s\}} \mathrm{P}\left(\widetilde{\text{po}}_B = 1 \,|\, \tilde{d} = d\right) \mathrm{P}\left(\tilde{d} = d\right) && (4.146) \\ &= \sum_{d \in \{m,s\}} \mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = B, \tilde{d} = d\right) \mathrm{P}\left(\tilde{d} = d\right). && (4.147) \end{aligned}$$

This only works if the conditional independence assumption holds. Intuitively, we are assuming that once we control for the severity, there are no other systematic differences between the patients that undergo each of the procedures.

c) Randomizing what patient undergoes each procedure would enable us to neutralize all confounding factors, even if we don't know what they are.

4.12 (Admissions)

a) The empirical conditional probabilities are

$$\begin{aligned} \mathrm{P}\left(\text{admitted} \,|\, \text{men}\right) &= \frac{40 + 25}{500 + 500} && (4.148) \\ &= 0.065, && (4.149) \\ \mathrm{P}\left(\text{admitted} \,|\, \text{women}\right) &= \frac{50 + 25}{800 + 200} && (4.150) \\ &= 0.075. && (4.151) \end{aligned}$$

b) By the assumptions,

$$\begin{aligned}
\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1\right) &= \mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1,\tilde{x}=\mathrm{Med}\right) && (4.152)\\
&\quad+\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1,\tilde{x}=\mathrm{Art}\right) && (4.153)\\
&= \mathrm{P}\left(\tilde{x}=\mathrm{Med}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1\,|\,\tilde{x}=\mathrm{Med}\right) && (4.154)\\
&\quad+\mathrm{P}\left(\tilde{x}=\mathrm{Art}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1\,|\,\tilde{x}=\mathrm{Art}\right) && (4.155)\\
&= \mathrm{P}\left(\tilde{x}=\mathrm{Med}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1\,|\,\tilde{x}=\mathrm{Med},\tilde{s}=\mathrm{man}\right) && (4.156)\\
&\quad+\mathrm{P}\left(\tilde{x}=\mathrm{Art}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1\,|\,\tilde{x}=\mathrm{Art},\tilde{s}=\mathrm{man}\right) && (4.157)\\
&= \mathrm{P}\left(\tilde{x}=\mathrm{Med}\right)\mathrm{P}\left(\tilde{y}=1\,|\,\tilde{x}=\mathrm{Med},\tilde{s}=\mathrm{man}\right) && (4.158)\\
&\quad+\mathrm{P}\left(\tilde{x}=\mathrm{Art}\right)\mathrm{P}\left(\tilde{y}=1\,|\,\tilde{x}=\mathrm{Art},\tilde{s}=\mathrm{man}\right) && (4.159)\\
&= \frac{1300}{2000}\frac{40}{800}+\frac{700}{2000}\frac{25}{200} && (4.160)\\
&= 0.076, && (4.161)
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{woman}}=1\right) &= \mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{man}}=1,\tilde{x}=\mathrm{Med}\right) && (4.162)\\
&\quad+\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{woman}}=1,\tilde{x}=\mathrm{Art}\right) && (4.163)\\
&= \mathrm{P}\left(\tilde{x}=\mathrm{Med}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{woman}}=1\,|\,\tilde{x}=\mathrm{Med}\right) && (4.164)\\
&\quad+\mathrm{P}\left(\tilde{x}=\mathrm{Art}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{woman}}=1\,|\,\tilde{x}=\mathrm{Art}\right) && (4.165)\\
&= \mathrm{P}\left(\tilde{x}=\mathrm{Med}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{woman}}=1\,|\,\tilde{x}=\mathrm{Med},\tilde{s}=\mathrm{woman}\right) && (4.166)\\
&\quad+\mathrm{P}\left(\tilde{x}=\mathrm{Art}\right)\mathrm{P}\left(\widetilde{\mathrm{po}}_{\mathrm{woman}}=1\,|\,\tilde{x}=\mathrm{Art},\tilde{s}=\mathrm{woman}\right) && (4.167)\\
&= \mathrm{P}\left(\tilde{x}=\mathrm{Med}\right)\mathrm{P}\left(\tilde{y}=1\,|\,\tilde{x}=\mathrm{Med},\tilde{s}=\mathrm{woman}\right) && (4.168)\\
&\quad+\mathrm{P}\left(\tilde{x}=\mathrm{Art}\right)\mathrm{P}\left(\tilde{y}=1\,|\,\tilde{x}=\mathrm{Art},\tilde{s}=\mathrm{woman}\right) && (4.169)\\
&= \frac{1300}{2000}\frac{25}{500}+\frac{700}{2000}\frac{50}{500} && (4.170)\\
&= 0.068. && (4.171)
\end{aligned}$$

Once we adjust by the school they apply to, men are more likely to be admitted than women, so the data do not support their hypothesis.

4.13 (Shop)

a) From the data the probability of a purchase if there is music is 120/160=0.75. If there is no music it is $109/190 = 0.57$.

b) We define random variables $\tilde{a}$ and $\tilde{m}$ to represent age and music respectively:

$$\begin{aligned}
p_{\widetilde{\mathrm{po}}_0}(1) &= \sum_{a\in\{\mathrm{young,mid,old}\}} p_{\tilde{a}}(a)\,p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a}}(1\,|\,a) && (4.172)\\
&= \sum_{a\in\{\mathrm{young,mid,old}\}} p_{\tilde{a}}(a)\,p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a},\tilde{m}}(1\,|\,a,0) && (4.173)\\
&= \sum_{a\in\{\mathrm{young,mid,old}\}} p_{\tilde{a}}(a)\,p_{\tilde{y}\,|\,\tilde{a},\tilde{m}}(1\,|\,a,0) && (4.174)\\
&= \frac{100}{350}\cdot\frac{10}{20}+\frac{110}{350}\cdot\frac{90}{100}+\frac{140}{350}\cdot\frac{20}{40} && (4.175)\\
&= 0.626. && (4.176)
\end{aligned}$$

$$p_{\widetilde{\text{po}}_1}(1) = \sum_{a \in \{\text{young,mid,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\text{po}}_1 \mid \tilde{a}}(1 \mid a) \tag{4.177}$$

$$= \sum_{a \in \{\text{young,mid,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\text{po}}_1 \mid \tilde{a}, \tilde{m}}(1 \mid a, 1) \tag{4.178}$$

$$= \sum_{a \in \{\text{young,mid,old}\}} p_{\tilde{a}}(a) p_{\tilde{y} \mid \tilde{a}, \tilde{m}}(1 \mid a, 1) \tag{4.179}$$

$$= \frac{100}{350} \cdot \frac{40}{80} + \frac{110}{350} \cdot \frac{9}{10} + \frac{140}{350} \cdot \frac{60}{100} \tag{4.180}$$

$$= 0.666. \tag{4.181}$$

Playing music decreases purchases under the conditional-independence assumption.

4.14 (Rackets)

a) True. Let $\tilde{s}$ be a random variable that represents the surface. If $\alpha := \mathrm{P}(\tilde{s} = \text{clay} \mid \tilde{r} = A)$ equals $\mathrm{P}(\tilde{s} = \text{clay})$, then $1 - \alpha := \mathrm{P}(\tilde{s} = \text{grass} \mid \tilde{r} = A)$ equals $\mathrm{P}(\tilde{s} = \text{grass})$, so that

$$\mathrm{P}(\tilde{y} = 1 \mid \tilde{r} = A) = \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{y} = 1, \tilde{s} = s \mid \tilde{r} = A) \tag{4.182}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\widetilde{\text{po}}_A = 1, \tilde{s} = s \mid \tilde{r} = A) \tag{4.183}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{s} = s \mid \tilde{r} = A) \mathrm{P}(\widetilde{\text{po}}_A = 1 \mid \tilde{r} = A, \tilde{s} = s) \tag{4.184}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{s} = s \mid \tilde{r} = A) \mathrm{P}(\widetilde{\text{po}}_A = 1 \mid \tilde{s} = s) \tag{4.185}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{s} = s) \mathrm{P}(\widetilde{\text{po}}_A = 1 \mid \tilde{s} = s) \tag{4.186}$$

$$= \mathrm{P}(\widetilde{\text{po}}_A = 1). \tag{4.187}$$

b)

$$\mathrm{P}(\tilde{y} = 1 \mid \tilde{r} = A) = \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{y} = 1, \tilde{s} = s \mid \tilde{r} = A) \tag{4.188}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\widetilde{\text{po}}_A = 1, \tilde{s} = s \mid \tilde{r} = A) \tag{4.189}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{s} = s \mid \tilde{r} = A) \mathrm{P}(\widetilde{\text{po}}_A = 1 \mid \tilde{r} = A, \tilde{s} = s) \tag{4.190}$$

$$= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}(\tilde{s} = s \mid \tilde{r} = A) \mathrm{P}(\widetilde{\text{po}}_A = 1 \mid \tilde{s} = s) \tag{4.191}$$

$$= 0.8\alpha + 0.2(1 - \alpha) = 0.2 + 0.6\alpha. \tag{4.192}$$

Similarly,

$$\begin{aligned}
\mathrm{P}\left(\tilde{y}=1 \,|\, \tilde{r}=B\right) &= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}\left(\tilde{y}=1, \tilde{s}=s \,|\, \tilde{r}=B\right) && (4.193)\\
&= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}\left(\widetilde{\text{po}}_A=1, \tilde{s}=s \,|\, \tilde{r}=B\right) && (4.194)\\
&= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}\left(\tilde{s}=s \,|\, \tilde{r}=A\right) \mathrm{P}\left(\widetilde{\text{po}}_B=1 \,|\, \tilde{r}=B, \tilde{s}=s\right) && (4.195)\\
&= \sum_{s \in \{\text{clay,grass}\}} \mathrm{P}\left(\tilde{s}=s \,|\, \tilde{r}=A\right) \mathrm{P}\left(\widetilde{\text{po}}_B=1 \,|\, \tilde{s}=s\right) && (4.196)\\
&= 0.7 \cdot 0.5 + 0.1 \cdot 0.5 = 0.4. && (4.197)
\end{aligned}$$

We need $0.2 + 0.6\alpha < 0.4$, which occurs for $\alpha < 1/3$.

4.15 (Missing data)

a) No, otherwise there wouldn't be such a large discrepancy between the number of missing data for men and women.

b) 70 out of 140 observed cases have side effects, so the probability is 0.5.

c) Let $\tilde{\text{se}}$, $\tilde{o}$ and $\tilde{s}$ be Bernoulli random variables representing the side effects, whether the data are observed ($\tilde{o}=1$) or not ($\tilde{o}=1$) and the sex respectively. We have

$$p_{\tilde{\text{se}}}(1) = \sum_{s \in \{\text{man,woman}\}} p_{\tilde{s}}(s) p_{\tilde{\text{se}} \,|\, \tilde{s}}(1 \,|\, s). \tag{4.198}$$

If the side effects are conditionally independent from the data being observed or missing given the patient sex, then

$$\begin{aligned}
p_{\tilde{\text{se}} \,|\, \tilde{s}}(1 \,|\, \text{man}) &= p_{\tilde{\text{se}} \,|\, \tilde{o}, \tilde{s}}(1 \,|\, 1, \text{man}) && (4.199)\\
&= \frac{40}{50} = 0.8, && (4.200)\\
p_{\tilde{\text{se}} \,|\, \tilde{s}}(1 \,|\, \text{woman}) &= p_{\tilde{\text{se}} \,|\, \tilde{o}, \tilde{s}}(1 \,|\, 1, \text{woman}) && (4.201)\\
&= \frac{30}{90} = 0.33. && (4.202)
\end{aligned}$$

Consequently, since $p_{\tilde{s}}(\text{man}) = p_{\tilde{s}}(\text{woman}) = \frac{1}{2}$,

$$p_{\tilde{\text{se}}}(1) = \frac{0.8 + 0.33}{2} = 0.57. \tag{4.203}$$

4.16 (Three players)

a) We cannot because there are no games were all three players are absent, so we cannot compute what fraction are wins or losses.

b) Let us represent the presence or absence of the players using a random vector $\tilde{x}$: $\tilde{x}[1]$, $\tilde{x}[2]$ and $\tilde{x}[3]$ are 1 if James, Kevin and Kyrie are present respectively, and 0 if they're absent. We represent the result of the game using a Bernoulli random variable $\tilde{y}$ (win is $\tilde{y}=1$ and loss $\tilde{y}=0$). Using the naive Bayes assumption and empirical estimates of

$p_{\tilde{y}}$ and $p_{\tilde{x}[i] \mid \tilde{y}}$

$$\begin{aligned} p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(1,0,0,0) &= p_{\tilde{y}}(1) p_{\tilde{x}[1] \mid \tilde{y}}(0 \mid 1) p_{\tilde{x}[2] \mid \tilde{y}}(0 \mid 1) p_{\tilde{x}[3] \mid \tilde{y}}(0 \mid 1) \\ &= \frac{6}{10}\frac{2}{6}\frac{1}{6}\frac{2}{6} = \frac{1}{90}, \end{aligned} \tag{4.204}$$

$$\begin{aligned} p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(0,0,0,0) &= p_{\tilde{y}}(0) p_{\tilde{x}[1] \mid \tilde{y}}(0 \mid 0) p_{\tilde{x}[2] \mid \tilde{y}}(0 \mid 0) p_{\tilde{x}[3] \mid \tilde{y}}(0 \mid 0) \\ &= \frac{4}{10}\frac{1}{4}\frac{3}{4}\frac{3}{4} = \frac{9}{160}. \end{aligned} \tag{4.205}$$

As a result,

$$p_{\tilde{y} \mid \tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(1 \mid 0,0,0) \tag{4.206}$$

$$= \frac{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(1,0,0,0)}{p_{\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(0,0,0)} \tag{4.207}$$

$$= \frac{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(1,0,0,0)}{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(1,0,0,0) + p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3]}(0,0,0,0)} \tag{4.208}$$

$$= \frac{\frac{1}{90}}{\frac{1}{90} + \frac{9}{160}} \tag{4.209}$$

$$= 0.165. \tag{4.210}$$

4.17 (Spam detector)

a) The problem is that there are $2^4 = 16$ different possible values for the entries of $\tilde{x}$, and hence 16 different conditional distributions. However, we only have 10 data points.

b) The email corresponds to $\tilde{x}[1] = 0$, $\tilde{x}[2] = 0$, $\tilde{x}[3] = 1$, $\tilde{x}[4] = 1$. By the naive Bayes assumption combined with empirical estimates of $p_{\tilde{y}}$ and $p_{\tilde{x}[j] \mid \tilde{y}}$

$$\begin{aligned} p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,0,1,1) &= p_{\tilde{y}}(1) p_{\tilde{x}[1] \mid \tilde{y}}(0 \mid 1) p_{\tilde{x}[2] \mid \tilde{y}}(0 \mid 1) p_{\tilde{x}[3] \mid \tilde{y}}(1 \mid 1) p_{\tilde{x}[4] \mid \tilde{y}}(1 \mid 1) \\ &= \frac{5}{10}\frac{1}{5}\frac{3}{5}\frac{2}{5}\frac{1}{5}, \end{aligned} \tag{4.211}$$

$$\begin{aligned} p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0,0,0,1,1) &= p_{\tilde{y}}(0) p_{\tilde{x}[1] \mid \tilde{y}}(0 \mid 0) p_{\tilde{x}[2] \mid \tilde{y}}(0 \mid 0) p_{\tilde{x}[3] \mid \tilde{y}}(1 \mid 0) p_{\tilde{x}[4] \mid \tilde{y}}(1 \mid 0) \\ &= \frac{5}{10}\frac{4}{5}\frac{2}{5}\frac{2}{5}\frac{4}{5}. \end{aligned} \tag{4.212}$$

As a result,

$$p_{\tilde{y} \mid \tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1 \mid 0,0,1,1) \tag{4.213}$$

$$= \frac{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,0,1,1)}{p_{\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0,0,1,1)} \tag{4.214}$$

$$= \frac{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,0,1,1)}{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,0,1,1) + p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0,0,0,1,1)} \tag{4.215}$$

$$= \frac{3}{3 + 4 \cdot 2 \cdot 4} \tag{4.216}$$

$$= \frac{3}{35}, \tag{4.217}$$

so $p_{\tilde{y} \mid \tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0 \mid 0,0,1,1) = \frac{32}{35}$ and we classify the email as not spam.

c) The email corresponds to $\tilde{x}[1] = 0$, $\tilde{x}[2] = 1$, $\tilde{x}[3] = 1$, $\tilde{x}[4] = 0$. By the naive Bayes

assumption combined with empirical estimates of $p_{\tilde{y}}$ and $p_{\tilde{x}[j] \mid \tilde{y}}$

$$p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,1,1,0) = p_{\tilde{y}}(1) p_{\tilde{x}[1] \mid \tilde{y}}(0 \mid 1) p_{\tilde{x}[2] \mid \tilde{y}}(1 \mid 1) p_{\tilde{x}[3] \mid \tilde{y}}(1 \mid 1) p_{\tilde{x}[4] \mid \tilde{y}}(0 \mid 1) = \frac{5}{10} \frac{1}{5} \frac{2}{5} \frac{2}{5} \frac{4}{5}, \tag{4.218}$$

$$p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0,0,1,1,0) = p_{\tilde{y}}(0) p_{\tilde{x}[1] \mid \tilde{y}}(0 \mid 0) p_{\tilde{x}[2] \mid \tilde{y}}(1 \mid 0) p_{\tilde{x}[3] \mid \tilde{y}}(1 \mid 0) p_{\tilde{x}[4] \mid \tilde{y}}(0 \mid 0) = \frac{5}{10} \frac{4}{5} \frac{3}{5} \frac{2}{5} \frac{1}{5}. \tag{4.219}$$

As a result,

$$p_{\tilde{y} \mid \tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1 \mid 0,1,1,0) \tag{4.220}$$

$$= \frac{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,1,1,0)}{p_{\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0,1,1,0)} \tag{4.221}$$

$$= \frac{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,1,1,0)}{p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(1,0,1,1,0) + p_{\tilde{y},\tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0,0,1,1,0)} \tag{4.222}$$

$$= \frac{2 \cdot 2 \cdot 4}{4 \cdot 3 \cdot 2 + 2 \cdot 2 \cdot 4} \tag{4.223}$$

$$= \frac{2}{5}, \tag{4.224}$$

so $p_{\tilde{y} \mid \tilde{x}[1],\tilde{x}[2],\tilde{x}[3],\tilde{x}[4]}(0 \mid 0,1,1,0) = \frac{3}{5}$ and we classify the email as not spam.
In our data the words *alternative* and *medicine* only appear together in spam emails, but the naive Bayes classifier ignores these dependencies.

4.18 (Symbols)

a) We use empirical probabilities to estimate the transition matrix.

$*$	†	□	
0	$\frac{1}{6}$	$\frac{4}{11}$	$*$
$\frac{3}{7}$	$\frac{2}{3}$	$\frac{3}{11}$	†
$\frac{4}{7}$	$\frac{1}{6}$	$\frac{4}{11}$	□

The Markov assumption is not consistent with the data. Under the assumption, the transition † → † is equally likely no matter what symbols occur before, but this is clearly violated in the data: if the previous three symbols are □ † † or $*$ † † then the following symbol is always †, but if the previous three symbols are ††† then the following symbol is never †.

b) The next symbol is either † or □ because the transition $* \to *$ has probability zero. If the next symbol is † the most likely sequence is ††, which has probability

$$\mathrm{P}(\text{next 2 symbols} = ††) = \mathrm{P}(* \to †)\, \mathrm{P}(† \to †) \tag{4.225}$$

$$= \frac{3}{7} \cdot \frac{2}{3} = \frac{2}{7}. \tag{4.226}$$

If the next symbol is □ the most likely sequence is □□ or □$*$, which have the same probability, equal to

$$\mathrm{P}(\text{next 2 symbols} = □□) = \mathrm{P}(* \to □)\, \mathrm{P}(□ \to □) \tag{4.227}$$

$$= \frac{4}{7} \cdot \frac{4}{11} = \frac{16}{77} < \frac{22}{77} = \frac{2}{7}. \tag{4.228}$$

We conclude that the most likely next two symbols according to the Markov-chain model are ††.

4.19 (The Markov property) Following the hint, we first establish that

$$p_{\tilde{a}_{i+1}\,|\,\tilde{a}_{i-1},\tilde{a}_i}\left(a_{i+1}\,|\,a_{i-1},a_i\right) = \frac{p_{\tilde{a}_{i-1},\tilde{a}_i,\tilde{a}_{i+1}}\left(a_{i-1},a_i,a_{i+1}\right)}{p_{\tilde{a}_{i-1},\tilde{a}_i}\left(a_{i-1},a_i\right)} \tag{4.229}$$

$$= \frac{\sum_{a_1,\ldots,a_{i-2}} p_{\tilde{a}_1,\ldots,\tilde{a}_{i+1}}\left(a_1,\ldots,a_{i+1}\right)}{\sum_{a_1,\ldots,a_{i-2}} p_{\tilde{a}_1,\ldots,\tilde{a}_i}\left(a_1,\ldots,a_i\right)} \tag{4.230}$$

$$= \frac{\sum_{a_1,\ldots,a_{i-2}} p_{\tilde{a}_1}\left(a_1\right)\prod_{j=1}^{i} p_{\tilde{a}_{j+1}\,|\,\tilde{a}_j}\left(a_{j+1}\,|\,a_j\right)}{\sum_{a_1,\ldots,a_{i-2}} p_{\tilde{a}_1}\left(a_1\right)\prod_{j=1}^{i-1} p_{\tilde{a}_{j+1}\,|\,\tilde{a}_j}\left(a_{j+1}\,|\,a_j\right)} \tag{4.231}$$

$$= \frac{p_{\tilde{a}_{i+1}\,|\,\tilde{a}_i}\left(a_{i+1}\,|\,a_i\right)\sum_{a_1,\ldots,a_{i-2}} p_{\tilde{a}_1}\left(a_1\right)\prod_{j=1}^{i-1} p_{\tilde{a}_{j+1}\,|\,\tilde{a}_j}\left(a_{j+1}\,|\,a_j\right)}{\sum_{a_1,\ldots,a_{i-2}} p_{\tilde{a}_1}\left(a_1\right)\prod_{j=1}^{i-1} p_{\tilde{a}_{j+1}\,|\,\tilde{a}_j}\left(a_{j+1}\,|\,a_j\right)}$$

$$= p_{\tilde{a}_{i+1}\,|\,\tilde{a}_i}\left(a_{i+1}\,|\,a_i\right). \tag{4.232}$$

We then use the result to conclude

$$p_{\tilde{a}_{i+1},\tilde{a}_{i-1}\,|\,\tilde{a}_i}\left(a_{i+1},a_{i-1}\,|\,a_i\right) = \frac{p_{\tilde{a}_{i-1},\tilde{a}_i,\tilde{a}_{i+1}}\left(a_{i-1},a_i,a_{i+1}\right)}{p_{\tilde{a}_i}\left(a_i\right)} \tag{4.233}$$

$$= \frac{p_{\tilde{a}_{i-1},\tilde{a}_i}\left(a_{i-1},a_i\right)p_{\tilde{a}_{i+1}\,|\,\tilde{a}_{i-1},\tilde{a}_i}\left(a_{i+1}\,|\,a_{i-1},a_i\right)}{p_{\tilde{a}_i}\left(a_i\right)} \tag{4.234}$$

$$= \frac{p_{\tilde{a}_i}\left(a_i\right)p_{\tilde{a}_{i-1}\,|\,\tilde{a}_i}\left(a_{i-1}\,|\,a_i\right)p_{\tilde{a}_{i+1}\,|\,\tilde{a}_i}\left(a_{i+1}\,|\,a_i\right)}{p_{\tilde{a}_i}\left(a_i\right)} \tag{4.235}$$

$$= p_{\tilde{a}_{i+1}\,|\,\tilde{a}_i}\left(a_{i+1}\,|\,a_i\right)p_{\tilde{a}_{i-1}\,|\,\tilde{a}_i}\left(a_{i-1}\,|\,a_i\right). \tag{4.236}$$

4.20 (Employment dynamics)

a) The pmf of $\tilde{s}$ equals

$$p_{\tilde{s}}(s) = \mathrm{P}\left(\tilde{x}_1 = \text{student},\ldots,\tilde{x}_s = \text{student},\tilde{x}_{s+1} = \text{not a student}\right) \tag{4.237}$$

$$= 0.8^s 0.2. \tag{4.238}$$

b)

$$\mathrm{P}\left(\tilde{x}_1 = \text{student}\,|\,\tilde{x}_3 = \text{employed}\right) \tag{4.239}$$

$$= \frac{\mathrm{P}\left(\tilde{x}_1 = \text{student},\tilde{x}_3 = \text{employed}\right)}{\mathrm{P}\left(\tilde{x}_3 = \text{employed}\right)} \tag{4.240}$$

$$= \frac{\sum_{x_2\in\{\text{st., em., un.}\}}\mathrm{P}\left(\tilde{x}_1 = \text{student},\tilde{x}_2 = x_2,\tilde{x}_3 = \text{employed}\right)}{\sum_{x_1\in\{\text{st., em., un.}\}}\sum_{x_2\in\{\text{st., em., un.}\}}\mathrm{P}\left(\tilde{x}_1 = x_1,\tilde{x}_2 = x_2,\tilde{x}_3 = \text{employed}\right)} \tag{4.241}$$

$$= \frac{0.8(0.8\cdot 0.2 + 0.2\cdot 0.9)}{0.8(0.8\cdot 0.2 + 0.2\cdot 0.9) + 0.2(0.9\cdot 0.9 + 0.1\cdot 0.4)} \tag{4.242}$$

$$= 0.615. \tag{4.243}$$

c) A stationary distribution should satisfy $T\pi_* = \pi_*$, which implies $0.8\pi_*[1] = \pi_*[1]$. This is only possible if $\pi_*[1] = 0$, which implies

$$\pi_* = \begin{bmatrix} 0 \\ \alpha \\ 1-\alpha \end{bmatrix}, \tag{4.244}$$

for some $0 \leq \alpha \leq 1$. Setting $T\pi_* = \pi_*$ yields the equations,

$$0.9\alpha + 0.4(1-\alpha) = \alpha, \tag{4.245}$$

$$0.1\alpha + 0.6(1-\alpha) = 1-\alpha. \tag{4.246}$$

Solving the equations, we obtain $\alpha = 0.8$.

4.21 (Cellphones)

a) The initial state vector and the transition matrix of the Markov chain are

$$\pi_1 := \begin{bmatrix} 0.9 \\ 0.1 \\ 0 \end{bmatrix}, \qquad T = \begin{bmatrix} 0.7 & 0 & 0 \\ 0.2 & 1 & 0 \\ 0.1 & 0 & 1 \end{bmatrix}. \tag{4.247}$$

The transition matrix T has three eigenvectors

$$q_1 := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad q_2 := \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad q_3 := \begin{bmatrix} 0.80 \\ -0.53 \\ -0.27 \end{bmatrix}. \tag{4.248}$$

The corresponding eigenvalues are $\lambda_1 := 1$, $\lambda_2 := 1$ and $\lambda_3 := 0.7$. We gather the eigenvectors and eigenvalues into two matrices

$$Q := \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix}, \qquad \Lambda := \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \tag{4.249}$$

so that the eigendecomposition of T is

$$T := Q\Lambda Q^{-1}. \tag{4.250}$$

We compute

$$\alpha := Q^{-1}\pi_1 = \begin{bmatrix} 0.3 \\ 0.7 \\ 1.122 \end{bmatrix}. \tag{4.251}$$

By Theorem 4.39,

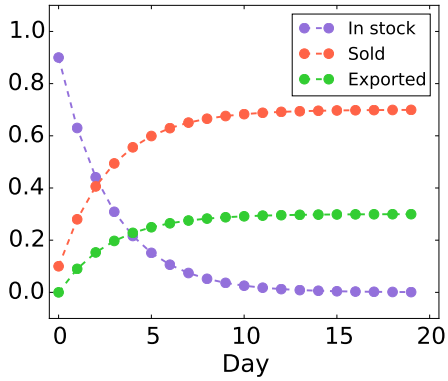
$$\lim_{i\to\infty} \pi_i = \lim_{i\to\infty} \sum_{j=1}^{3} \lambda^{i-1} \alpha_j q_j \tag{4.252}$$

$$= 0.3\, q_1 + 0.7\, q_2 \tag{4.253}$$

$$= \begin{bmatrix} 0 \\ 0.7 \\ 0.3 \end{bmatrix}. \tag{4.254}$$

This means that eventually the probability that each phone has been sold locally converges to 0.7 and the probability that it has been exported converges to 0.3.

b) The following graph shows the evolution of the state vector of the simulated Markov chain.



4.22 (Empirical state vector and transition matrix) For $1 \leq j \leq m$, $\pi_X[j]$ is the fraction of data equal to s_j. Since n_j is equal to the number of transitions starting at s_j, $\pi_X[j]$ is equal to $(n_j - 1)/n$ if x_n is equal to s_j or to n_j/n otherwise, so unless n_j is very small

$$\pi_X[j] \approx \frac{n_j}{n}. \tag{4.255}$$

For any l between 2 and n,

$$\sum_{j=1}^{m} 1\left(x_{l-1} = s_j, x_l = s_i\right) = 1\left(x_l = s_i\right), \tag{4.256}$$

because x_{l-1} is equal to exactly one of the possible values $s_1, \ldots, s_m$. Consequently,

$$\begin{aligned}
\left(T_X \pi_X\right)[i] &= \sum_{j=1}^{m} T_X[i,j] \pi_X[j] && (4.257)\\
&= \sum_{j=1}^{m} \frac{\pi_X[j]}{n_j} \sum_{l=2}^{n} 1\left(x_{l-1} = s_j, x_l = s_i\right) && (4.258)\\
&\approx \frac{1}{n} \sum_{l=2}^{n} \sum_{j=1}^{m} 1\left(x_{l-1} = s_j, x_l = s_i\right) && (4.259)\\
&= \frac{1}{n} \sum_{l=2}^{n} 1\left(x_l = s_i\right) && (4.260)\\
&\approx \frac{1}{n} \sum_{l=1}^{n} 1\left(x_l = s_i\right) && (4.261)\\
&= \pi_X[i]. && (4.262)
\end{aligned}$$

5

Multiple Continuous Variables

Exercises

5.1 (Identities) The only identities that hold are

$$\int_{x=-\infty}^{\infty} f_{\tilde{x}\,|\,\tilde{y},\tilde{z}}\,(x\,|\,y,z)\;\mathrm{d}x = 1, \tag{5.1}$$

$$\int_{x=-\infty}^{\infty} f_{\tilde{x},\tilde{y}\,|\,\tilde{z}}\,(x,y\,|\,z)\;\mathrm{d}x = f_{\tilde{y}\,|\,\tilde{z}}\,(y\,|\,z)\,. \tag{5.2}$$

The function $f_{\tilde{x}\,|\,\tilde{y},\tilde{z}}(x\,|\,y,z)$ is a valid (conditional) probability density of $\tilde{x}$ for any fixed y and z (but not of $\tilde{y}$ or $\tilde{z}$), so if we integrate it with respect to x, the integral equals one. Similarly, the function $f_{\tilde{x},\tilde{y}\,|\,\tilde{z}}\,(x,y\,|\,z)$ is a valid (conditional) joint probability density of $\tilde{x}$ and $\tilde{y}$ for any fixed z (but not of $\tilde{z}$), so if we integrate it with respect to x, we obtain the (conditional) marginal pdf of $\tilde{y}$.

5.2 (Cross)

a) The value of the pdf is 1/12, because the area of the cross equals 12. If $x < -2$ or $x > 2$, $f_{\tilde{x}}(x) = 0$. If $-2 \leq x \leq -1$ or $1 \leq x \leq 2$,

$$f_{\tilde{x}}(x) = \int_{y\in\mathbb{R}} f_{\tilde{x},\tilde{y}}(x,y)\;\mathrm{d}y \tag{5.3}$$

$$= \int_{-1}^{1} \frac{1}{12}\;\mathrm{d}y \tag{5.4}$$

$$= \frac{1}{6}. \tag{5.5}$$

If $-1 \leq x \leq 1$

$$f_{\tilde{x}}(x) = \int_{y\in\mathbb{R}} f_{\tilde{x},\tilde{y}}(x,y)\;\mathrm{d}y \tag{5.6}$$

$$= \int_{-2}^{2} \frac{1}{12}\;\mathrm{d}y \tag{5.7}$$

$$= \frac{1}{3}. \tag{5.8}$$

b) If $x < -2$ or $x > 2$ the conditional pdf is not defined. If $-2 \leq x \leq -1$ or $1 \leq x \leq 2$, the conditional pdf is uniform between -1 and 1, and equal to

$$f_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x) = \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)} \tag{5.9}$$

$$= \frac{\frac{1}{12}}{\frac{1}{6}} = \frac{1}{2}. \tag{5.10}$$

If $-1 \leq x \leq 1$ the conditional pdf is uniform between -2 and 2, and equal to

$$f_{\tilde{y} \mid \tilde{x}}(y \mid x) = \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)} \tag{5.11}$$

$$= \frac{\frac{1}{12}}{\frac{1}{3}} = \frac{1}{4}. \tag{5.12}$$

Since the conditional distributions are different, $\tilde{x}$ and $\tilde{y}$ are not independent.

5.3 (Two random variables)

a) The marginal pdf of $\tilde{b}$ equals

$$f_{\tilde{b}}(b) = \int_{a=0}^{2} f_{\tilde{a}}(a) f_{\tilde{b} \mid \tilde{a}}(b \mid a) \, \mathrm{d}a = \begin{cases} \int_{a=0}^{1} \frac{1}{2} \, \mathrm{d}a + \int_{a=1}^{2} \frac{1}{2} \cdot \frac{1}{2} \, \mathrm{d}a = \frac{3}{4}, & \text{if } 0 \leq b \leq 1, \\ \int_{a=1}^{2} \frac{1}{2} \cdot \frac{1}{2} \, \mathrm{d}a = \frac{1}{4}, & \text{if } 1 \leq b \leq 2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.13}$$

If $0 \leq b \leq 1$,

$$f_{\tilde{a} \mid \tilde{b}}(a \mid b) = \frac{f_{\tilde{a}}(a) f_{\tilde{b} \mid \tilde{a}}(b \mid a)}{f_{\tilde{b}}(b)} = \begin{cases} \frac{1/2}{3/4} = \frac{2}{3}, & \text{if } 0 \leq a \leq 1, \\ \frac{1/2 \cdot 1/2}{3/4} = \frac{1}{3}, & \text{if } 1 \leq a \leq 2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.14}$$

If $1 \leq b \leq 2$,

$$f_{\tilde{a} \mid \tilde{b}}(a \mid b) = \frac{f_{\tilde{a}}(a) f_{\tilde{b} \mid \tilde{a}}(b \mid a)}{f_{\tilde{b}}(b)} = \begin{cases} \frac{1/2 \cdot 1/2}{1/4} = 1, & \text{if } 1 \leq a \leq 2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.15}$$

b) We first sample from the marginal pdf of $\tilde{a}$. The cdf of $\tilde{a}$ equals

$$F_{\tilde{a}}(a) = \int_{x=0}^{a} f_{\tilde{a}}(a) \, \mathrm{d}a \tag{5.16}$$

$$= \frac{a}{2}, \tag{5.17}$$

so its inverse is $F_{\tilde{a}}^{-1}(u) = 2u$. The sample a from $\tilde{a}$ can therefore be obtained by setting:

$$a := 2u_1. \tag{5.18}$$

In our example, that yields $a = 1.4$.

Then, we sample from the conditional distribution of $\tilde{b}$ given $\tilde{a} = a$, where a is the sample from $\tilde{a}$. If $0 \leq 1 \leq a$, then we just set $b := u_2$ because the conditional pdf of $\tilde{b}$ given $\tilde{a} = a$ is uniform in $[0, 1]$. If $1 \leq a \leq 2$, as is the case for our example, the conditional pdf of $\tilde{b}$ given $\tilde{a} = a$ is uniform in $[0, 2]$, just like the marginal distribution of $\tilde{a}$, so we set $b := 2u_2$, which for our example yields $b = 0.4$.

5.4 (Piecewise-constant density)

a) If $0 < a \leq 1$, the marginal pdf of $\tilde{a}$ at a equals

$$f_{\tilde{a}}(a) = \int_{b=0}^{2} f_{\tilde{a},\tilde{b}}(a,b) \, \mathrm{d}b \tag{5.19}$$

$$= \int_{b=0}^{1} \frac{1}{10} \, \mathrm{d}b + \int_{b=1}^{2} \frac{3}{10} \, \mathrm{d}b \tag{5.20}$$

$$= \frac{2}{5}. \tag{5.21}$$

If $1 < a \leq 2$,

$$f_{\tilde{a}}(a) = \int_{b=0}^{2} f_{\tilde{a},\tilde{b}}(a,b)\ \mathrm{d}b \tag{5.22}$$

$$= \int_{b=0}^{1} \frac{1}{10}\ \mathrm{d}b + \int_{b=1}^{2} \frac{5}{10}\ \mathrm{d}b \tag{5.23}$$

$$= \frac{3}{5}. \tag{5.24}$$

For $a \notin [0,2]$ $f_{\tilde{a},\tilde{b}}(a,b) = 0$, so the marginal pdf equals zero.

b) If $0 < b \leq 1$, the marginal cdf of $\tilde{b}$ at b equals

$$F_{\tilde{b}}(b) = \mathrm{P}\left(\tilde{b} \leq b\right) \tag{5.25}$$

$$= \int_{a=0}^{2} \int_{b=0}^{b} f_{\tilde{a},\tilde{b}}(a,b)\ \mathrm{d}a\ \mathrm{d}b \tag{5.26}$$

$$= \int_{a=0}^{1} \int_{b=0}^{b} \frac{1}{10}\ \mathrm{d}a\ \mathrm{d}b + \int_{a=1}^{2} \int_{b=0}^{b} \frac{1}{10}\ \mathrm{d}a\ \mathrm{d}b \tag{5.27}$$

$$= \frac{b}{5}. \tag{5.28}$$

If $1 < b \leq 2$,

$$F_{\tilde{b}}(b) = \mathrm{P}\left(\tilde{b} \leq b\right) \tag{5.29}$$

$$= \int_{a=0}^{2} \int_{b=0}^{b} f_{\tilde{a},\tilde{b}}(a,b)\ \mathrm{d}a\ \mathrm{d}b \tag{5.30}$$

$$= \int_{a=0}^{1} \int_{b=0}^{1} \frac{1}{10}\ \mathrm{d}a\ \mathrm{d}b + \int_{a=0}^{1} \int_{b=0}^{1} \frac{1}{10}\ \mathrm{d}a\ \mathrm{d}b \tag{5.31}$$

$$+ \int_{a=1}^{2} \int_{b=0}^{1} \frac{3}{10}\ \mathrm{d}a + \int_{a=1}^{2} \int_{b=0}^{1} \frac{5}{10}\ \mathrm{d}a \tag{5.32}$$

$$= \frac{1}{5} + \frac{4(b-1)}{5} = \frac{4b-3}{5}. \tag{5.33}$$

For $b < 0$ $F_{\tilde{a},\tilde{b}}(a,b) = 0$ and for $b > 2$ $F_{\tilde{a},\tilde{b}}(a,b) = 1$.

c) The conditional pdf of $\tilde{b}$ given $\tilde{a} = 1.6$ when $0 < b \leq 1$ equals,

$$f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,1.6) = \frac{f_{\tilde{a},\tilde{b}}(1.6,b)}{f_{\tilde{a}}(1.6)} \tag{5.34}$$

$$= \frac{\frac{1}{10}}{\frac{3}{5}} = \frac{1}{6}, \tag{5.35}$$

so the desired probability is

$$\mathrm{P}\left(\tilde{b} < 0.5\,|\,\tilde{a} = 1.6\right) = \int_{b=0}^{0.5} \frac{1}{6}\ \mathrm{d}b \tag{5.36}$$

$$= \frac{1}{12}. \tag{5.37}$$

d) We first obtain a sample from $\tilde{b}$ using inverse transform sampling. Since $u_1 = 0.1$, this yields $b = F_{\tilde{b}}^{-1}(u_1) = 0.5$. Then we obtain a sample from $\tilde{a}$ applying inverse transform sampling to the conditional cdf of $\tilde{a}$ given $\tilde{b} = 0.5$. Differentiating (5.28)

yields $f_{\tilde{b}}(0.5) = 1/5$, so the conditional pdf of $\tilde{a}$ given $\tilde{b} = 0.5$ equals

$$f_{\tilde{a}\,|\,\tilde{b}}(a\,|\,0.5) = \frac{f_{\tilde{a},\tilde{b}}(a, 0.5)}{f_{\tilde{b}}(0.5)} \tag{5.38}$$

$$= \frac{1}{2} \tag{5.39}$$

for $0 < a \leq 2$ (and 0 otherwise). The conditional cdf therefore equals $a/2$. We conclude that the second sample is $a = F_{\tilde{a}}^{-1}(u_2) = 1.4$.

5.5 (Samples and independence) In (1) and (2) the density of the samples in the vertical direction for $0 \leq a \leq 1$ is completely different to the density for $1 \leq a \leq 2$, so the random variables are dependent. In (3) the density of the samples in the vertical direction looks the same for all values of a between 0 and 1, and is zero otherwise. This suggests that the conditional pdf of $\tilde{b}$ given $\tilde{a} = a$ is the same for all values of a, which would imply that the random variables are independent.

5.6 (Independence of continuous random variables)

a) True. By the assumption we can define a function g such that $g(y) := f_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x)$ for all x and y. Marginalizing the joint pdf we obtain

$$f_{\tilde{y}}(y) = \int_{x=-\infty}^{\infty} f_{\tilde{x},\tilde{y}}(x, y)\ \mathrm{d}x \tag{5.40}$$

$$= \int_{x=-\infty}^{\infty} f_{\tilde{x}}(x) f_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x)\ \mathrm{d}x \tag{5.41}$$

$$= \int_{x=-\infty}^{\infty} f_{\tilde{x}}(x) g(y)\ \mathrm{d}x \tag{5.42}$$

$$= g(y) \int_{x=-\infty}^{\infty} f_{\tilde{x}}(x)\ \mathrm{d}x \tag{5.43}$$

$$= g(y) = f_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x), \tag{5.44}$$

for any x, so the random variables are independent.

b) True. At points like $x = 2$ and $y = 8$ we have $f_{\tilde{x},\tilde{y}}(x, y) = 0$ but both $f_{\tilde{x}}(x)$ and $f_{\tilde{y}}(y)$ are nonzero (the joint pdf is nonzero on the lines $x = 2$ and $y = 8$, so it will integrate to nonzero values when we marginalize). This implies that $f_{\tilde{x},\tilde{y}}(x, y) = f_{\tilde{x}}(x) f_{\tilde{y}}(y)$ cannot hold.

c) False. If the variables are independent then

$$f_{\tilde{x},\tilde{y}}(x_1, y) = f_{\tilde{x}}(x_1) f_{\tilde{y}}(y), \tag{5.45}$$

$$f_{\tilde{x},\tilde{y}}(x_2, y) = f_{\tilde{x}}(x_2) f_{\tilde{y}}(y), \tag{5.46}$$

so $f_{\tilde{x},\tilde{y}}(x_1, y) \neq f_{\tilde{x},\tilde{y}}(x_2, y)$ can hold for all y as long as $f_{\tilde{x}}(x_1) \neq f_{\tilde{x}}(x_2)$. For a concrete example, take $\tilde{y}$ uniform between 0 and 1, and $\tilde{x}$ exponential with parameter λ. If they are independent, for any $y \in [0, 1]$,

$$f_{\tilde{x},\tilde{y}}(1, y) = \lambda \exp(-\lambda) \tag{5.47}$$

$$\neq \lambda \exp(-2\lambda) \tag{5.48}$$

$$= f_{\tilde{x},\tilde{y}}(2, y). \tag{5.49}$$

5.7 (Regression to the mean)

a) Let q denote the xth percentile of the distribution of $\tilde{a}$ and $\tilde{b}$. By independence between

$\tilde{a}$ and $\tilde{b}$, the assumption that they have the same distribution and the definition of percentile,

$$\begin{aligned} \mathrm{P}\left(\tilde{b} \leq \tilde{a} \,|\, \tilde{a} = q\right) &= \mathrm{P}\left(\tilde{b} \leq q \,|\, \tilde{a} = q\right) && (5.50)\\ &= \mathrm{P}\left(\tilde{b} \leq q\right) && (5.51)\\ &= x\%. && (5.52) \end{aligned}$$

In particular if x is equal to 95, the probability that $\tilde{b}$ is smaller or equal to $\tilde{a}$ is 0.95.

b) The previous answer shows that even if two measurements are completely independent, and are generated by the exact same distribution, the more extreme the first measurement is, the more likely it is for the second to be less extreme.

c) The value of $\tilde{a}$ has no effect on the distribution of $\tilde{b}$. In fact we use this for our proof! It only affects the probability we are computing, because we are comparing $\tilde{b}$ to $\tilde{a}$.

5.8 (Frog)

a) From the assumptions we know that the density is zero outside of the ponds, equal to a constant c_L in the large pond and to another constant c_S in the small pond. The probability of the frog being in the large pond is 1/4 so

$$\begin{aligned} P\left(\text{large pond}\right) &= \int_0^{10}\int_0^{10} c_L \,\mathrm{d}x\,\mathrm{d}y && (5.53)\\ &= 100c_L = \frac{1}{4}, && (5.54) \end{aligned}$$

which implies $c_L = \frac{1}{400}$. Similarly

$$\begin{aligned} P\left(\text{small pond}\right) &= \int_{12}^{17}\int_0^{5} c_S \,\mathrm{d}x\,\mathrm{d}y && (5.55)\\ &= 25c_S = \frac{3}{4}, && (5.56) \end{aligned}$$

which implies $c_S = \frac{3}{100}$. Therefore,

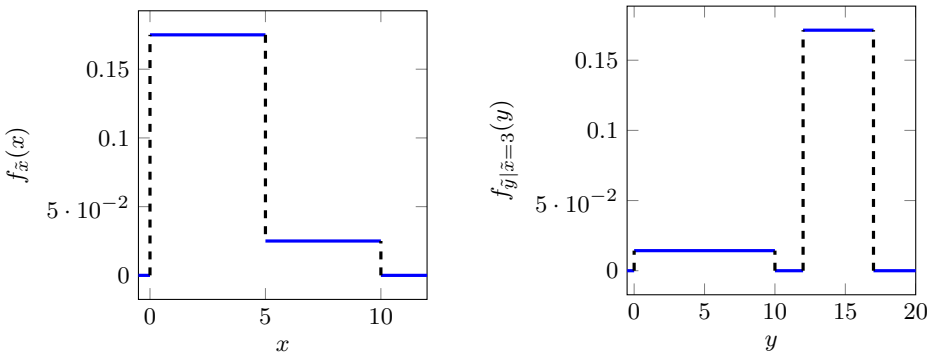
$$f_{\tilde{x},\tilde{y}}(x,y) = \begin{cases} \frac{1}{400} & \text{if } 0 \leq x, y \leq 10,\\ \frac{3}{100} & \text{if } 0 \leq x \leq 5, 12 \leq y \leq 17,\\ 0 & \text{otherwise.} \end{cases} \quad (5.57)$$

b) The marginal pdf of the horizontal position of the frog is

$$f_{\tilde{x}}(x) = \int_0^{17} f_{\tilde{x},\tilde{y}}(x,y)dy = \begin{cases} \int_0^{10} \frac{1}{400}dy + \int_{12}^{17} \frac{3}{100}dy = \frac{7}{40} & \text{if } 0 \leq x \leq 5,\\ \int_0^{10} \frac{1}{400}dy = \frac{1}{40} & \text{if } 5 \leq x \leq 10,\\ 0 & \text{otherwise.} \end{cases} \quad (5.58)$$

c)

$$\begin{aligned} f_{\tilde{y}|\tilde{x}=3}(y) = \frac{f(3,y)}{f_{\tilde{x}}(3)} &= \frac{40}{7}\begin{cases} \frac{1}{400} & \text{if } 0 \leq y \leq 10\\ \frac{3}{100} & \text{if } 12 \leq y \leq 17\\ 0 & \text{otherwise} \end{cases} && (5.59)\\ &= \begin{cases} \frac{1}{70} & \text{if } 0 \leq y \leq 10\\ \frac{6}{35} & \text{if } 12 \leq y \leq 17\\ 0 & \text{otherwise} \end{cases} && (5.60) \end{aligned}$$



d) The vertical position of the frog is not independent from the horizontal position since

$$f_{\tilde{y}}(13) = \int_0^5 f_{\tilde{x},\tilde{y}}\,(x, y)\;\mathrm{d}x \tag{5.61}$$

$$= 5\frac{3}{100} = \frac{3}{20} \neq \frac{6}{35} = f_{\tilde{y}|\tilde{x}=3}(13). \tag{5.62}$$

e) Conditioned on the frog being in the small pond the joint pdf is a constant c such that

$$P\,(\text{small pond}) = \int_{12}^{17}\int_0^5 c\;\mathrm{d}x\;\mathrm{d}y \tag{5.63}$$

$$= 25c = 1, \tag{5.64}$$

so $c = \frac{1}{25}$. The joint pdf is consequently

$$f_{\tilde{x},\tilde{y}|\text{small pond}}\,(x, y) = \begin{cases} \frac{1}{25} & \text{if } 0 \leq x \leq 5 \text{ and } 12 \leq y \leq 17, \\ 0 & \text{otherwise.} \end{cases} \tag{5.65}$$

We marginalize to find the pdfs of $\tilde{x}$ and $\tilde{y}$.

$$f_{\tilde{x}|\text{small pond}}\,(x) = \int_{12}^{17} f_{\tilde{x},\tilde{y}|\text{small pond}}\,(x, y)\;\mathrm{d}y = \begin{cases} \frac{1}{5} & \text{if } 0 \leq x \leq 5, \\ 0 & \text{otherwise.} \end{cases} \tag{5.66}$$

$$f_{\tilde{y}|\text{small pond}}\,(y) = \int_0^5 f_{\tilde{x},\tilde{y}|\text{small pond}}\,(x, y)\;\mathrm{d}x = \begin{cases} \frac{1}{5} & \text{if } 12 \leq y \leq 17, \\ 0 & \text{otherwise.} \end{cases} \tag{5.67}$$

The two variables are conditionally independent because $f_{\tilde{x}|\text{small pond}}\,(x)\,f_{\tilde{y}|\text{small pond}}\,(y) = f_{\tilde{x},\tilde{y}|\text{small pond}}\,(x, y)$ for any values of x and y.

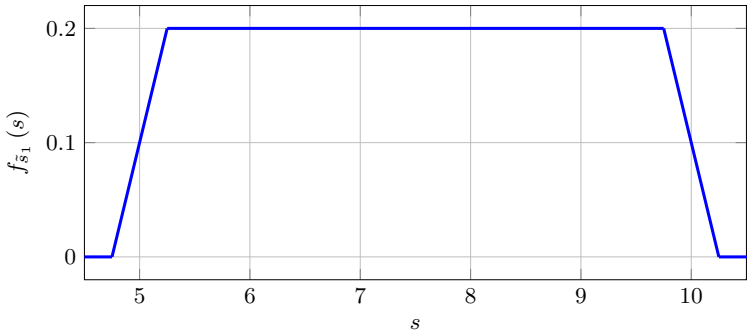
5.9 (Sonar)

a) Note that for fixed x $f_{\tilde{s}_1 \mid \tilde{x}}(s \mid x) = 0$ if $x < s - 0.25$ or $x > s + 0.25$. This implies

$$f_{\tilde{s}_1}(s) = \int_5^{10} f_{\tilde{x}}(x) f_{\tilde{s}_1 \mid \tilde{x}}(s \mid x) \, \mathrm{d}x \tag{5.68}$$

$$= \int_{\max\{5, s-0.25\}}^{\min\{10, s+0.25\}} \frac{2}{5} \, \mathrm{d}x \tag{5.69}$$

$$= \begin{cases} 0 & \text{if } s < 4.75 \\ \frac{2(s-4.75)}{5} & \text{if } 4.75 \leq s \leq 5.25 \\ \frac{1}{5} & \text{if } 5.25 \leq s \leq 9.75 \\ \frac{2(10.25-s)}{5} & \text{if } 9.75 \leq s \leq 10.25 \\ 0 & \text{if } x > 10.25. \end{cases} \tag{5.70}$$



b) For fixed s_1 $f_{\tilde{s}_1 \mid \tilde{x}}(s_1 \mid x) = 0$ if $x < s_1 - 0.25$ or $x > s_1 + 0.25$. For fixed s_2 $f_{\tilde{s}_2 \mid \tilde{x}}(s_2 \mid x) = 0$ if $x < s_2 - 0.25$ or $x > s_2 + 0.25$, so

$$f_{\tilde{s}_1, \tilde{s}_2}(7, 7.1) = \int_5^{10} f_{\tilde{x}}(x) f_{\tilde{s}_1 \mid \tilde{x}}(7 \mid x) f_{\tilde{s}_2 \mid \tilde{x}}(7.1 \mid x) \, \mathrm{d}x \tag{5.71}$$

$$= \int_{6.85}^{7.25} \frac{4}{5} \, \mathrm{d}x \tag{5.72}$$

$$= 0.32. \tag{5.73}$$

$$f_{\tilde{x} \mid \tilde{s}_1, \tilde{s}_2}(x \mid 7, 7.1) = \frac{f_{\tilde{x}, \tilde{s}_1, \tilde{s}_2}(x, 7, 7.1)}{f_{\tilde{s}_1, \tilde{s}_2}(7, 7.1)} \tag{5.74}$$

$$= \begin{cases} \frac{4/5}{0.32} = 2.5 & \text{if } 6.85 \leq x \leq 7.25, \\ 0 & \text{otherwise.} \end{cases} \tag{5.75}$$

c) For fixed s_1 $f_{\tilde{s}_1 \mid \tilde{x}}(s_1 \mid x) = 0$ if $x < s_1 - 0.25$ or $x > s_1 + 0.25$. For fixed s_2 $f_{\tilde{s}_2 \mid \tilde{x}}(s_2 \mid x) = 0$ if $x < s_2 - 0.25$ or $x > s_2 + 0.25$, so

$$f_{\tilde{s}_1, \tilde{s}_2}(s_1, s_2) = \int_5^{10} f_{\tilde{x}}(x) f_{\tilde{s}_1 \mid \tilde{x}}(s_1 \mid x) f_{\tilde{s}_2 \mid \tilde{x}}(s_2 \mid x) \, \mathrm{d}x \tag{5.76}$$

$$= \int_{\max\{5, s_1-0.25, s_2-0.25\}}^{\min\{10, s_1+0.25, s_2+0.25\}} \frac{4}{5} \, \mathrm{d}x \tag{5.77}$$

$$= 0.8 \left(\min\{10, s_1 + 0.25, s_2 + 0.25\} - \max\{5, s_1 - 0.25, s_2 - 0.25\} \right)$$

if $4.75 \leq s_1, s_2 \leq 10.25$ and $|s_2 - s_1| \leq 0.5$, and 0 otherwise.

The two measurements are not independent. For example $f_{\tilde{s}_1}(5) \neq 0$ and $f_{\tilde{s}_2}(10) \neq 0$ but $f_{\tilde{s}_1,\tilde{s}_2}(5, 10) = 0$ so the joint pdf is not the product of the marginals. This makes sense since $\tilde{s}_1$ provides information about $\tilde{x}$, which in turn provides information about $\tilde{s}_2$.

5.10 (Rufus)

a) By assumption, the joint probability density is equal to a constant c. It must integrate to one, so

$$\int_{\text{garden}} c \, \mathrm{d}x \, \mathrm{d}y = c \, \text{Area}\,(\text{garden}) = 1. \tag{5.78}$$

The area of the garden is equal to $100^2 - 40^2 = 10000 - 1600 = 8400$, which implies

$$f_{\tilde{x},\tilde{y}}(x, y) = \begin{cases} \frac{1}{8400} & \text{if } \{(x, y) : -50 \leq x \leq 50, \, -50 \leq y \leq 50\} \,/\, \{(x, y) : -20 \leq x \leq 20, \, -20 \leq y \leq 20\}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.79}$$

b) For y in $[-50, -20]$ or $[20, 50]$

$$f_{\tilde{y}}(y) = \int_{x=-\infty}^{\infty} f_{\tilde{x},\tilde{y}}(x, y) \, \mathrm{d}x \tag{5.80}$$

$$= \int_{x=-50}^{50} \frac{1}{8400}(x, y) \, \mathrm{d}x \tag{5.81}$$

$$= \frac{50 + 50}{8400} = \frac{1}{84}. \tag{5.82}$$

For y in $[-20, 20]$

$$f_{\tilde{y}}(y) = \int_{x=-\infty}^{\infty} f_{\tilde{x},\tilde{y}}(x, y) \, \mathrm{d}x \tag{5.83}$$

$$= \int_{x=-50}^{-20} \frac{1}{8400}(x, y) \, \mathrm{d}x + \int_{x=20}^{50} \frac{1}{8400}(x, y) \, \mathrm{d}x \tag{5.84}$$

$$= 2 \cdot \frac{50 - 20}{8400} = \frac{1}{140}. \tag{5.85}$$

For other values of y, $f_{\tilde{y}}(y) = 0$. The pdf is shown in Figure 5.1.

c) In order to compute the conditional pdf of $\tilde{x}$ given $\tilde{y}$ we need to compute the marginal pdf of $\tilde{y}$ by marginalizing over $\tilde{x}$. The conditional density is only defined for values of y such that $f_{\tilde{y}}(y) \neq 0$. For y in $[-50, -20]$ or $[20, 50]$

$$f_{\tilde{x}|\tilde{y}}(x|y) = \frac{f_{\tilde{x},\tilde{y}}(x, y)}{f_{\tilde{y}}(y)} = \begin{cases} \frac{1}{100} & \text{if } -50 \leq x \leq 50, \\ 0 & \text{otherwise.} \end{cases} \tag{5.86}$$

For y in $[-20, 20]$

$$f_{\tilde{x}|\tilde{y}}(x|y) = \begin{cases} \frac{1}{60} & \text{if } -50 \leq x \leq -20 \text{ or } 20 \leq x \leq 50, \\ 0 & \text{otherwise.} \end{cases} \tag{5.87}$$

The conditional pdfs are plotted in Figure 5.2.

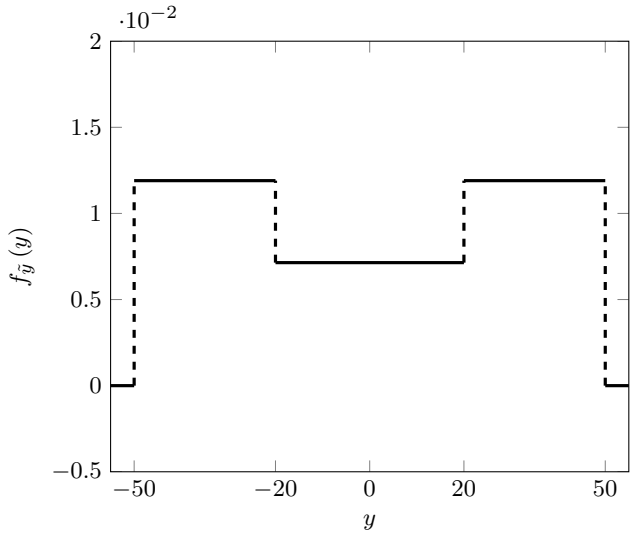


Figure 5.1 Pdf of the vertical position $\tilde{y}$ of Rufus.

d) $\tilde{x}$ and $\tilde{y}$ are not independent. By (5.85) $f_{\tilde{y}}(0) = \frac{60}{8400} \neq 0$ and by symmetry $f_{\tilde{x}}(0) = \frac{60}{8400} \neq 0$. However $(0, 0)$ is not in the garden, so

$$f_{\tilde{x},\tilde{y}}(0,0) = 0 \neq f_{\tilde{x}}(0) f_{\tilde{y}}(0). \tag{5.88}$$

e) The cdf of $\tilde{y}$ equals

$$F_{\tilde{y}}(y) = \begin{cases} 0 & \text{for } y < -50, \\ \frac{y-(-50)}{84} = \frac{y+50}{84} & \text{for } -50 \leq y < -20, \\ \frac{30}{84} + \frac{y-(-20)}{140} = \frac{5}{14} + \frac{y+20}{140} & \text{for } -20 \leq y < 20, \\ \frac{30}{84} + \frac{40}{140} + \frac{y-20}{84} = \frac{9}{14} + \frac{y-20}{84} & \text{for } 20 \leq y < 50, \\ 1 & \text{for } y > 50. \end{cases} \tag{5.89}$$

We obtain a sample of $\tilde{y}$ by applying the inverse transform method. Since $0.1 < 5/14$, we set

$$F_{\tilde{y}}(y) = \frac{y + 50}{84} = 0.1, \tag{5.90}$$

which yields $y = 8.4 - 50 = -41.6$. Now we consider the conditional cdf of $\tilde{x}$ given $\tilde{y} = -41.6$, which is equal to

$$F_{\tilde{x}|\tilde{y}}(x \mid -41.6) = \begin{cases} 0 & \text{for } x < -50, \\ \frac{x+50}{100} & \text{for } -50 \leq x \leq 50, \\ 1 & \text{for} x > 50. \end{cases} \tag{5.91}$$

To obtain the sample from x we set $F_{\tilde{x}|\tilde{y}}(x \mid -41.6) = 0.3$, which yields $x = -20$.

5.11 (Spider on a wall)

a) Let $\tilde{x}$ be the horizontal position and $\tilde{y}$ be the vertical position, or height. We have the

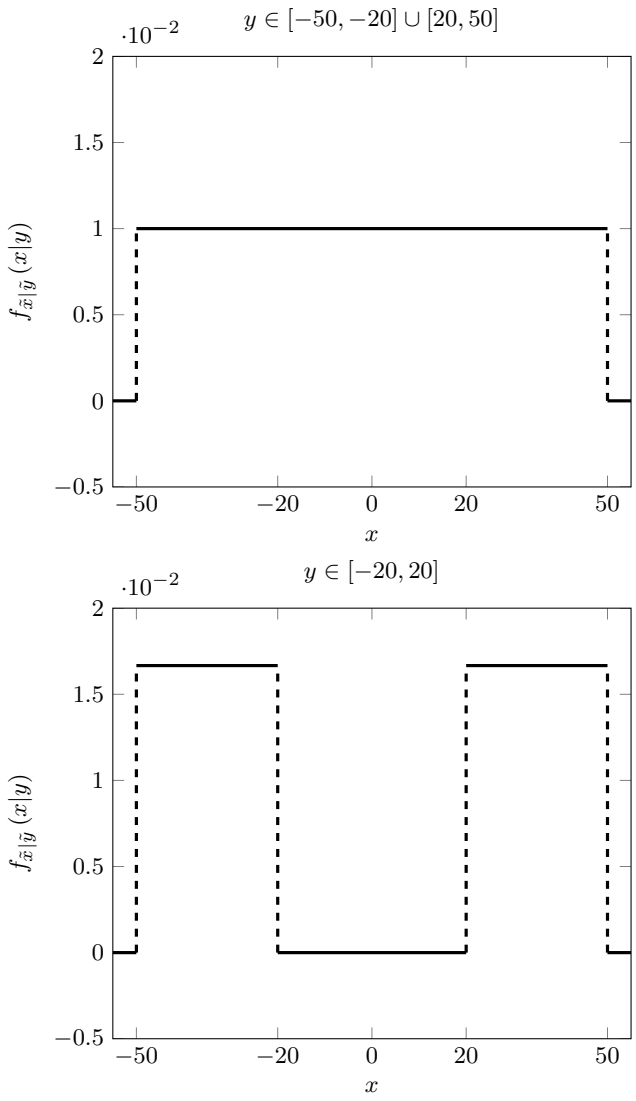


Figure 5.2 Conditional pdf of the position $\tilde{x}$ of Rufus given his $\tilde{y}$ position.

following two equations:

$$\mathrm{P}\left(\text{behind the painting}\right) = 2\mathrm{P}\left(\text{not behind the painting}\right), \tag{5.92}$$

$$\mathrm{P}\left(\text{behind the painting}\right) + \mathrm{P}\left(\text{not behind the painting}\right) = 1. \tag{5.93}$$

Under the assumptions, the probability density is constant in the region behind the painting and also in the rest of the wall. We denote these two densities by c_1 and c_2 respectively. Since the painting has area 4 and the rest of the wall has area 96, the

equations imply

$$4c_1 = 2 \cdot 96c_2, \tag{5.94}$$

$$4c_1 + 96c_2 = 1. \tag{5.95}$$

This implies $c_1 = 1/6$ and $c_2 = 1/288$, so

$$f_{(\tilde{x},\tilde{y})}(x,y) = \begin{cases} 1/6 \text{ if } (x,y) \in \text{painting} \\ 1/288 \text{ if } (x,y) \notin \text{painting} \end{cases} \tag{5.96}$$

b) The marginal distribution of $\tilde{y}$ is supported on $(0, 10)$:

$$f_{\tilde{y}}(y) = \int f_{(\tilde{x},\tilde{y})}(x,y)dx = \begin{cases} \int_0^{10} \frac{1}{288} dx = 5/144 & \text{if } y \notin (6,8)\ , \\ \int_{(0,4)\cup(6,10)} \frac{1}{288} dx + \int_{(4,6)} 1/6 dx = 13/36 & \text{if } y \in (6,8)\ , \\ 0 & \text{otherwise.} \end{cases}$$

c) There are possible three cases, the spider is either below the paint, above the paint or at the paint level:

$$F_{\tilde{y}|(\tilde{x},\tilde{y})\notin(4,6)\times(6,8)}(y) = \mathrm{P}\left(\tilde{y} < y | (\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right) \tag{5.97}$$

$$= \frac{\mathrm{P}\left(\tilde{y} < y, (\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right)}{\mathrm{P}\left((\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right)} \tag{5.98}$$

$$= \begin{cases} 30y/288 & \text{if } y \in (0,6)\ , \\ 180/288 + 24(y-6)/288 & \text{if } y \in (6,8)\ , \\ 228/288 + 30(y-8)/288 & \text{if } y \in (8,10)\ . \end{cases} \tag{5.99}$$

We have used the following intermediate results.

$$\mathrm{P}\left((\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right) = 1 - \mathrm{P}\left((\tilde{x},\tilde{y}) \in (4,6)\times(6,8)\right) \tag{5.100}$$

$$= 1 - \int_{x=4}^{6}\int_{y=6}^{8} f_{(\tilde{x},\tilde{y})}(x,y)\ \mathrm{d}x\ \mathrm{d}y \tag{5.101}$$

$$= 1 - \int_{x=4}^{6}\int_{y=6}^{8} \frac{1}{6}\ \mathrm{d}x\ \mathrm{d}y \tag{5.102}$$

$$= \frac{1}{3}. \tag{5.103}$$

If $y \in (0,6)$

$$\mathrm{P}\left(\tilde{y} < y, (\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right) = \mathrm{P}\left(\tilde{y} < y\right) \tag{5.104}$$

$$= \int_{u=0}^{y} f_{\tilde{y}}(u)\ \mathrm{d}u \tag{5.105}$$

$$= \frac{5y}{144}. \tag{5.106}$$

If $y \in (6,8)$

$$\mathrm{P}\left(\tilde{y} < y, (\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right) = \mathrm{P}\left(\tilde{y} < 6\right) + \mathrm{P}\left(6 \le \tilde{y} < y, (\tilde{x},\tilde{y}) \notin (4,6)\times(6,8)\right)$$

$$= \frac{30}{144} + \int_{x=0}^{4}\int_{u=6}^{y} \frac{1}{288}\ \mathrm{d}x\ \mathrm{d}y + \int_{x=6}^{10}\int_{u=6}^{y} \frac{1}{288}\ \mathrm{d}x\ \mathrm{d}u$$

$$= \frac{30}{144} + \frac{8\,(y-6)}{288}. \tag{5.107}$$

If $y \in (8, 10)$

$$\begin{aligned}
\mathrm{P}\left(\tilde{y} < y, (\tilde{x}, \tilde{y}) \notin (4,6) \times (6,8)\right) &= \mathrm{P}\left(\tilde{y} < 6\right) + \mathrm{P}\left(6 \leq \tilde{y} < 8, (\tilde{x}, \tilde{y}) \notin (4,6) \times (6,8)\right) \\
&\quad + \mathrm{P}\left(8 \leq \tilde{y} < y\right) && (5.108) \\
&= \frac{30}{144} + \frac{16}{288} + \int_{u=8}^{y} f_{\tilde{y}}(u) \, \mathrm{d}u && (5.109) \\
&= \frac{76}{288} + \frac{5\,(y-8)}{144}. && (5.110)
\end{aligned}$$

5.12 (Simulating a constant density) There are different ways of generating the samples. Arguably the easiest is to generate two independent uniform samples u_1, u_2 and only use them if the vector $\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right)$ is in the region of interest, which we denote by $\mathcal{R}$. Let $\tilde{u}_1$ and $\tilde{u}_2$ denote two independent samples from a uniform distribution in $[0, 1]$. The cdf of the resulting samples is

$$\begin{aligned}
F_{\tilde{u}_1, \tilde{u_2} \,|\, \left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R}}(u_1, u_2) &= \mathrm{P}(\tilde{u}_1 \leq u_1, \tilde{u_2} \leq u_2 \,|\, \left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R}) && (5.111) \\
&= \frac{\mathrm{P}(\tilde{u}_1 \leq u_1, \tilde{u_2} \leq u_2, \left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R})}{\mathrm{P}(\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R})} && (5.112) \\
&= \frac{F_{\tilde{u}_1, \tilde{u_2}}(u_1, u_2)}{\mathrm{P}(\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R})} && (5.113)
\end{aligned}$$

as long as $\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R}$. Differentiating, we obtain

$$\begin{aligned}
f_{\tilde{u}_1, \tilde{u_2} \,|\, \left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R}}(u_1, u_2) &= \frac{f_{\tilde{u}_1, \tilde{u_2}}(u_1, u_2)}{\mathrm{P}(\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R})} && (5.114) \\
&= \frac{f_{\tilde{u}_1}(u_1) f_{\tilde{u_2}}(u_2)}{\mathrm{P}(\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R})} && (5.115) \\
&= \frac{1}{\mathrm{P}(\left(\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right) \in \mathcal{R})}. && (5.116)
\end{aligned}$$

The resulting samples therefore have constant density in the region of interest.

5.13 (Multivariate Gaussian pdf) We denote the multivariate Gaussian pdf by $f_{\tilde{x}}$ and its mean vector by $\mu \in \mathbb{R}^d$.

a) Let σ_i^2 be the ith diagonal entry of the covariance matrix Σ for $1 \leq i \leq d$. If the covariance matrix is diagonal, the joint pdf is equal to the product of the marginal pdfs,

$$\begin{aligned}
f_{\tilde{x}}\left(x\right) &= \frac{1}{\sqrt{\left(2\pi\right)^d \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1} \left(x-\mu\right)\right) && (5.117) \\
&= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(x[i]-\mu_i\right)^2}{2\sigma_i^2}\right). && (5.118)
\end{aligned}$$

Consequently, by Exercise 3.9

$$\begin{aligned}
\int_{x \in \mathbb{R}^d} f_{\tilde{x}}\left(x\right) \mathrm{d}x &= \prod_{i=1}^{d} \int_{x[i]=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\left(x[i]-\mu_i\right)^2}{2\sigma_i^2}\right) \mathrm{d}x[i] && (5.119) \\
&= \prod_{i=1}^{d} 1 = 1. && (5.120)
\end{aligned}$$

b) Using a non-diagonal covariance matrix just rotates the joint pdf with respect to the standard axes. Consequently, its integral should not change and must therefore equal one.

More formally, by the spectral theorem, $\Sigma = U\Lambda U^T$ for an orthogonal $d \times d$ matrix U and a diagonal $d \times d$ matrix Λ. If we perform the change of variable $z = U^T(x - \mu)$, then $\mathrm{d}z = \mathrm{d}x$, since $|U| = 1$, because U is orthogonal. Consequently,

$$\int_{x\in\mathbb{R}^d} f_{\tilde{x}}(x)\,\mathrm{d}x = \int_{x\in\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\mathrm{d}x \tag{5.121}$$

$$= \int_{x\in\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d\,|U|\,|\Lambda|\,|U|}} \exp\left(-\frac{1}{2}(x-\mu)^T U\Lambda^{-1}U^T(x-\mu)\right)\mathrm{d}x$$

$$= \int_{z\in\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d\,|\Lambda|}} \exp\left(-\frac{1}{2}z^T\Lambda^{-1}z\right)\mathrm{d}z, \tag{5.122}$$

which equals one because Λ is diagonal, as established above.

5.14 (Exotic fruit)

a) The problem is that it runs into the curse of dimensionality. We need to condition on 4 features, but we only have 10 data points. Consequently it will be very difficult to find a relevant subset of the data, as the vast majority of possible combinations of feature values are unobserved.

b) The maximum-likelihood estimate of the mean parameter is

$$\mu_{\mathrm{ML}} = \frac{1}{10}\sum_{i=1}^{n} x_i = \begin{bmatrix} 1.53 \\ 16.5 \\ 10.6 \\ 98.6 \\ 15.4 \end{bmatrix}. \tag{5.123}$$

The maximum-likelihood estimate of the covariance-matrix parameter is

$$\Sigma_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathrm{ML}})(x_i - \mu_{\mathrm{ML}})^T \tag{5.124}$$

$$= \begin{bmatrix} 0.51 & -1.61 & 1.45 & -7.89 & 3.97 \\ -1.61 & 7.05 & -3.9 & 0.6 & -12.25 \\ 1.45 & -3.9 & 9.44 & -52.46 & 12.62 \\ -7.89 & 0.6 & -52.46 & 616.84 & -90.18 \\ 3.97 & -12.25 & 12.62 & -90.18 & 39.01 \end{bmatrix}. \tag{5.125}$$

We represent the parameters as follows, to separate the fruit weight and the rest of the features,

$$\mu_{\mathrm{ML}} = \begin{bmatrix} \mu_{\mathrm{fruit}} \\ \mu_{\mathrm{rest}} \end{bmatrix}, \tag{5.126}$$

$$\Sigma_{\mathrm{ML}} = \begin{bmatrix} \Sigma_{\mathrm{fruit}} & \Sigma_{\mathrm{fruit,rest}}^T \\ \Sigma_{\mathrm{fruit,rest}} & \Sigma_{\mathrm{rest}} \end{bmatrix}, \tag{5.127}$$

where

$$\mu_{\text{fruit}} = 1.53, \qquad \Sigma_{\text{fruit}} = 0.51, \tag{5.128}$$

$$\mu_{\text{rest}} = \begin{bmatrix} 1.53 \\ 16.5 \\ 10.6 \\ 98.6 \\ 15.4 \end{bmatrix}, \qquad \Sigma_{\text{rest}} = \begin{bmatrix} 7.05 & -3.9 & 0.6 & -12.25 \\ -3.9 & 9.44 & -52.46 & 12.62 \\ 0.6 & -52.46 & 616.84 & -90.18 \\ -12.25 & 12.62 & -90.18 & 39.01 \end{bmatrix}, \tag{5.129}$$

$$\Sigma_{\text{fruit,rest}} = \begin{bmatrix} -1.61 \\ 1.45 \\ -7.89 \\ 3.97 \end{bmatrix}. \tag{5.130}$$

The conditional pdf of the fruit weight given the rest of the features is Gaussian with mean

$$\mu_{\text{cond}} = \mu_{\text{fruit}} + \Sigma_{\text{fruit,rest}}^T \Sigma_{\text{rest}}^{-1} \left(\begin{bmatrix} 15 \\ 20 \\ 120 \\ 8 \end{bmatrix} - \mu_{\text{rest}} \right) \tag{5.131}$$

$$= 1.77, \tag{5.132}$$

and variance

$$\sigma^2_{\text{cond}} = \Sigma_{\text{cond}} \tag{5.133}$$

$$= \Sigma_{\text{fruit}} - \Sigma_{\text{fruit,rest}}^T \Sigma_{\text{rest}}^{-1} \Sigma_{\text{fruit,rest}} \tag{5.134}$$

$$= 0.037. \tag{5.135}$$

5.15 (Gaussian Bayesian model)

a) By the chain rule

$$f_{\tilde{\mu},\tilde{y}}(\mu, y) = f_{\tilde{\mu}}(\mu) f_{\tilde{y} \,|\, \tilde{\mu}}(y \,|\, \mu) \tag{5.136}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2} \right) \tag{5.137}$$

$$= \frac{1}{2\pi\sigma} \exp\left(-\frac{1}{2} \left(\frac{\mu^2}{\sigma^2} + \mu^2 - 2\mu y + y^2 \right) \right) \tag{5.138}$$

$$= \frac{1}{2\pi\sigma} \exp\left(-\frac{1}{2} \begin{bmatrix} \mu \\ y \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} \mu \\ y \end{bmatrix} \right), \tag{5.139}$$

where

$$\Sigma^{-1} = \begin{bmatrix} 1 + \frac{1}{\sigma^2} & -1 \\ -1 & 1 \end{bmatrix}. \tag{5.140}$$

This is the joint pdf of a Gaussian random vector with mean zero and covariance-matrix parameter equal to

$$\Sigma = \frac{1}{1 + \frac{1}{\sigma^2} - 1} \begin{bmatrix} 1 & 1 \\ 1 & 1 + \frac{1}{\sigma^2} \end{bmatrix} \tag{5.141}$$

$$= \begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & 1 + \sigma^2 \end{bmatrix}. \tag{5.142}$$

You can check that the determinant of Σ is equal to σ.

b) We apply Theorem 5.24 in the notes on multiple continuous random variables, setting $\tilde{a} := \tilde{y}$, $\tilde{b} := \tilde{\mu}$. We have $\sigma_{\tilde{a}}^2 = 1 + \sigma^2$, $\sigma_{\tilde{b}}^2 = \sigma^2$ and $\rho = \frac{\sigma^2}{\sigma_{\tilde{a}}\sigma_{\tilde{b}}} = \frac{\sigma}{\sqrt{1+\sigma^2}}$. By the theorem, conditioned on $\tilde{y} = y$, μ is Gaussian with mean

$$\mu_{\text{cond}} = \frac{\rho\sigma_{\tilde{b}} y}{\sigma_{\tilde{a}}} \tag{5.143}$$

$$= \frac{\sigma^2 y}{1+\sigma^2} \tag{5.144}$$

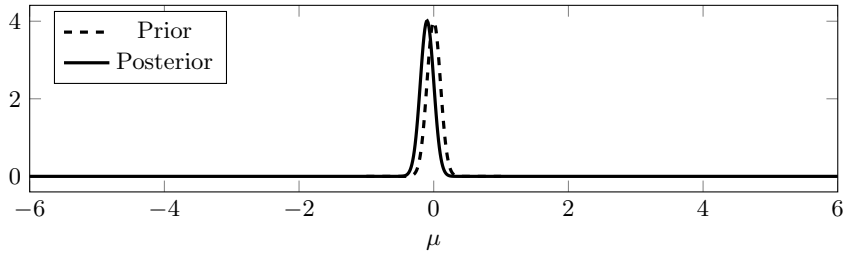
and variance

$$\sigma_{\text{cond}}^2 = (1-\rho^2)\sigma_{\tilde{b}}^2 \tag{5.145}$$

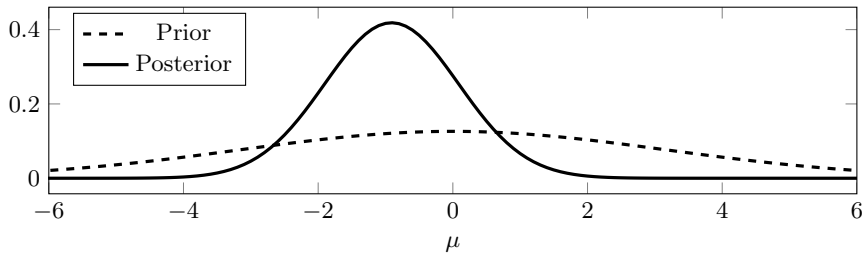
$$= \left(1 - \frac{\sigma^2}{1+\sigma^2}\right)\sigma^2 \tag{5.146}$$

$$= \frac{\sigma^2}{1+\sigma^2}. \tag{5.147}$$

c) When $y = -1$ and $\sigma^2 := 0.01$, $\mu_{\text{cond}} = -0.0099$ and $\sigma_{\text{cond}}^2 = 0.0099$. The scientist is very confident in the prior. The posterior and prior pdfs are very similar.



d) When $y = -1$ and $\sigma^2 := 10$, $\mu_{\text{cond}} = -0.909$ and $\sigma_{\text{cond}}^2 = 0.909$. The scientist is not confident in the prior. The posterior and prior pdfs are very different.



6

Discrete and Continuous Variables

Exercises

6.1 (Shared car)

a) Let $\tilde{a}$ be a random variable with range $\{1, 2, 3\}$. If $\tilde{a} = 1$ Carlos drives. If $\tilde{a} = 2$ Dani drives. If $\tilde{a} = 3$ Felix drives. We have

$$f_{\tilde{d}}(d) = \sum_{a=1}^{3} p_{\tilde{a}}(a) f_{\tilde{d}\,|\,\tilde{a}}(d\,|\,a) = \begin{cases} 0 & \text{if } d < 0, \\ \frac{1}{3}\left(\frac{1}{10} + \frac{1}{20} + \frac{1}{30}\right) = \frac{11}{180} & \text{if } 0 \leq d \leq 10, \\ \frac{1}{3}\left(\frac{1}{20} + \frac{1}{30}\right) = \frac{5}{180} & \text{if } 10 \leq d \leq 20, \\ \frac{1}{90} & \text{if } 20 \leq d \leq 30, \\ 0 & \text{if } d > 30. \end{cases} \tag{6.1}$$

b) By the chain rule of discrete and continuous random variables,

$$p_{\tilde{a}\,|\,\tilde{d}}(2\,|\,15) = \frac{p_{\tilde{a}}(2) f_{\tilde{d}\,|\,\tilde{a}}(15\,|\,2)}{f_{\tilde{d}}(15)} \tag{6.2}$$

$$= \frac{\frac{1}{3}\frac{1}{20}}{\frac{5}{180}} \tag{6.3}$$

$$= 0.6. \tag{6.4}$$

6.2 (Buckets)

a) It is easier to compute the probability that $\tilde{s}_1 \geq 1$, as this cannot happen if the bucket is small. Since $f_{\tilde{s}_1\,|\,\tilde{b}}(s\,|\,b) = 1/2$ between 0 and 2,

$$\mathrm{P}\left(\tilde{s}_1 \geq 1\right) = \mathrm{P}\left(\tilde{s}_1 \geq 1\,|\,\tilde{b} = 0\right)\mathrm{P}\left(\tilde{b} = 0\right) + \mathrm{P}\left(\tilde{s}_1 \geq 1\,|\,\tilde{b} = 1\right)\mathrm{P}\left(\tilde{b} = 1\right) \tag{6.5}$$

$$= \mathrm{P}\left(\tilde{s}_1 \geq 1\,|\,\tilde{b} = 1\right)\mathrm{P}\left(\tilde{b} = 1\right) \tag{6.6}$$

$$= \frac{1}{4}\int_{s=1}^{2} \frac{1}{2}\,\mathrm{d}s \tag{6.7}$$

$$= \frac{1}{8}. \tag{6.8}$$

We conclude that

$$\mathrm{P}\left(\tilde{s}_1 < 1\right) = 1 - \mathrm{P}\left(\tilde{s}_1 \geq 1\right) \tag{6.9}$$

$$= \frac{7}{8}. \tag{6.10}$$

b) Under the conditional independence assumption,

$$\begin{aligned}
&\mathrm{P}\left(\tilde{b}=1 \,|\, \tilde{s}_1=1/2, \tilde{s}_2=3/4\right) && (6.11)\\
&= \frac{p_{\tilde{b}}(1) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 1) f_{\tilde{s}_2 \,|\, \tilde{b}}(3/4 \,|\, 1)}{f_{\tilde{s}_1, \tilde{s}_2}(1/2, 3/4)} && (6.12)\\
&= \frac{p_{\tilde{b}}(1) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 1) f_{\tilde{s}_2 \,|\, \tilde{b}}(3/4 \,|\, 1)}{p_{\tilde{b}}(0) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 0) f_{\tilde{s}_2 \,|\, \tilde{b}}(3/4 \,|\, 0) + p_{\tilde{b}}(1) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 1) f_{\tilde{s}_2 \,|\, \tilde{b}}(3/4 \,|\, 1)} && (6.13)\\
&= \frac{\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4} \cdot 1 \cdot 1 + \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2}} && (6.14)\\
&= \frac{1}{13}. && (6.15)
\end{aligned}$$

c) Under the assumptions $\tilde{b}$ and $\tilde{s}_2$ are conditionally independent given $\tilde{s}_1$, so

$$\begin{aligned}
\mathrm{P}\left(\tilde{b}=1 \,|\, \tilde{s}_1=1/2, \tilde{s}_2=3/4\right) &= \mathrm{P}\left(\tilde{b}=1 \,|\, \tilde{s}_1=1/2\right) && (6.16)\\
&= \frac{p_{\tilde{b}}(1) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 1)}{f_{\tilde{s}_1}(1/2)} && (6.17)\\
&= \frac{p_{\tilde{b}}(1) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 1)}{p_{\tilde{b}}(0) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 0) + p_{\tilde{b}}(1) f_{\tilde{s}_1 \,|\, \tilde{b}}(1/2 \,|\, 1)} && (6.18)\\
&= \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{3}{4} \cdot 1 + \frac{1}{4} \cdot \frac{1}{2}} && (6.19)\\
&= \frac{1}{7}. && (6.20)
\end{aligned}$$

6.3 (Balls)

a) Let $\tilde{d}$ be a random variable representing the distance and $\tilde{h}$ a Bernoulli random variable indicating whether the ball is hollow ($\tilde{h}=1$) or not ($\tilde{h}=0$). The conditional pdf $f_{\tilde{d} \,|\, \tilde{h}}(d \,|\, 1)$ of $\tilde{d}$ given $\tilde{h}=1$ equals $1/5$ if d is in $[5, 10]$ and zero otherwise. The conditional pdf $f_{\tilde{d} \,|\, \tilde{h}}(d \,|\, 0)$ of $\tilde{d}$ given $\tilde{h}=0$ equals $1/5$ if d is in $[1, 6]$ and zero otherwise. Consequently, the marginal pdf of $\tilde{d}$ equals

$$\begin{aligned}
f_{\tilde{d}}(d) &= p_{\tilde{h}}(0) f_{\tilde{d} \,|\, \tilde{h}}(d \,|\, 0) + p_{\tilde{h}}(1) f_{\tilde{d} \,|\, \tilde{h}}(d \,|\, 1) && (6.21)\\
&= \begin{cases} \frac{1}{20} & \text{if } 1 \leq d < 5, \\ \frac{1}{5} & \text{if } 5 \leq d < 6, \\ \frac{3}{20} & \text{if } 6 \leq d < 10, \\ 0 & \text{otherwise.} \end{cases} && (6.22)
\end{aligned}$$

b) Let $\tilde{d}_1$ be the first distance and $\tilde{d}_2$ the second. Notice that the conditional pdf $f_{\tilde{d} \,|\, \tilde{h}}(d \,|\, 1)$ of $\tilde{d}$ given $\tilde{h}=1$ equals 0 if $d=3$. Consequently, under the conditional independence

assumptions,

$$f_{\tilde{d}_2 \,|\, \tilde{d}_1}(d \,|\, 3) = \frac{f_{\tilde{d}_1,\tilde{d}_2}(3,d)}{f_{\tilde{d}_1}(3)} \tag{6.23}$$

$$= \frac{p_{\tilde{h}}(0) f_{\tilde{d}_1 \,|\, \tilde{h}}(3 \,|\, 0) f_{\tilde{d}_2 \,|\, \tilde{h}}(d \,|\, 0) + p_{\tilde{h}}(1) f_{\tilde{d}_1 \,|\, \tilde{h}}(3 \,|\, 1) f_{\tilde{d}_2 \,|\, \tilde{h}}(d \,|\, 1)}{p_{\tilde{h}}(0) f_{\tilde{d}_1 \,|\, \tilde{h}}(3 \,|\, 0) + p_{\tilde{h}}(1) f_{\tilde{d}_1 \,|\, \tilde{h}}(3 \,|\, 1)} \tag{6.24}$$

$$= \frac{p_{\tilde{h}}(0) f_{\tilde{d}_1 \,|\, \tilde{h}}(3 \,|\, 0) f_{\tilde{d}_2 \,|\, \tilde{h}}(d \,|\, 0)}{p_{\tilde{h}}(0) f_{\tilde{d}_1 \,|\, \tilde{h}}(3 \,|\, 0)} \tag{6.25}$$

$$= f_{\tilde{d}_2 \,|\, \tilde{h}}(d \,|\, 0) \tag{6.26}$$

$$= \frac{1}{5} \qquad \text{if } 1 \leq d < 6, \tag{6.27}$$

$$\tag{6.28}$$

and zero otherwise. Intuitively, if the first distance is three, then it the ball must be solid, so the conditional pdf of the second distance given this information is the same as the conditional pdf of the distance given that the ball is not hollow.

6.4 (Computer defect)

a)

$$\mathrm{P}(\tilde{t} > 2) = \mathrm{P}(\tilde{t} > 2, \tilde{d} = 0) + \mathrm{P}(\tilde{t} > 2, \tilde{d} = 1) \tag{6.29}$$

$$= \mathrm{P}(\tilde{d} = 0)\mathrm{P}(\tilde{t} > 2 \,|\, \tilde{d} = 0) + \mathrm{P}(\tilde{d} = 1)\mathrm{P}(\tilde{t} > 2 \,|\, \tilde{d} = 1) \tag{6.30}$$

$$= 0.9 \int_2^\infty \exp(-t)\ \mathrm{d}t + 0.1 \int_2^\infty 2\exp(-2t)\ \mathrm{d}t \tag{6.31}$$

$$= 0.9\exp(-2) + 0.1\exp(-4) \tag{6.32}$$

$$= 0.124. \tag{6.33}$$

b)

$$p_{\tilde{d} \,|\, \tilde{t}}(1 \,|\, 2) = \frac{p_{\tilde{d}}(1) f_{\tilde{t} \,|\, \tilde{d}}(2 \,|\, 1)}{f_{\tilde{t}}(2)} \tag{6.34}$$

$$= \frac{p_{\tilde{d}}(1) f_{\tilde{t} \,|\, \tilde{d}}(2 \,|\, 1)}{p_{\tilde{d}}(0) f_{\tilde{t} \,|\, \tilde{d}}(2 \,|\, 0) + p_{\tilde{d}}(1) f_{\tilde{t} \,|\, \tilde{d}}(2 \,|\, 1)} \tag{6.35}$$

$$= \frac{0.1 \cdot 2\exp(-4)}{0.9 \cdot \exp(-1) + 0.1 \cdot 2\exp(-4)} \tag{6.36}$$

$$= 1.09\%. \tag{6.37}$$

6.5 (Wolf)

$$p_{\tilde{s} \,|\, \tilde{x}}\left(1 \middle| \begin{bmatrix}-1\\-1\end{bmatrix}\right) = \frac{p_{\tilde{s}}(1) f_{\tilde{x} \,|\, \tilde{s}}\left(\begin{bmatrix}-1\\-1\end{bmatrix} \middle| 1\right)}{p_{\tilde{s}}(1) f_{\tilde{x} \,|\, \tilde{s}}\left(\begin{bmatrix}-1\\-1\end{bmatrix} \middle| 1\right) + p_{\tilde{s}}(0) f_{\tilde{x} \,|\, \tilde{s}}\left(\begin{bmatrix}-1\\-1\end{bmatrix} \middle| 0\right)} \tag{6.38}$$

$$= \frac{0.25 \cdot \frac{1}{50}}{0.25 \cdot \frac{1}{50} + 0.75 \cdot \frac{1}{100}} \tag{6.39}$$

$$= 0.4. \tag{6.40}$$

6.6 (Potatoes)

a) We have

$$f_{\tilde{x}}(x) = \sum_{b=0}^{1} \sum_{w=0}^{1} p_{\tilde{b},\tilde{w}}(b,w)\, f_{\tilde{x}\,|\,\tilde{b},\tilde{w}}(x\,|\,b,w)\,. \tag{6.41}$$

The pdf is therefore a mixture of Gaussians, which is not Gaussian.

b) We use the empirical pmfs

$$p_{\tilde{w}}(1) = \frac{1}{3}, \qquad p_{\tilde{w}}(0) = \frac{2}{3}, \tag{6.42}$$

$$p_{\tilde{b}}(1) = \frac{2}{3}, \qquad p_{\tilde{b}}(0) = \frac{1}{3}. \tag{6.43}$$

c) The random variables are independent, since the empirical joint pmf factorizes into the product of the marginal pmfs

$$p_{\tilde{b},\tilde{w}}(0,0) = \frac{2}{9} = p_{\tilde{w}}(0)\, p_{\tilde{b}}(0)\,, \tag{6.44}$$

$$p_{\tilde{b},\tilde{w}}(0,1) = \frac{1}{9} = p_{\tilde{w}}(1)\, p_{\tilde{b}}(0)\,, \tag{6.45}$$

$$p_{\tilde{b},\tilde{w}}(1,0) = \frac{4}{9} = p_{\tilde{w}}(0)\, p_{\tilde{b}}(1)\,, \tag{6.46}$$

$$p_{\tilde{b},\tilde{w}}(1,1) = \frac{2}{9} = p_{\tilde{w}}(1)\, p_{\tilde{b}}(1)\,. \tag{6.47}$$

d) We compare

$$p_{\tilde{b}\,|\,\tilde{x},\tilde{w}}(1\,|\,40,1) = \frac{p_{\tilde{b},\tilde{w}}(1,1)\, f_{\tilde{x}\,|\,\tilde{b},\tilde{w}}(40\,|\,1,1)}{p_{\tilde{w}}(1)\, f_{\tilde{x}\,|\,\tilde{w}}(40\,|\,1)} \tag{6.48}$$

$$= \frac{\frac{2}{9} \cdot 0.02}{p_{\tilde{w}}(1)\, f_{\tilde{x}\,|\,\tilde{w}}(40\,|\,1)} \tag{6.49}$$

with

$$p_{\tilde{b}\,|\,\tilde{x},\tilde{w}}(0\,|\,40,1) = \frac{p_{\tilde{b},\tilde{w}}(0,1)\, f_{\tilde{x}\,|\,\tilde{b},\tilde{w}}(40\,|\,0,1)}{p_{\tilde{w}}(1)\, f_{\tilde{x}\,|\,\tilde{w}}(40\,|\,1)} \tag{6.50}$$

$$= \frac{\frac{1}{9} \cdot 0.01}{p_{\tilde{w}}(1)\, f_{\tilde{x}\,|\,\tilde{w}}(40\,|\,1)}. \tag{6.51}$$

Consequently, according to the model, it is more likely that the beetle was present that year.

6.7 (Halloween parade)

a)

$$\mathrm{P}(\tilde{w} \neq \tilde{r}) = p_{\tilde{w},\tilde{r}}(0,1) + p_{\tilde{w},\tilde{r}}(1,0) \tag{6.52}$$

$$= p_{\tilde{r}}(1)\, p_{\tilde{w}\,|\,\tilde{r}}(0\,|\,1) + p_{\tilde{r}}(0)\, p_{\tilde{w}\,|\,\tilde{r}}(1\,|\,0) \tag{6.53}$$

$$= 0.24. \tag{6.54}$$

b) It is not reasonable to assume that the forecast and the humidity are independent, even if we know that the forecast does not take the humidity into account. The reason is that both variables are linked through the rain. For example, if $\tilde{w} = 1$ then the humidity is probably high (because it probably will rain) and if $\tilde{w} = 0$ the humidity is probably low (because it probably won't rain). In contrast, conditioned on whether it rains or

not, it is quite reasonable to assume independence of $\tilde{w}$ and $\tilde{h}$, because $\tilde{h}$ is not used to produce the forecast.

c) By the chain rule for discrete and continuous random variables

$$p_{\tilde{r}\,|\,\tilde{w},\tilde{h}}(r\,|\,w,h) = \frac{f_{\tilde{h}\,|\,\tilde{w},\tilde{r}}(h\,|\,w,r)p_{\tilde{w}\,|\,\tilde{r}}(w\,|\,r)p_{\tilde{r}}(r)}{p_{\tilde{w}}(w)f_{\tilde{h}\,|\,\tilde{w}}(h\,|\,w)} \tag{6.55}$$

$$= \frac{f_{\tilde{h}\,|\,\tilde{r}}(h\,|\,r)p_{\tilde{w}\,|\,\tilde{r}}(w\,|\,r)p_{\tilde{r}}(r)}{\sum_{r=0}^{1} f_{\tilde{h}\,|\,\tilde{r}}(h\,|\,r)p_{\tilde{w}\,|\,\tilde{r}}(w\,|\,r)p_{\tilde{r}}(r)}. \tag{6.56}$$

This expression equals to zero unless h is between 0.5 and 0.7. If h is between 0.6 and 0.7 then it is equal to 1. If h is between 0.5 and 0.6 then

$$p_{\tilde{r}\,|\,\tilde{w},\tilde{h}}(1\,|\,1,h) = \frac{0.2 \cdot 0.8 \cdot 5}{0.2 \cdot 0.8 \cdot 5 + 0.8 \cdot 0.25 \cdot 2} \tag{6.57}$$

$$= 0.667, \tag{6.58}$$

$$p_{\tilde{r}\,|\,\tilde{w},\tilde{h}}(1\,|\,0,h) = \frac{0.2 \cdot 0.2 \cdot 5}{0.2 \cdot 0.2 \cdot 5 + 0.8 \cdot 0.75 \cdot 2} \tag{6.59}$$

$$= 0.143. \tag{6.60}$$

The conditional pmf is only well defined for $0.1 \leq h \leq 0.7$,

$$p_{\tilde{r}\,|\,\tilde{w},\tilde{h}}(1\,|\,w,h) = \begin{cases} 0 & \text{if } 0.1 \leq h \leq 0.5 \\ 0.667 & \text{if } 0.5 \leq h \leq 0.6 \quad \text{and } w = 1 \\ 0.143 & \text{if } 0.5 \leq h \leq 0.6 \quad \text{and } w = 0 \\ 1 & \text{if } 0.6 \leq h \leq 0.7. \end{cases} \tag{6.61}$$

We predict no rain (a) if the humidity is between 0.1 and 0.5 or (b) if $0.5 \leq h \leq 0.6$ and $w = 0$. Otherwise, we predict rain.

d) According to the model, if the humidity is between 0.1 and 0.5 then we are always right. Similarly, if the humidity is between 0.6 and 0.7 we are also always right. For h between 0.5 and 0.6 we make a mistake under two scenarios: (1) $w = 0$ and it rains and (2) $w = 1$ and it does not rain. The probability of these two events happening equals

$$\begin{aligned} \mathrm{P}(\text{Error}) &= \mathrm{P}(\tilde{r} = 0, w = 1, 0.5 \leq \tilde{h} \leq 0.6) + \mathrm{P}(\tilde{r} = 1, w = 0, 0.5 \leq \tilde{h} \leq 0.6) \\ &= \int_{h=0.5}^{0.6} f_{\tilde{h}\,|\,\tilde{r}}\,(h\,|\,0)\,p_{\tilde{w}\,|\,\tilde{r}}(1\,|\,0)p_{\tilde{r}}\,(0)\;\mathrm{d}h + \int_{h=0.5}^{0.6} f_{\tilde{h}\,|\,\tilde{r}}\,(h\,|\,1)\,p_{\tilde{w}\,|\,\tilde{r}}(0\,|\,1)p_{\tilde{r}}\,(1)\;\mathrm{d}h \\ &= \int_{h=0.5}^{0.6} 0.8 \cdot 2 \cdot 0.25\;\mathrm{d}h + \int_{h=0.5}^{0.6} 0.2 \cdot 5 \cdot 0.2\;\mathrm{d}h = 0.06. \end{aligned}$$

6.8 (Fish)

a) The position is a uniform random variable since

$$f_{\tilde{x}}x = p_{\tilde{s}}(a)f_{\tilde{x}\,|\,\tilde{s}}(x\,|\,a) + p_{\tilde{s}}(b)f_{\tilde{x}\,|\,\tilde{s}}(x\,|\,b) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.62}$$

b) By the chain rule for discrete and continuous random variables

$$p_{\tilde{s}\,|\,\tilde{x}}(b\,|\,0.25) = \frac{p_{\tilde{s}}\,(b)\,f_{\tilde{x}\,|\,\tilde{s}}(0.25\,|\,b)}{f_{\tilde{x}}\,(0.25)} \tag{6.63}$$

$$= 0.25. \tag{6.64}$$

c) Integrating the conditional pdfs we obtain

$$F_{\tilde{x}\,|\,\tilde{s}}\,(x\,|\,a) = \begin{cases} 0 & \text{if } x < 0, \\ 2x - x^2 & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1. \end{cases} \tag{6.65}$$

$$F_{\tilde{x}\,|\,\tilde{s}}\,(x\,|\,b) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1, \end{cases} \tag{6.66}$$

d) Let u_1 and u_2 be the two uniform samples. We use the first one to sample from $\tilde{s}$ by assigning $\tilde{s} = a$ if u_1 is below 0.5, and $\tilde{s} = b$ otherwise. In this case, $u_1 = 0.8$ so we obtain $\tilde{s} = b$. Then we sample from the conditional distribution of $\tilde{x}$ given $\tilde{s} = b$ using u_2 via the inverse transform method. We have $F_{\tilde{x}\,|\,\tilde{b}}\,(0.8\,|\,b) = 0.64$ (computed by solving for $x^2 = 0.64$), so the sample of $\tilde{x}$ equals 0.8.

6.9 (Samples and conditional independence) In (1) if we only consider the samples for which $c = 0$, the density of the samples in the vertical direction looks the same for all values of a between 0 and 1, and is zero otherwise, which suggests that the conditional pdf of $\tilde{b}$ given $\tilde{a} = a$ and $\tilde{c} = 0$ is the same for all values of a. Similarly, when $c = 1$, the vertical density again looks the same for all values of a between 1 and 2, and is zero otherwise, which suggests that the conditional pdf of $\tilde{b}$ given $\tilde{a} = a$ and $\tilde{c} = 1$ is the same for all values of a. Since $\tilde{c}$ only has these two possible values, this would imply that $\tilde{a}$ and $\tilde{b}$ are conditionally independent given $\tilde{c}$.

In (2) $\tilde{a}$ and $\tilde{b}$ again look conditionally independent given $\tilde{c} = 0$, but not given $\tilde{c} = 1$, because for those samples the vertical density for a between 0.5 and 1 is clearly different to the density for a between 1 and 1.5, or between 1.5 and 2. Consequently, $\tilde{a}$ and $\tilde{b}$ are not conditionally independent given $\tilde{c}$.

In (3) if we only consider the samples for which $c = 0$, the vertical density for a between 0 and 1 is clearly different to the density for a between 1 and 2. The same occurs for the samples for which $c = 1$. Consequently, $\tilde{a}$ and $\tilde{b}$ are not conditionally independent given $\tilde{c}$.

6.10 (Chad)

a) The kernel density estimate is of the form

$$f_{\tilde{t}\,|\,\tilde{c}}\,(t\,|\,0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{1}{2}\,\Pi\left(\frac{t - d_{0,i}}{2}\right), \tag{6.67}$$

$$f_{\tilde{t}\,|\,\tilde{c}}\,(t\,|\,1) = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{2}\,\Pi\left(\frac{t - d_{1,i}}{2}\right), \tag{6.68}$$

$$\tag{6.69}$$

where Π is a rectangular kernel with unit width, $d_{1,0}, \ldots, d_{n_1,0}$ the temperatures when Chad is not there and $d_{1,1}, \ldots, d_{n_1,1}$ the temperatures when he is there. The estimate is shown in Figure 6.1.

b) We have

$$f_{\tilde{t}\,|\,\tilde{c}}\,(68\,|\,0) = 0.2 > 0 = f_{\tilde{t}\,|\,\tilde{c}}\,(68\,|\,1)\,, \tag{6.70}$$

so the ML estimate is that Chad is not at the office.

c) That probability does not make sense because the parameter representing the presence of Chad is deterministic.

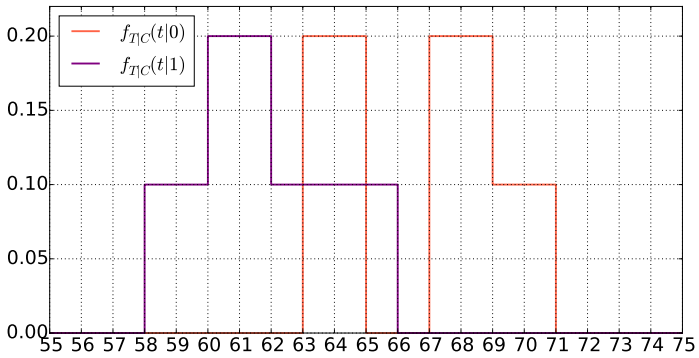


Figure 6.1 Kernel density estimate for Exercise 6.10.

d) The empirical pmf is

$$p_{\tilde{c}}(0) = \frac{5}{15} = \frac{1}{3}, \tag{6.71}$$

$$p_{\tilde{c}}(1) = \frac{10}{15} = \frac{2}{3}. \tag{6.72}$$

e) Applying Bayes' rule,

$$p_{\tilde{c}\,|\,\tilde{t}}(0\,|\,64) = \frac{p_{\tilde{c}}(0)\, f_{\tilde{t}\,|\,\tilde{c}}(64\,|\,0)}{p_{\tilde{c}}(0)\, f_{\tilde{t}\,|\,\tilde{c}}(64\,|\,0) + p_{\tilde{c}}(1)\, f_{\tilde{t}\,|\,\tilde{c}}(64\,|\,1)} \tag{6.73}$$

$$= \frac{\frac{1}{3}0.2}{\frac{1}{3}0.2 + \frac{2}{3}0.1} \tag{6.74}$$

$$= \frac{1}{2}, \tag{6.75}$$

$$p_{\tilde{c}\,|\,\tilde{t}}(1\,|\,64) = 1 - p_{\tilde{c}\,|\,\tilde{t}}(0\,|\,64) \tag{6.76}$$

$$= \frac{1}{2}. \tag{6.77}$$

According to the posterior pmf there is a 50 % chance that Chad is there.

f) Both $f_{\tilde{t}\,|\,\tilde{c}}(57\,|\,0)$ and $f_{\tilde{t}\,|\,\tilde{c}}(68\,|\,1)$ are zero so both the ML estiamate and Bayesian posterior are inconclusive. If we use a parametric distribution such as a Gaussian to fit the data, then $f_{\tilde{t}\,|\,\tilde{c}}(57\,|\,0)$ and $f_{\tilde{t}\,|\,\tilde{c}}(68\,|\,1)$ would not be set to zero as long as the distribution has nonzero values on all of the real line (as is the case for a Gaussian pdf). This would allow us to apply MAP or ML estimation. A nonparametric solution would be to use a kernel with a larger width.

6.11 (Mixture model with fixed mean) We have

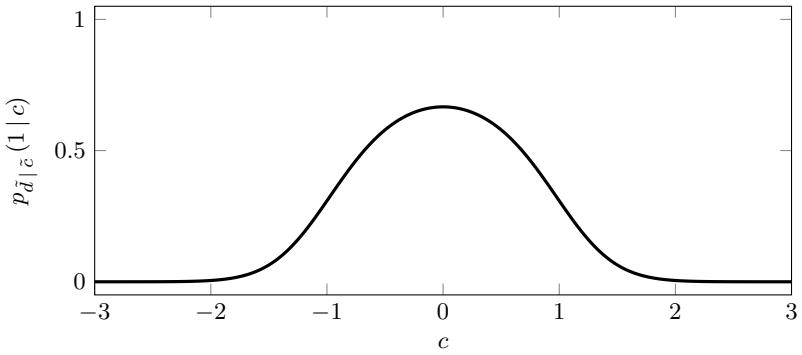
$$p_{\tilde{s}\,|\,\tilde{c}}\,(1\,|\,c) = \frac{p_{\tilde{s}}\,(1)\,f_{\tilde{c}\,|\,\tilde{s}}\,(c\,|\,1)}{p_{\tilde{s}}\,(0)\,f_{\tilde{c}\,|\,\tilde{s}}\,(c\,|\,0) + p_{\tilde{s}}\,(1)\,f_{\tilde{c}\,|\,\tilde{s}}\,(c\,|\,1)} \tag{6.78}$$

$$= \frac{\frac{\theta}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{c^2}{2\sigma_1^2}\right)}{\frac{\theta}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{c^2}{2\sigma_1^2}\right) + \frac{1-\theta}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{c^2}{2\sigma_0^2}\right)} \tag{6.79}$$

$$= \frac{1}{1 + \frac{(1-\theta)\sigma_1}{\theta\sigma_0}\exp\left(\frac{c^2}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\right)} \tag{6.80}$$

$$\tag{6.81}$$

For $\theta := 0.5$, $\sigma_0 := 1$, and $\sigma_1 := 0.5$, the function looks like this:



6.12 (K-means and Gaussian mixture models) There are two main differences. First, when fitting a Gaussian mixture model we make *soft* assignments to each cluster, represented by the membership probabilities, as opposed to the *hard* assignment in Lloyd's algorithm, where each data point is assigned to a single cluster. Second, the Gaussian mixture model estimates the covariance matrix of each cluster, as well as its mean, whereas in k-means the covariance structure of each cluster is implicitly assumed to be the same.

6.13 (Alternative model for coin flip)

a) The model makes sense because we suspect that the coin that is more prone to be heads than tails, but we are uncertain about the exact bias so we just assume it is uniform between 1/2 (fair coin) and 1 (coin that always lands heads). To compute the probability of heads, we integrate over all the possible values of the Bernouilli parameter. Let $\tilde{r}$ be the result of the coin flip and $\tilde{\theta}$ the parameter of the Bernouilli. Since $\tilde{r}$ is uniform between 1/2 and 1 $f_{\tilde{\theta}}\,(\theta)$ is equal to 2 for $1/2 \leq \theta \leq 1$ and zero otherwise.

$$P\,(\text{heads}) = p_{\tilde{r}}\,(1) = \int_{\theta=1/2}^{1} f_{\tilde{\theta}}\,(\theta)\,p_{\tilde{r}\,|\,\tilde{\theta}}\,(1\,|\,\theta)\;\mathrm{d}\theta = \int_{\theta=1/2}^{1} 2\,\theta\;\mathrm{d}\theta = 1 - \frac{1}{4} = \frac{3}{4}, \tag{6.82}$$

$$P\,(\text{tails}) = 1 - P\,(\text{heads}) = \frac{1}{4}. \tag{6.83}$$

b) The conditional pdf on the bias of the coin flip conditioned on tails equals

$$f_{\tilde{\theta}\,|\,\tilde{r}}\,(\theta\,|\,0) = \frac{f_{\tilde{\theta}}\,(\theta)\,p_{\tilde{r}\,|\,\tilde{\theta}}\,(0)}{p_{\tilde{r}}\,(0)} \tag{6.84}$$

$$= \frac{f_{\tilde{\theta}}\,(\theta)\,(1-\theta)}{1/4} = \begin{cases} 8\,(1-\theta) & 1/2 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.85}$$

Note that the conditional pdf is more skewed towards 1/2, as shown on Figure 6.2. It makes sense because the coin flip is tails. Intuitively the model should be adjusted towards the coin flip being less biased.

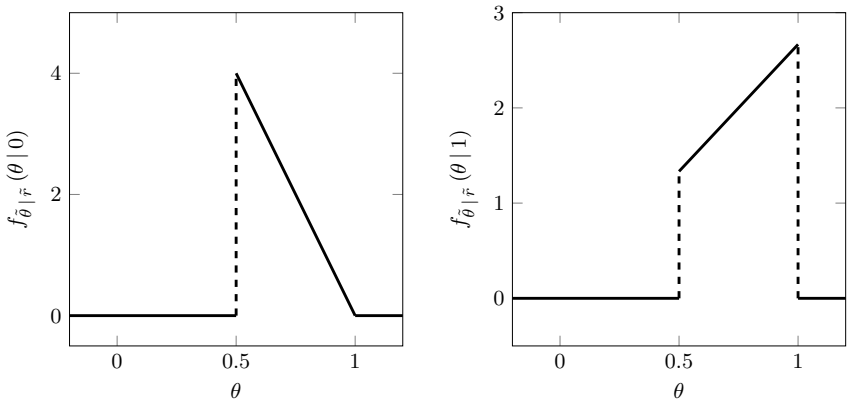


Figure 6.2 Conditional pdf of the bias of the coin flip given tails (left) and heads (right).

The conditional pdf on the bias of the coin flip conditioned on heads equals

$$f_{\tilde{\theta}\,|\,\tilde{r}}\,(\theta\,|\,1) = \frac{f_{\tilde{\theta}}\,(\theta)\,p_{\tilde{r}\,|\,\tilde{\theta}}\,(1)}{p_{\tilde{r}}\,(1)} = \frac{f_{\tilde{\theta}}\,(\theta)\,\theta}{3/4} = \begin{cases} \frac{8\,\theta}{3} & 1/2 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.86}$$

In this case the conditional pdf is more skewed towards 1/2, as shown on Figure 6.2. Intuitively, we are adjusting the model by incorporating some evidence that supports that the coin might be biased towards heads.

c) The data seem to indicate that the probability of heads is very small. However, no matter what evidence we observe, our model will always assign zero probability density to any value of the bias below 0.5, so we should probably consider changing the prior.

6.14 (Two coin flips)

a)

$$p_{\tilde{x}_1,\tilde{x}_2}(1,1) = \int_{\theta=-\infty}^{\infty} f_{\tilde{\theta}}(\theta) p_{\tilde{x}_1 \,|\, \tilde{\theta}}(1 \,|\, \theta) p_{\tilde{x}_2 \,|\, \tilde{\theta}}(1 \,|\, \theta) \, \mathrm{d}\theta \tag{6.87}$$

$$= \int_0^1 \theta^2 \, \mathrm{d}\theta = \frac{1}{3}, \tag{6.88}$$

$$p_{\tilde{x}_1,\tilde{x}_2}(0,1) = \int_0^1 \theta(1-\theta) \, \mathrm{d}\theta = \frac{1}{6}, \tag{6.89}$$

$$p_{\tilde{x}_1,\tilde{x}_2}(1,0) = \int_0^1 \theta(1-\theta) \, \mathrm{d}\theta = \frac{1}{6}, \tag{6.90}$$

$$p_{\tilde{x}_1,\tilde{x}_2}(0,0) = \int_0^1 (1-\theta)^2 \, \mathrm{d}\theta = \frac{1}{3}. \tag{6.91}$$

b) The posterior pdf equals

$$f_{\tilde{\theta} \,|\, \tilde{x}_1,\tilde{x}_2}(\theta \,|\, 1,1) = \frac{f_{\tilde{\theta}}(\theta) p_{\tilde{x}_1 \,|\, \tilde{\theta}}(1 \,|\, \theta) p_{\tilde{x}_2 \,|\, \tilde{\theta}}(1 \,|\, \theta)}{p_{\tilde{x}_1,\tilde{x}_2}(1,1)} \tag{6.92}$$

$$= \frac{\theta^2}{1/3} = 3\theta^2. \tag{6.93}$$

The conditional probability that $\tilde{\theta} < 1/2$ equals

$$\mathrm{P}\left(\tilde{\theta} < \frac{1}{2}\right) = \int_{\theta=0}^{1/2} f_{\tilde{\theta} \,|\, \tilde{x}_1,\tilde{x}_2}(\theta \,|\, 1,1) \, \mathrm{d}\theta \tag{6.94}$$

$$= \int_{\theta=0}^{1/2} 3\theta^2 \, \mathrm{d}\theta \tag{6.95}$$

$$= \frac{1}{8}. \tag{6.96}$$

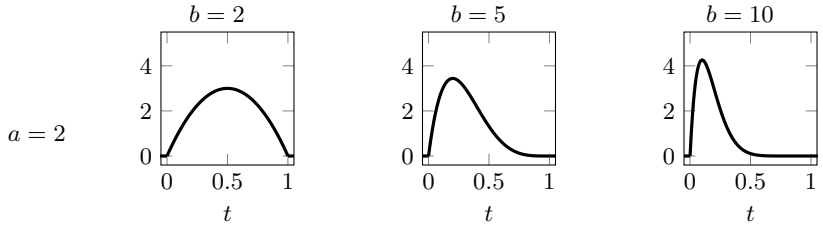
6.15 (PT Cruiser) Let $\tilde{t}$ denote the number of times the car is used. Under the assumptions, given $\tilde{\theta} = \theta$, the conditional distribution of $\tilde{t}$ is geometric with parameter θ. The posterior distribution given $\tilde{t} = t$ is therefore

$$f_{\tilde{\theta} \,|\, \tilde{t}}\,(\theta \,|\, t) = \frac{f_{\tilde{\theta}}\,(\theta)\, p_{\tilde{t} \,|\, \tilde{\theta}}\,(t \,|\, \theta)}{p_{\tilde{t}}\,(t)} \tag{6.97}$$

$$= \frac{f_{\tilde{\theta}}\,(\theta)\, p_{\tilde{t} \,|\, \tilde{\theta}}\,(t \,|\, \theta)}{\int_u f_{\tilde{\theta}}\,(u)\, p_{\tilde{t} \,|\, \tilde{\theta}}\,(t \,|\, u) \, \mathrm{d}u} \tag{6.98}$$

$$= \frac{\theta\,(1-\theta)^{t-1}}{\int_{u=0}^1 u\,(1-u)^{t-1} \, \mathrm{d}u}, \tag{6.99}$$

which is a beta distribution with parameters $a = 2$ and $b = t$. In order to understand how it changes with t, we need to look at the shape of the beta distribution for $a = 2$ and different values of b:



For small t the posterior is very flat (see $a = 2$, $b = 1$). As t grows, the pdf is increasingly more skewed towards zero and more concentrated, indicating that the car is likely to be in good condition after all.

7

Averaging

Exercises

7.1 (Race)

a) By the formula for the mean of a function of two random variables,

$$\mathbb{E}\left[\tilde{d}-\tilde{a}\right] = \sum_{d\in\{5,10,20\}} \sum_{a\in\{5,10,20\}} (d-a) p_{\tilde{d},\tilde{a}}(d,a) \tag{7.1}$$

$$= -5\cdot 0.1 + 5\cdot 0.1 + 15\cdot 0.1 + 10\cdot 0.2 \tag{7.2}$$

$$= 3.5, \tag{7.3}$$

so the mean difference is 3.5 km.

b) The mean equals

$$\mathbb{E}\left[\tilde{a}\right] = \sum_{d\in\{5,10,20\}} \sum_{a\in\{5,10,20\}} a p_{\tilde{d},\tilde{a}}(d,a) \tag{7.4}$$

$$= 5\left(0.2+0.1+0.1\right) + 10\left(0.1+0.2+0.2\right) + 20\cdot 0.1 \tag{7.5}$$

$$= 2+5+2 = 9. \tag{7.6}$$

The mean square equals

$$\mathbb{E}\left[\tilde{a}^2\right] = \sum_{d\in\{5,10,20\}} \sum_{a\in\{5,10,20\}} a^2 p_{\tilde{d},\tilde{a}}(d,a) \tag{7.7}$$

$$= 25\left(0.2+0.1+0.1\right) + 100\left(0.1+0.2+0.2\right) + 400\cdot 0.1 \tag{7.8}$$

$$= 10+50+40 = 100. \tag{7.9}$$

Consequently the variance equals

$$\mathrm{Var}\left[\tilde{a}\right] = \mathbb{E}\left[\tilde{a}^2\right] - \mathbb{E}\left[\tilde{a}\right]^2 \tag{7.10}$$

$$= 100 - 81 = 19. \tag{7.11}$$

7.2 (Mean of a function)

$$\mathbb{E}[\tilde{b}] := \sum_{b \in B} b p_{\tilde{b}}(b) \tag{7.12}$$

$$= \sum_{b \in B} b \mathrm{P}(\tilde{b} = b) \tag{7.13}$$

$$= \sum_{b \in B} b \sum_{\{a:h(a)=b\}} \mathrm{P}(\tilde{a} = a) \tag{7.14}$$

$$= \sum_{b \in B} \sum_{\{a:h(a)=b\}} h(a) \mathrm{P}(\tilde{a} = a) \tag{7.15}$$

$$= \sum_{a \in A} h(a) p_{\tilde{a}}(a). \tag{7.16}$$

Where the last step follows from the fact that each $a \in A$ is in exactly one of the sets $\{a : h(a) = b\}$.

7.3 (Presents) We define a random variable $\tilde{a}_i$ that is equal to one when kid i gets the present bought by their own parents, and to zero otherwise. The probability of the event of kid i getting their own present is $1/n$ so

$$p_{\tilde{a}_i}(0) = 1 - \frac{1}{n}, \tag{7.17}$$

$$p_{\tilde{a}_i}(1) = \frac{1}{n}. \tag{7.18}$$

By linearity of expectation

$$\mathbb{E}\left[\text{Number of kids that get their own present}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \tilde{a}_i\right] \tag{7.19}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\tilde{a}_i\right] \tag{7.20}$$

$$= \sum_{i=1}^{n} p_{\tilde{a}_i}(1) \tag{7.21}$$

$$= 1. \tag{7.22}$$

On average one kid gets their own present.

7.4 (Cats and dogs) The mean is equal to 1.05 (from the example) and the mean square equals

$$\mathbb{E}[(\tilde{c} + \tilde{d})^2] = \sum_{c=0}^{3} \sum_{d=0}^{2} (c+d)^2 p_{\tilde{c},\tilde{d}}(c,d) \tag{7.23}$$

$$= 0.15 + 4 \cdot 0.1 + 9 \cdot 0.05 + 0.2 + 4 \cdot 0.05 + 9 \cdot 0.03 + 4 \cdot 0.05 + 9 \cdot 0.02$$

$$= 2.05. \tag{7.24}$$

The standard deviation therefore equals

$$\sqrt{\mathrm{Var}[\tilde{c} + \tilde{d}]} = \sqrt{\mathbb{E}[(\tilde{c} + \tilde{d})^2] - \mathbb{E}[\tilde{c} + \tilde{d}]} \tag{7.25}$$

$$= \sqrt{2.05 - 1.05^2} = 0.973. \tag{7.26}$$

7.5 (Mean squared difference)

$$\mathbb{E}\left[(\tilde{a}-\tilde{b})^2\right] = \mathbb{E}\left[\tilde{a}^2+\tilde{b}^2-2\tilde{a}\tilde{b}\right] \tag{7.27}$$

$$= \mathbb{E}[\tilde{a}^2]+\mathbb{E}[\tilde{b}^2]-2\mathbb{E}\left[\tilde{a}\tilde{b}\right] \quad \text{by linearity} \tag{7.28}$$

$$= \mathbb{E}[\tilde{a}^2]+\mathbb{E}[\tilde{b}^2]-2\mathbb{E}\left[\tilde{x}_1\right]\mathbb{E}\left[\tilde{b}\right] \quad \text{by independence} \tag{7.29}$$

$$= \mathbb{E}[\tilde{a}^2]-\mu^2+\mathbb{E}[\tilde{b}^2]-\mu^2 \tag{7.30}$$

$$= 2\sigma^2. \tag{7.31}$$

7.6 (Basketball player) Let $\tilde{x}$ be a random variable that represents the points scored by the player. We have

$$\mathrm{P}(\tilde{x}=2) = \mathrm{P}(\{\text{attempts 2 point shot}\}\cap\{\text{makes shot}\}) \tag{7.32}$$

$$= \mathrm{P}(\{\text{attempts 2 point shot}\})\mathrm{P}(\{\text{makes shot}\} \mid \{\text{attempts 2 point shot}\})$$

$$= 0.5\alpha, \tag{7.33}$$

$$\mathrm{P}(\tilde{x}=3) = \mathrm{P}(\{\text{attempts 3 point shot}\}\cap\{\text{makes shot}\}) \tag{7.34}$$

$$= \mathrm{P}(\{\text{attempts 3 point shot}\})\mathrm{P}(\{\text{makes shot}\} \mid \{\text{attempts 3 point shot}\})$$

$$= 0.4(1-\alpha), \tag{7.35}$$

$$\mathrm{P}(\tilde{x}=0) = 1-0.5\alpha-0.4(1-\alpha). \tag{7.36}$$

The mean, mean square and variance equal

$$\mathbb{E}[\tilde{x}] = 2\mathrm{P}(\tilde{x}=2)+3\mathrm{P}(\tilde{x}=3) \tag{7.37}$$

$$= \alpha+1.2-1.2\alpha \tag{7.38}$$

$$= 1.2-0.2\alpha, \tag{7.39}$$

$$\mathbb{E}[\tilde{x}^2] = 4\mathrm{P}(\tilde{x}=2)+9\mathrm{P}(\tilde{x}=3) \tag{7.40}$$

$$= 2\alpha+3.6-3.6\alpha \tag{7.41}$$

$$= 3.6-1.6\alpha, \tag{7.42}$$

$$\mathrm{Var}[\tilde{x}] = \mathbb{E}[\tilde{x}^2]-\mathbb{E}[\tilde{x}]^2 \tag{7.43}$$

$$= 3.6-1.6\alpha-(1.2-0.2\alpha)^2 \tag{7.44}$$

$$= -0.04\alpha^2-1.12\alpha+2.16. \tag{7.45}$$

Therefore, the standard deviation is equal to $\sqrt{-0.04\alpha^2-1.12\alpha+2.16}$. The derivative of the standard deviation with respect to α equals

$$\frac{-0.08\alpha-1.12}{2\sqrt{-0.04\alpha^2-1.12\alpha+2.16}}, \tag{7.46}$$

which is negative for $\alpha > 0$. Consequently, for $0 \leq \alpha \leq 1$ the standard deviation decreases as we increase α. It is minimized at $\alpha = 1$ (the player only attempts 2-point shots), where it is equal to 1. It is maximized at $\alpha = 0$ (the player only attempts 3-point shots), where it is equal to 1.47.

7.7 (Computer defect)

a) The random variable $\tilde{d}$ is discrete, so the conditional mean of $\tilde{t}$ is also discrete (it is a function of $\tilde{d}$). The mean of an exponential random variable with parameter λ equal $\frac{1}{\lambda}$.

As a result, we have $\mu_{\tilde{t}\,|\,\tilde{d}}(1) = 0.5$ and $\mu_{\tilde{t}\,|\,\tilde{d}}(0) = 1$, so the pmf of $\mu_{\tilde{t}\,|\,\tilde{d}}(\tilde{d})$ is

$$p_{\mu_{\tilde{t}\,|\,\tilde{d}}(\tilde{d})}(0.5) = \mathrm{P}(\tilde{d} = 1) \tag{7.47}$$

$$= 0.1, \tag{7.48}$$

$$p_{\mu_{\tilde{t}\,|\,\tilde{d}}(\tilde{d})}(1) = \mathrm{P}(\tilde{d} = 0) \tag{7.49}$$

$$= 0.9, \tag{7.50}$$

$$p_{\tilde{a}}(a) = 0, \qquad \text{for } \tilde{a} \notin \{0.5, 1\}\,. \tag{7.51}$$

b) From the previous question,

$$\mathbb{E}\left[\tilde{t}\right] = \mathrm{P}(\tilde{d} = 0)\mu_{\tilde{t}\,|\,\tilde{d}}(0) + \mathrm{P}(\tilde{d} = 1)\mu_{\tilde{t}\,|\,\tilde{d}}(1) \tag{7.52}$$

$$= 0.95. \tag{7.53}$$

The mean and variance of an exponential random variable with parameter λ equal $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$ respectively, so its mean square is $\frac{2}{\lambda^2}$. Therefore,

$$\mathbb{E}\left[\tilde{t}^2 \,|\, \tilde{d} = 0\right] = 2, \tag{7.54}$$

$$\mathbb{E}\left[\tilde{t}^2 \,|\, \tilde{d} = 1\right] = \frac{1}{2}. \tag{7.55}$$

By iterated expectation,

$$\mathbb{E}\left[\tilde{t}^2\right] = \mathrm{P}(\tilde{d} = 0)\mu_{\tilde{t}^2\,|\,\tilde{d}}(0) + \mathrm{P}(\tilde{d} = 1)\mu_{\tilde{t}^2\,|\,\tilde{d}}(1) \tag{7.56}$$

$$= 1.85. \tag{7.57}$$

We conclude that

$$\mathrm{Var}\left[\tilde{t}\right] = \mathbb{E}\left[\tilde{t}^2\right] - \mathbb{E}\left[\tilde{t}\right]^2 \tag{7.58}$$

$$= 0.9475. \tag{7.59}$$

c) The number of computers that break down during the first year is binomial with parameters $n = 100$ and θ, which represents the probability that $\tilde{t} \leq 1$:

$$\theta = \mathrm{P}(\tilde{t} \leq 1) = \int_0^1 f_{\tilde{t}}(t)\ \mathrm{d}t \tag{7.60}$$

$$= \int_0^1 \left(0.9\exp(-t) + 0.1\exp(-2t)\right)\ \mathrm{d}t \tag{7.61}$$

$$= -0.9\exp(-t)]_0^1 - 0.1\exp(-2t)]_0^1 \tag{7.62}$$

$$= 0.9(1 - \exp(-1)) + 0.1(1 - \exp(-2)) \tag{7.63}$$

$$= 0.655. \tag{7.64}$$

The binomial therefore has mean $n\theta = 65.5$ and variance $n\theta(1-\theta) = 22.6$.

7.8 (Hen breeds)

a) Let $\tilde{b}_1$, ..., $\tilde{b}_4$ represent the eggs laid by the bantam hens and $\tilde{s}_1$, $\tilde{s}_2$, $\tilde{s}_3$ the ones laid by the Sussex hens. Since it's always warm, the mean of the eggs laid by each bantam

chicken equals,

$$\mu_{\text{bantam}} = \sum_{i=4}^{6} i\mathrm{P}\left(\text{eggs} = i\right) \tag{7.65}$$

$$= 4 \cdot 0.25 + 5 \cdot 0.5 + 6 \cdot 0.25 \tag{7.66}$$

$$= 5, \tag{7.67}$$

and the mean of the eggs laid by each Sussex hen is

$$\mu_{\text{sussex}} = \sum_{i=3}^{6} i\mathrm{P}\left(\text{eggs} = i\right) \tag{7.68}$$

$$= 3 \cdot 0.25 + 4 \cdot 0.25 + 5 \cdot 0.25 + 6 \cdot 0.25 \tag{7.69}$$

$$= 4.5. \tag{7.70}$$

By linearity of expectation, which does not require independence assumptions, the mean of the total number of eggs $\tilde{t}$ is

$$\mathbb{E}\left[\tilde{t}\right] = \mathbb{E}\left[\sum_{i=1}^{4} \tilde{b}_i + \sum_{i=1}^{3} \tilde{s}_i\right] \tag{7.71}$$

$$= \sum_{i=1}^{4} \mathbb{E}\left[\tilde{b}_i\right] + \sum_{i=1}^{3} \mathbb{E}\left[\tilde{s}_i\right] \tag{7.72}$$

$$= 4\mu_{\text{bantam}} + 3\mu_{\text{sussex}} \tag{7.73}$$

$$= 33.5 \text{ eggs}. \tag{7.74}$$

b) Let $\tilde{w}$ represent the weather, $\tilde{b}$ the eggs laid by a bantam hen and $\tilde{s}$ the eggs laid by a Sussex hen. The conditional mean function of $\tilde{b}$ given $\tilde{w} = \text{cold}$ is

$$\mu_{\tilde{b}\,|\,\tilde{w}}(\text{cold}) = \sum_{i=0}^{1} i \cdot p_{\tilde{b}\,|\,\tilde{w}}\left(i\,|\,\text{cold}\right) \tag{7.75}$$

$$= 0.25. \tag{7.76}$$

The conditional mean function of $\tilde{s}$ given $\tilde{w} = \text{cold}$ is

$$\mu_{\tilde{s}\,|\,\tilde{w}}(\text{cold}) = \sum_{i=0}^{2} i \cdot p_{\tilde{s}\,|\,\tilde{w}}\left(i\,|\,\text{cold}\right) \tag{7.77}$$

$$= 0.5 + 2 \cdot 0.25 = 1. \tag{7.78}$$

The conditional mean function of $\tilde{b}$ given $\tilde{w} = \text{warm}$ is 5, and conditional mean function of $\tilde{s}$ given $\tilde{w} = \text{warm}$ is 4.5 as derived above. By iterated expectation,

$$\mathbb{E}\left[\tilde{b}\right] = \mathbb{E}\left[\mu_{\tilde{b}\,|\,\tilde{w}}(\tilde{w})\right] \tag{7.79}$$

$$= \mathrm{P}\left(\tilde{w} = \text{cold}\right)\mu_{\tilde{b}\,|\,\tilde{w}}(\text{cold}) + \mathrm{P}\left(\tilde{w} = \text{warm}\right)\mu_{\tilde{b}\,|\,\tilde{w}}(\text{warm}) \tag{7.80}$$

$$= 0.25\left(1 - \alpha\right) + 5\alpha = 0.25 + 4.75\alpha. \tag{7.81}$$

Similarly

$$\mathbb{E}\left[\tilde{s}\right] = \mathbb{E}\left[\mu_{\tilde{s}\,|\,\tilde{w}}(\tilde{w})\right] \tag{7.82}$$

$$= \mathrm{P}\left(\tilde{w} = \text{cold}\right)\mu_{\tilde{s}\,|\,\tilde{w}}(\text{cold}) + \mathrm{P}\left(\tilde{w} = \text{warm}\right)\mu_{\tilde{s}\,|\,\tilde{w}}(\text{warm}) \tag{7.83}$$

$$= 1 - \alpha + 4.5\alpha = 1 + 3.5\alpha. \tag{7.84}$$

The Sussex hen has a higher mean if

$$0.25 + 4.75\alpha \leq 1 + 3.5\alpha, \tag{7.85}$$

which is equivalent to $\alpha \leq 0.75/1.25 = 0.6$.

7.9 (Hotel questionnaire)

a) By iterated expectation

$$\mathbb{E}\left[s\right] = p_{\tilde{r}}(0)\mu_{\tilde{s}\,|\,\tilde{r}}(0) + p_{\tilde{r}}(1)\mu_{\tilde{s}\,|\,\tilde{r}}(1) \tag{7.86}$$

$$= 0.8 \cdot 4 + 0.2 \cdot 2 = 3.6. \tag{7.87}$$

b) Given the conditional independence assumptions, the conditional pmf of $\tilde{s}$ given $\tilde{q} = 1$ equals

$$p_{\tilde{s}\,|\,\tilde{q}}(s\,|\,1) = \frac{p_{\tilde{s},\tilde{q}}(s,1)}{p_{\tilde{q}}(1)} \tag{7.88}$$

$$= \frac{\sum_{r=0}^{1} p_{\tilde{r},\tilde{s},\tilde{q}}(r,s,1)}{\sum_{r=0}^{1} p_{\tilde{r},\tilde{q}}(r,1)} \tag{7.89}$$

$$= \frac{\sum_{r=0}^{1} p_{\tilde{r}}(r)p_{\tilde{q}\,|\,\tilde{r}}(q\,|\,1)p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,r)}{\sum_{r=0}^{1} p_{\tilde{r}}(r)p_{\tilde{q}\,|\,\tilde{r}}(q\,|\,1)} \tag{7.90}$$

$$= \frac{0.8 \cdot 0.25 \cdot p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,0) + 0.2 \cdot 0.9 \cdot p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,1)}{0.8 \cdot 0.25 + 0.2 \cdot 0.9} \tag{7.91}$$

$$= \frac{0.2p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,0) + 0.18p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,1)}{0.38}. \tag{7.92}$$

The conditional mean given $\tilde{q} = 1$ therefore equals

$$\mu_{\tilde{s}\,|\,\tilde{q}}(1) = \sum_{s=1}^{5} s p_{\tilde{s}\,|\,\tilde{q}}(s\,|\,1) \tag{7.93}$$

$$= \sum_{s=1}^{5} s\left(\frac{0.2p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,0) + 0.18p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,1)}{0.38}\right) \tag{7.94}$$

$$= \frac{0.2}{0.38}\sum_{s=1}^{5} s p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,0) + \frac{0.18}{0.38}\sum_{s=1}^{5} s p_{\tilde{s}\,|\,\tilde{r}}(s\,|\,1) \tag{7.95}$$

$$= \frac{0.2}{0.38}\mu_{\tilde{s}\,|\,\tilde{r}}(0) + \frac{0.18}{0.38}\mu_{\tilde{s}\,|\,\tilde{r}}(1) \tag{7.96}$$

$$= \frac{0.2 \cdot 4}{0.38} + \frac{0.18 \cdot 2}{0.38} \tag{7.97}$$

$$= 3.05. \tag{7.98}$$

The observed mean is 0.55 less than the true mean.

7.10 (Another basketball player) Let $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{x}_3$ represent the number of free throws, 2-point shots, 3-point shots, and free throws respectively. We denote the number of made

free throws, 2-point shots, and 3-point shots by $\tilde{m}_1$, $\tilde{m}_2$ and $\tilde{m}_3$ respectively, and set $p_1 = 0.8$, $p_2 = 0.5$ and $p_3 = 0.3$. The number of points $\tilde{y}$ is equal to

$$\tilde{y} = \tilde{m}_1 + 2\tilde{m}_2 + 3\tilde{m}_3. \tag{7.99}$$

By linearity of expectation (which does not require independence)

$$\mathbb{E}[\tilde{y}] = \mathbb{E}[\tilde{m}_1] + 2\mathbb{E}[\tilde{m}_2] + 3\mathbb{E}[\tilde{m}_3]. \tag{7.100}$$

We now apply iterated expectation to compute each term. For this we require the conditional means. Conditioned on $\tilde{x}_i = x$, $\tilde{m}_i = \sum_{j=1}^{x} \tilde{s}_i$ where $\tilde{s}_1$ is a Bernoulli random variable that equals one if the ith shot is made. By linearity of expectation

$$\mu_{\tilde{m}_i \mid \tilde{x}_i}(x) = \mathbb{E}\left[\sum_{j=1}^{x} \tilde{s}_i\right] \tag{7.101}$$

$$= \sum_{j=1}^{x} \mathbb{E}\left[\tilde{s}_i\right] \tag{7.102}$$

$$= x p_i. \tag{7.103}$$

By iterated expectation,

$$\mathbb{E}[\tilde{m}_i] = \mathbb{E}\left[\mu_{\tilde{m}_i \mid \tilde{x}_i}(\tilde{x}_i)\right] \tag{7.104}$$

$$= \mathbb{E}\left[p_i \tilde{x}_i\right] \tag{7.105}$$

$$= p_i \mathbb{E}(\tilde{x}_i) \tag{7.106}$$

$$= 2p_i, \tag{7.107}$$

where we have used the fact that

$$\mathbb{E}[\tilde{x}_i] = \sum_{j=0}^{4} j p_{\tilde{x}}(j) \tag{7.108}$$

$$= \frac{1+2+3+4}{5} \tag{7.109}$$

$$= 2. \tag{7.110}$$

We conclude that

$$\mathbb{E}[\tilde{y}] = \mathbb{E}[\tilde{m}_1] + 2\mathbb{E}[\tilde{m}_2] + 3\mathbb{E}[\tilde{m}_3] \tag{7.111}$$

$$= 2(p_1 + 2 \cdot p_2 + 3 \cdot p_3) \tag{7.112}$$

$$= 5.4 \text{ points}. \tag{7.113}$$

7.11 (Life expectancy) We represent the age of death as a random variable $\tilde{x}$ and define a Bernoulli random variable $\tilde{d}$ with parameter 0.25 to represent whether a person died ($\tilde{d} = 1$) in the first year or not ($\tilde{d} = 0$). We are interested in the conditional mean function $\mu_{\tilde{x} \mid \tilde{d}}$ evaluated at 0. If $\tilde{d} = 1$, then $\tilde{x} = 0$, so $\mu_{\tilde{x} \mid \tilde{d}}(1) = 0$. By iterated expectation,

$$37.5 = \mathbb{E}\left[\tilde{x}\right] = \mathbb{E}\left[\mu_{\tilde{x} \mid \tilde{d}}(\tilde{d})\right] \tag{7.114}$$

$$= 0 \cdot \mathrm{P}\left(\tilde{d} = 1\right) + \mu_{\tilde{x} \mid \tilde{d}}(0) \mathrm{P}\left(\tilde{d} = 0\right) \tag{7.115}$$

$$= 0.75 \mu_{\tilde{x} \mid \tilde{d}}(0). \tag{7.116}$$

We conclude that $\mu_{\tilde{x} \mid \tilde{d}}(0) = 50$.

7.12 (Buckets) Let 0 indicate the small bucket and 1 the large one. The mean of $\tilde{s}_1$ equals

$$\mathbb{E}\left[\tilde{s}_1\right] = \sum_{b=0}^{1} \int_{s=-\infty}^{\infty} s p_{\tilde{b}}\left(b\right) f_{\tilde{s}_1 \,|\, \tilde{b}}(s \,|\, b) \, \mathrm{d}s \tag{7.117}$$

$$= p_{\tilde{b}}\left(0\right) \int_{s=0}^{1} s \, \mathrm{d}s + p_{\tilde{b}}\left(0\right) \int_{s=0}^{2} \frac{s}{2} \, \mathrm{d}s \tag{7.118}$$

$$= \frac{3}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot 1 \tag{7.119}$$

$$= \frac{5}{8}. \tag{7.120}$$

The mean square equals

$$\mathbb{E}\left[\tilde{s}_1^2\right] = \sum_{b=0}^{1} \int_{s=-\infty}^{\infty} s^2 p_{\tilde{b}}\left(b\right) f_{\tilde{s}_1 \,|\, \tilde{b}}(s \,|\, b) \, \mathrm{d}s \tag{7.121}$$

$$= p_{\tilde{b}}\left(0\right) \int_{s=0}^{1} s^2 \, \mathrm{d}s + p_{\tilde{b}}\left(0\right) \int_{s=0}^{2} \frac{s^2}{2} \, \mathrm{d}s \tag{7.122}$$

$$= \frac{3}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2^3}{3} \tag{7.123}$$

$$= \frac{11}{12}. \tag{7.124}$$

The variance equals

$$\mathrm{Var}\left[\tilde{s}_1\right] = \mathbb{E}\left[\tilde{s}_1^2\right] - \mathbb{E}\left[\tilde{s}_1\right]^2 \tag{7.125}$$

$$= 0.526. \tag{7.126}$$

7.13 (Potatoes)

a) The minimum MSE estimate is $\mu_{\tilde{x} \,|\, \tilde{b}, \tilde{w}}(1, 1) = 50$ tons (from the corresponding graph).

b) Since we know the conditional distribution of $\tilde{x}$ given the weather and the presence of the beetle, we express the desired conditional mean function including $\tilde{b}$:

$$\mu_{\tilde{x} \,|\, \tilde{w}}(0) = \sum_{b=0}^{1} p_{\tilde{b} \,|\, \tilde{w}}(b \,|\, 0) \int_{x=-\infty}^{\infty} x f_{\tilde{x} \,|\, \tilde{w}, \tilde{b}} \, \mathrm{d}x \tag{7.127}$$

$$= \sum_{b=0}^{1} p_{\tilde{b} \,|\, \tilde{w}}(b \,|\, 0) \mu_{\tilde{x} \,|\, \tilde{w}, \tilde{b}}(0, b). \tag{7.128}$$

The conditional distribution of $\tilde{x}$ given $\tilde{b}$ and $\tilde{w}$ is Gaussian, so their means are equal to their mode, right at the center of symmetry of the pdf. From the graphs we have $\mu_{\tilde{x} \,|\, \tilde{w}, \tilde{b}}(0, 0) = 40$ and $\mu_{\tilde{x} \,|\, \tilde{w}, \tilde{b}}(0, 1) = 30$. In addition, from the table $p_{\tilde{b} \,|\, \tilde{w}}(1 \,|\, 0) = 20/30 = 2/3$ and $p_{\tilde{b} \,|\, \tilde{w}}(0 \,|\, 0) = 10/30 = 1/3$. We conclude that

$$\mu_{\tilde{x} \,|\, \tilde{w}}(0) = \frac{40}{3} + \frac{2 \cdot 30}{3} \tag{7.129}$$

$$= \frac{100}{3}. \tag{7.130}$$

7.14 (Wolf)

a)

$$\mu_{\tilde{x}[2] \mid \tilde{s}}(0) = \int_{a=-5}^{5} \int_{b=-5}^{5} \frac{b}{100} \, \mathrm{d}a \, \mathrm{d}b \tag{7.131}$$

$$= 0, \tag{7.132}$$

$$\mu_{\tilde{x}[2] \mid \tilde{s}}(1) = \int_{a=-5}^{5} \int_{b=-5}^{0} \frac{b}{50} \, \mathrm{d}a \, \mathrm{d}b \tag{7.133}$$

$$= \frac{10}{50} \left(0 - \frac{(-5)^2}{2} \right) \tag{7.134}$$

$$= -2.5. \tag{7.135}$$

By iterated expectation,

$$\mathbb{E}[\tilde{x}[2]] = \mathbb{E}\left[\mu_{\tilde{x}[2] \mid \tilde{s}}(\tilde{s}) \right] \tag{7.136}$$

$$= 0.25 \cdot (-2.5) \tag{7.137}$$

$$= -0.625. \tag{7.138}$$

b) The minimum MSE estimator of $\tilde{x}[1]$ given $\tilde{x}[2]$ is the conditional mean function of $\tilde{x}[1]$ given $\tilde{x}[2]$. We first need to compute the conditional pdf of $\tilde{x}[1]$ given $\tilde{x}[2]$.

If $0 \leq x_2 \leq 5$,

$$f_{\tilde{x}[1] \mid \tilde{x}[2]}(x_1 \mid x_2) = \frac{f_{\tilde{x}}(x_1, x_2)}{f_{\tilde{x}[2]}(x_2)} \tag{7.139}$$

$$= \frac{\sum_{s=0}^{1} p_{\tilde{s}}(s) f_{\tilde{x} \mid \tilde{s}}(x_1, x_2 \mid s)}{\int_{x_1=-\infty}^{\infty} \sum_{s=0}^{1} p_{\tilde{s}}(s) f_{\tilde{x} \mid \tilde{s}}(x_1, x_2 \mid s) \, \mathrm{d}x_1} \tag{7.140}$$

$$= \frac{0.75 \cdot \frac{1}{100}}{\int_{x_1=-5}^{5} 0.75 \cdot \frac{1}{100} \, \mathrm{d}x_1} \tag{7.141}$$

$$= \frac{1}{10}, \tag{7.142}$$

for $-5 \leq x_1 \leq 5$ and 0 otherwise.

If $-5 \leq x_2 \leq 0$,

$$f_{\tilde{x}[1] \mid \tilde{x}[2]}(x_1 \mid x_2) = \frac{f_{\tilde{x}}(x_1, x_2)}{f_{\tilde{x}[2]}(x_2)} \tag{7.143}$$

$$= \frac{\sum_{s=0}^{1} p_{\tilde{s}}(s) f_{\tilde{x} \mid \tilde{s}}(x_1, x_2 \mid s)}{\int_{x_1=-\infty}^{\infty} \sum_{s=0}^{1} p_{\tilde{s}}(s) f_{\tilde{x} \mid \tilde{s}}(x_1, x_2 \mid s) \, \mathrm{d}x_1} \tag{7.144}$$

$$= \frac{0.75 \cdot \frac{1}{100} + 0.25 \cdot \frac{1}{50}}{\int_{x_1=-5}^{5} \left(0.75 \cdot \frac{1}{100} + 0.25 \cdot \frac{1}{50} \right) \mathrm{d}x_1} \tag{7.145}$$

$$= \frac{1}{10}, \tag{7.146}$$

for $-5 \leq x_1 \leq 5$ and 0 otherwise.

It turns out that the two entries of $\tilde{x}$ are independent (you can check that the marginal pdf is also equal to 1/10). As a result, the best estimate is just the mean of $\tilde{x}[1]$, no

matter what the value of $\tilde{x}[2]$ is. For any $-5 \leq x_2 \leq 5$,

$$\mu_{\tilde{x}[1] \,|\, \tilde{x}[2]}(x_2) = \int_{x_1=-\infty}^{\infty} x_1 f_{\tilde{x}[1] \,|\, \tilde{x}[2]}(x_1 \,|\, x_2) \, \mathrm{d}x_1 \tag{7.147}$$

$$= \int_{x_1=--5}^{5} \frac{x_1}{10} \, \mathrm{d}x_1 \tag{7.148}$$

$$= 0. \tag{7.149}$$

7.15 (Law of conditional variance)

a) Recall that

$$p_{\tilde{b} \,|\, \tilde{a}}(1 \,|\, 1) = \frac{1}{7}, \qquad p_{\tilde{b} \,|\, \tilde{a}}(2 \,|\, 1) = \frac{4}{7}, \qquad p_{\tilde{b} \,|\, \tilde{a}}(3 \,|\, 1) = \frac{2}{7} \tag{7.150}$$

$$p_{\tilde{b} \,|\, \tilde{a}}(1 \,|\, 2) = \frac{2}{7}, \qquad p_{\tilde{b} \,|\, \tilde{a}}(2 \,|\, 2) = \frac{1}{7}, \qquad p_{\tilde{b} \,|\, \tilde{a}}(3 \,|\, 2) = \frac{4}{7} \tag{7.151}$$

$$p_{\tilde{b} \,|\, \tilde{a}}(1 \,|\, 3) = \frac{1}{3}, \qquad p_{\tilde{b} \,|\, \tilde{a}}(2 \,|\, 3) = \frac{1}{3}, \qquad p_{\tilde{b} \,|\, \tilde{a}}(3 \,|\, 3) = \frac{1}{3}, \tag{7.152}$$

and the conditional mean function of $\tilde{b}$ given $\tilde{a}$ is

$$\mu_{\tilde{b} \,|\, \tilde{a}}(1) = \frac{15}{7}, \tag{7.153}$$

$$\mu_{\tilde{b} \,|\, \tilde{a}}(2) = \frac{16}{7}, \tag{7.154}$$

$$\mu_{\tilde{b} \,|\, \tilde{a}}(3) = 2. \tag{7.155}$$

We have

$$\nu_{\tilde{b} \,|\, \tilde{a}}(1) = \sum_{b=1}^{3} \left(b - \frac{15}{7}\right)^2 p_{\tilde{b} \,|\, \tilde{a}}(b \,|\, 1) \tag{7.156}$$

$$= \frac{20}{49}, \tag{7.157}$$

$$\nu_{\tilde{b} \,|\, \tilde{a}}(2) = \sum_{b=1}^{3} \left(b - \frac{16}{7}\right)^2 p_{\tilde{b} \,|\, \tilde{a}}(b \,|\, 2) \tag{7.158}$$

$$= \frac{38}{49}, \tag{7.159}$$

$$\nu_{\tilde{b} \,|\, \tilde{a}}(3) = \sum_{b=1}^{3} (b-2)^2 p_{\tilde{b} \,|\, \tilde{a}}(b \,|\, 3) \tag{7.160}$$

$$= \frac{2}{3}. \tag{7.161}$$

b)

$$p_{\nu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a})}\left(\frac{20}{49}\right) = \mathrm{P}\left(\mu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a}) = \frac{20}{49}\right) \tag{7.162}$$

$$= \mathrm{P}\left(\tilde{a} = 1\right) \tag{7.163}$$

$$= 0.35. \tag{7.164}$$

Similarly,

$$p_{\mu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a})}\left(\frac{38}{49}\right) = 0.35, \tag{7.165}$$

$$p_{\mu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a})}\left(\frac{2}{3}\right) = 0.3. \tag{7.166}$$

c) We denote the conditional mean square function of $\tilde{b}$ given $\tilde{a} = a$ by $\mu_{\tilde{b}^2 \mid \tilde{a}}(a)$. By iterated expectation $\mathbb{E}\left[\mu_{\tilde{b}^2 \mid \tilde{a}}(\tilde{a})\right] = \mathbb{E}[\tilde{b}^2]$. We have

$$\begin{aligned}
\operatorname{Var}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right] &= \mathbb{E}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})^2\right] - \mathbb{E}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right]^2 && (7.167)\\
&= \mathbb{E}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})^2\right] - \mathbb{E}\left[\tilde{b}\right]^2, && (7.168)\\
\mathbb{E}\left[\nu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right] &= \mathbb{E}\left[\mu_{\tilde{b}^2 \mid \tilde{a}}(\tilde{a}) - \mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})^2\right] && (7.169)\\
&= \mathbb{E}[\tilde{b}^2] - \mathbb{E}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})^2\right]. && (7.170)
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\left[\nu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right] + \operatorname{Var}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right] &= \mathbb{E}[\tilde{b}^2] - \mathbb{E}\left[\tilde{b}\right]^2 && (7.171)\\
&= \operatorname{Var}[\tilde{b}]. && (7.172)
\end{aligned}$$

d) The pmf of the conditional mean is

$$\begin{aligned}
p_{\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})}\left(\frac{15}{7}\right) &= 0.35, && (7.173)\\
p_{\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})}\left(\frac{16}{7}\right) &= 0.35, && (7.174)\\
p_{\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})}\left(2\right) &= 0.3. && (7.175)
\end{aligned}$$

The mean square of $\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})$ therefore equals

$$\begin{aligned}
\mathbb{E}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})^2\right] &= \left(\frac{15}{7}\right)^2 \cdot 0.35 + \left(\frac{16}{7}\right)^2 \cdot 0.35 + 2^2 \cdot 0.3 && (7.176)\\
&= 4.636, && (7.177)
\end{aligned}$$

and its mean is 2.15, so the variance of the conditional mean equals

$$\operatorname{Var}\left[\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right] = 4.636 - 2.15^2 = 0.0132. \quad (7.178)$$

The mean of the conditional variance is

$$\begin{aligned}
\mathbb{E}\left[\nu_{\tilde{b} \mid \tilde{a}}(\tilde{a})\right] &= \sum_{a=1}^{3} \nu_{\tilde{b} \mid \tilde{a}}(a) p_{\tilde{a}}(a) && (7.179)\\
&= 0.35 \cdot \frac{20}{49} + 0.35 \cdot \frac{38}{49} + 0.3 \cdot \frac{2}{3} && (7.180)\\
&= 0.6143. && (7.181)
\end{aligned}$$

By the law of conditional variance the variance of $\tilde{b}$ should equal $0.0132 + 0.6143 = 0.6275$. We verify this using the marginal pmf of $\tilde{b}$. The mean square of $\tilde{b}$ equals

$$\begin{aligned}
\mathbb{E}\left[\tilde{b}^2\right] &= \sum_{b=1}^{3} b^2 p_{\tilde{b}}(b) && (7.182)\\
&= 1 \cdot 0.25 + 4 \cdot 0.35 + 9 \cdot 0.4 && (7.183)\\
&= 5.25, && (7.184)
\end{aligned}$$

so the variance equals $5.25 - 2.15^2 = 0.6275$.

7.16 (Faulty database)

a) Let $\tilde{c}$ be a random variable that represents whether the entry is corrupted, we have

$$\tilde{b} = \begin{cases} -x & \text{if } \tilde{c} = -1, \\ \tilde{a} & \text{if } \tilde{c} = 0, \\ x & \text{if } \tilde{c} = 1. \end{cases} \tag{7.185}$$

The conditional mean function of $\tilde{b}$ given $\tilde{c} = c$ equals

$$\mu_{\tilde{b} \mid \tilde{c}}(c) = \begin{cases} -x & \text{if } c = -1, \\ 0 & \text{if } c = 0, \\ x & \text{if } c = 1, \end{cases} \tag{7.186}$$

so by iterated expectation

$$\mathbb{E}[\tilde{b}] = \mathbb{E}\left[\mu_{\tilde{b} \mid \tilde{c}}(\tilde{c})\right] \tag{7.187}$$

$$= 0.1(-x) + 0.1x = 0. \tag{7.188}$$

The conditional mean function of $\tilde{b}^2$ given $\tilde{c} = c$ equals

$$\mu_{\tilde{b}^2 \mid \tilde{c}}(c) = \begin{cases} x^2 & \text{if } c = -1, \\ \sigma^2 & \text{if } c = 0, \\ x^2 & \text{if } c = 1, \end{cases} \tag{7.189}$$

so by iterated expectation

$$\mathbb{E}[\tilde{b}^2] = \mathbb{E}\left[\mu_{\tilde{b}^2 \mid \tilde{c}}(\tilde{c})\right] \tag{7.190}$$

$$= 0.2x^2 + 0.8\sigma^2. \tag{7.191}$$

We conclude that $\text{Var}[\tilde{b}] = \mathbb{E}[\tilde{b}^2] - \mathbb{E}[\tilde{b}]^2 = 0.2x^2 + 0.8\sigma^2$.

b) The minimum mean-squared error estimator of $\tilde{a}$ given $\tilde{b} = b$ is the conditional mean $\mu_{\tilde{a} \mid \tilde{b}}$ of $\tilde{a}$ given $\tilde{b} = b$. Conditioned on $\tilde{b} = -x$ or $\tilde{b} = x$, the distribution of $\tilde{a}$ is Gaussian with zero mean and variance σ^2 because the database failure is independent from the content of the database entry. Therefore $\mu_{\tilde{a} \mid \tilde{b}}(-x) = \mu_{\tilde{a} \mid \tilde{b}}(x) = 0$. Conditioned on $\tilde{b} = b \notin \{-x, x\}$, $\tilde{a}$ is equal to b with probability one, so $\mu_{\tilde{a} \mid \tilde{b}}(b) = b$ for all $b \notin \{-x, x\}$.

7.17 (Coffee and life expectancy)

a) Let $\tilde{\ell}$, $\tilde{c}$ and $\tilde{s}$ represent life span, smoking and coffee respectively. Since have access to the conditional mean given both smoking and coffee $\mu_{\tilde{\ell} \mid \tilde{c}, \tilde{s}}$,

$$\mu_{\tilde{\ell} \mid \tilde{c}}(1) = \int_{\ell=-\infty}^{\infty} \ell f_{\tilde{\ell} \mid \tilde{c}}(\ell \mid 1) \, \mathrm{d}\ell \tag{7.192}$$

$$= \int_{\ell=-\infty}^{\infty} \ell \sum_{s=0}^{1} p_{\tilde{s} \mid \tilde{c}}(s \mid 1) f_{\tilde{\ell} \mid \tilde{c}}(\ell \mid 1, s) \, \mathrm{d}\ell \tag{7.193}$$

$$= \sum_{s=0}^{1} p_{\tilde{s} \mid \tilde{c}}(s \mid 1) \mu_{\tilde{\ell} \mid \tilde{c}, \tilde{s}}(1, s) \tag{7.194}$$

$$= \frac{80}{120} \cdot 68 + \frac{40}{120} \cdot 82 \tag{7.195}$$

$$= 72.7. \tag{7.196}$$

Similarly,

$$\mu_{\tilde{\ell}\,|\,\tilde{c}}(0) = \sum_{s=0}^{1} p_{\tilde{s}\,|\,\tilde{c}}(s\,|\,0)\mu_{\tilde{\ell}\,|\,\tilde{c},\tilde{s}}(0,s) \tag{7.197}$$

$$= \frac{20}{80}\cdot 70 + \frac{60}{80}\cdot 80 \tag{7.198}$$

$$= 77.5. \tag{7.199}$$

The observed ATE is therefore $72.7 - 77.5 = -4.8$ years.

b) From the tables $p_{\tilde{s}}(0) = 100/200 = 1/2$ and $p_{\tilde{s}}(1) = 1/2$. By the formula for the adjusted ATE given the conditional independence assumption,

$$\text{ATE} = \sum_{s=0}^{1} p_{\tilde{s}}(s)\mu_{\tilde{\ell}\,|\,\tilde{c},\tilde{s}}(1,s) - \sum_{s=0}^{1} p_{\tilde{s}}(s)\mu_{\tilde{y}\,|\,\tilde{c},\tilde{s}}(0,s) \tag{7.200}$$

$$= \frac{1}{2}\cdot 68 + \frac{1}{2}\cdot 82 - \frac{1}{2}\cdot 70 + \frac{1}{2}\cdot 80 \tag{7.201}$$

$$= 0. \tag{7.202}$$

The adjusted ATE is zero, which indicates that coffee has no effect on life span.

7.18 (Weight-loss supplement)

a) The unadjusted ATE equals

$$\begin{aligned}&\text{unadjusted ATE}\\ &= \frac{0.5 - 4.6 - 7.5 + 1.1 - 3.4 - 1.4 + 3.2 - 1.9 - 3.6 + 2.4 + 1.2 + 4.2 - 2.5 - 3.5 + 4.7}{15}\\ &\quad - \frac{1.6 - 6.7 + 2.3 - 4.5 + 3.4 - 0.6 + 2.3 + 1.5 + 1.7 + 2.6 + 0.5 + 1.2 - 3.6 + 2.1 - 1.2 + 0.4}{16}\\ &= -0.93.\end{aligned} \tag{7.203}$$

b) We model the treatment as a random variable $\tilde{t}$, the weight loss as a random variable $\tilde{w}$ and the exercise as a random variable $\tilde{x}$. In order to adjust for the effect of exercise, we compute

$$p_{\tilde{x}}(1) = \frac{13}{31} = 0.42, \tag{7.204}$$

$$\mu_{\tilde{w}\,|\,\tilde{t},\tilde{x}}(0,0) = \frac{1.6 + 2.3 - 0.6 + 2.3 + 1.5 + 1.7 + 2.6 + 0.5 + 1.2 + 2.1 - 1.2 + 0.4}{12} \tag{7.205}$$

$$= 1.2, \tag{7.206}$$

$$\mu_{\tilde{w}\,|\,\tilde{t},\tilde{x}}(0,1) = \frac{-6.7 - 4.5 + 3.4 - 3.6}{4} = -2.85, \tag{7.207}$$

$$\mu_{\tilde{w}\,|\,\tilde{t},\tilde{x}}(1,0) = \frac{0.5 + 1.1 - 1.9 + 2.4 + 1.2 + 4.7}{6} = 1.33, \tag{7.208}$$

$$\mu_{\tilde{w}\,|\,\tilde{t},\tilde{x}}(1,1) = \frac{-4.6 - 7.5 - 3.4 - 1.4 + 3.2 - 3.6 + 4.2 - 2.5 - 3.5}{9} = -2.12, \tag{7.209}$$

The adjusted ATE equals

$$\text{adjusted ATE} = \sum_{x=0}^{1} p_{\tilde{x}}(x)\mu_{\tilde{w}\,|\,\tilde{t},\tilde{x}}(1,x) - \sum_{x=0}^{1} p_{\tilde{x}}(x)\mu_{\tilde{w}\,|\,\tilde{t},\tilde{x}}(0,x) \tag{7.210}$$

$$= (0.58\cdot 1.33 - 0.42\cdot 2.12) - (0.58\cdot 1.2 - 0.42\cdot 2.85) = 0.382. \tag{7.211}$$

This approximates the true ATE if the potential outcomes of the weight gain/loss for treatment and control are conditionally independent of the treatment given the exercise.

7.19 (Chocolate bar)

a) We model taking the chocolate bar as a treatment $\tilde{t}$, the cholesterol as a random variable $\tilde{y}$ and the sex of the participants as another random variable $\tilde{s}$. From the information in the problem,

$$\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(1, \text{man}) = 150, \tag{7.212}$$

$$\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(0, \text{man}) = 140, \tag{7.213}$$

$$\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(1, \text{woman}) = 130, \tag{7.214}$$

$$\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(0, \text{woman}) = 120. \tag{7.215}$$

and

$$p_{\tilde{s}\,|\,\tilde{t}}(\text{man}\,|\,1) = \frac{p_{\tilde{s}}(\text{man})p_{\tilde{t}\,|\,\tilde{s}}(1\,|\,\text{man})}{p_{\tilde{s}}(\text{man})p_{\tilde{t}\,|\,\tilde{s}}(1\,|\,\text{man}) + p_{\tilde{s}}(\text{woman})p_{\tilde{t}\,|\,\tilde{s}}(1\,|\,\text{woman})} \tag{7.216}$$

$$= \frac{\alpha}{\alpha + 0.5}, \tag{7.217}$$

$$p_{\tilde{s}\,|\,\tilde{t}}(\text{man}\,|\,0) = \frac{1-\alpha}{1-\alpha+0.5} \tag{7.218}$$

The average cholesterol in the treatment group is

$$\mu_{\tilde{y}\,|\,\tilde{t}}(1) = \sum_{s\in\{\text{man},\text{woman}\}} p_{\tilde{s}\,|\,\tilde{t}}(s\,|\,1) \int_{y=-\infty}^{\infty} y f_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(y\,|\,1,s)\ \mathrm{d}y \tag{7.219}$$

$$= p_{\tilde{s}\,|\,\tilde{t}}(\text{man}\,|\,1)\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(1,\text{man}) + p_{\tilde{s}\,|\,\tilde{t}}(\text{woman}\,|\,1)\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(1,\text{woman}) \tag{7.220}$$

$$= \frac{150\alpha}{\alpha+0.5} + \frac{0.5\cdot 130}{\alpha+0.5}. \tag{7.221}$$

By the same argument, the average cholesterol in the control group is

$$\mu_{\tilde{y}\,|\,\tilde{t}}(0) = p_{\tilde{s}\,|\,\tilde{t}}(\text{man}\,|\,0)\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(0,\text{man}) + p_{\tilde{s}\,|\,\tilde{t}}(\text{woman}\,|\,0)\mu_{\tilde{y}\,|\,\tilde{t},\tilde{s}}(0,\text{woman}) \tag{7.222}$$

$$= \frac{(1-\alpha)140}{1.5-\alpha} + \frac{0.5\cdot 120}{1.5-\alpha}. \tag{7.223}$$

The observed ATE equals

$$\text{observed ATE}(\alpha) = \frac{150\alpha+65}{\alpha+0.5} - \frac{200-140\alpha}{1.5-\alpha}. \tag{7.224}$$

b)

$$\text{observed ATE}(0.05) = -1.29, \tag{7.225}$$

$$\text{observed ATE}(0.5) = 10, \tag{7.226}$$

$$\text{observed ATE}(0.95) = 21.3. \tag{7.227}$$

8

Correlation

Exercises

8.1 (Three variables) Yes, it is possible. Let $\tilde{a}$, $\tilde{b}$ and $\tilde{c}$ be random variables with zero mean and unit variance, which are pairwise uncorrelated. We define:

$$\tilde{w} := \tilde{a} + \alpha\tilde{b}, \tag{8.1}$$
$$\tilde{y} := \tilde{b} + \beta\tilde{c}, \tag{8.2}$$
$$\tilde{z} := \gamma\tilde{a} + \tilde{c}, \tag{8.3}$$

where α, β and γ are real-valued constants. By linearity of expectation

$$\mathbb{E}\left[\tilde{w}\right] = \mathbb{E}\left[\tilde{a}\right] + \alpha\mathbb{E}[\tilde{b}] = 0, \tag{8.4}$$
$$\mathbb{E}\left[\tilde{y}\right] = \mathbb{E}[\tilde{b}] + \beta\mathbb{E}\left[\tilde{c}\right] = 0, \tag{8.5}$$
$$\mathbb{E}\left[\tilde{z}\right] = \gamma\mathbb{E}\left[\tilde{a}\right] + \mathbb{E}\left[\tilde{c}\right] = 0. \tag{8.6}$$

Since $\tilde{a}$, $\tilde{b}$ and $\tilde{c}$ have zero mean and are pairwise uncorrelated,

$$\mathbb{E}[\tilde{a}\tilde{b}] = \mathbb{E}\left[\tilde{a}\tilde{c}\right] = \mathbb{E}[\tilde{b}\tilde{c}] = 0. \tag{8.7}$$

Consequently, again by linearity of expectation,

$$\mathbb{E}\left[\tilde{w}\tilde{y}\right] = \mathbb{E}[\tilde{a}\tilde{b}] + \beta\mathbb{E}\left[\tilde{a}\tilde{c}\right] + \alpha\mathbb{E}[\tilde{b}^2] + \alpha\beta\mathbb{E}[\tilde{b}\tilde{c}] \tag{8.8}$$
$$= \alpha, \tag{8.9}$$
$$\mathbb{E}\left[\tilde{w}\tilde{z}\right] = \gamma\mathbb{E}\left[\tilde{a}^2\right] + \mathbb{E}\left[\tilde{a}\tilde{c}\right] + \gamma\alpha\mathbb{E}[\tilde{a}\tilde{b}] + \alpha\mathbb{E}[\tilde{b}\tilde{c}] \tag{8.10}$$
$$= \gamma, \tag{8.11}$$
$$\mathbb{E}\left[\tilde{y}\tilde{z}\right] = \gamma\mathbb{E}[\tilde{a}\tilde{b}] + \mathbb{E}[\tilde{b}\tilde{c}] + \gamma\beta\mathbb{E}\left[\tilde{a}\tilde{c}\right] + \beta\mathbb{E}\left[\tilde{c}^2\right] \tag{8.12}$$
$$= \beta. \tag{8.13}$$

If we choose α and β to be positive and γ negative, then $\tilde{w}$ and $\tilde{y}$ are positively correlated, and so are $\tilde{y}$ and $\tilde{z}$, but $\tilde{w}$ and $\tilde{y}$ are negatively correlated.

8.2 (Properties of covariance)

a) By the definition of covariance and linearity of expectation

$$\mathrm{Cov}[\beta\tilde{a} + \alpha, \tilde{b}] := \mathbb{E}\left[(\beta\tilde{a} + \alpha - \mathbb{E}[\beta\tilde{a} + \alpha])(\tilde{b} - \mathbb{E}[\tilde{b}])\right] \tag{8.14}$$
$$= \beta\mathbb{E}\left[(\tilde{a} - \mathbb{E}[\tilde{a}])(\tilde{b} - \mathbb{E}[\tilde{b}])\right] \tag{8.15}$$
$$= \beta\mathrm{Cov}[\tilde{a}, \tilde{b}]. \tag{8.16}$$

b) Since $-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$ and $\rho_{\tilde{a},\tilde{b}} = \frac{\text{Cov}[\tilde{a},\tilde{b}]}{\sqrt{\text{Var}[\tilde{a}]\text{Var}[\tilde{b}]}}$, it follows that

$$-\sqrt{\text{Var}[\tilde{a}]\text{Var}[\tilde{b}]} \leq \text{Cov}[\tilde{a},\tilde{b}] \leq \sqrt{\text{Var}[\tilde{a}]\text{Var}[\tilde{b}]}. \tag{8.17}$$

c) In the proof of Theorem 8.34 we show that the covariance is symmetric and that for any random variables $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{y}$, $\text{Cov}[\tilde{x}_1 + \tilde{x}_2, \tilde{y}] = \text{Cov}[\tilde{x}_1, \tilde{y}] + \text{Cov}[\tilde{x}_2, \tilde{y}]$. Consequently,

$$\begin{aligned}
\text{Cov}[\tilde{a} + \tilde{b}, \tilde{a} - \tilde{b}] &= \text{Cov}[\tilde{a}, \tilde{a} - \tilde{b}] + \text{Cov}[\tilde{b}, \tilde{a} - \tilde{b}] && (8.18)\\
&= \text{Cov}[\tilde{a}, \tilde{a}] - \text{Cov}[\tilde{a}, \tilde{b}] + \text{Cov}[\tilde{b}, \tilde{a}] - \text{Cov}[\tilde{b}, \tilde{b}] && (8.19)\\
&= \text{Var}[\tilde{a}] - \text{Var}[\tilde{b}], && (8.20)
\end{aligned}$$

because $\text{Cov}\left[\tilde{a}, \tilde{a}\right] = \mathbb{E}\left[\left(\tilde{a} - \mathbb{E}\left[\tilde{a}\right]\right)^2\right] = \text{Var}[\tilde{a}]$.

8.3 (Dunning-Kruger effect) Since the means of all variables are zero,

$$\begin{aligned}
\text{Cov}\left[\tilde{t}, \tilde{d}\right] &= \mathbb{E}\left[\tilde{t}\left(\tilde{s} - \tilde{t}\right)\right] && (8.21)\\
&= \mathbb{E}\left[\tilde{t}, \tilde{s}\right] - \mathbb{E}\left[\tilde{t}^2\right] && (8.22)\\
&= -\text{Var}\left[\tilde{t}\right] = 1, && (8.23)
\end{aligned}$$

and by the uncorrelation assumption

$$\begin{aligned}
\text{Var}\left[\tilde{d}\right] &= \text{Var}\left[\tilde{s} - \tilde{t}\right] && (8.24)\\
&= \text{Var}\left[\tilde{s}\right] + \text{Var}\left[-\tilde{t}\right] && (8.25)\\
&= \text{Var}\left[\tilde{s}\right] + \text{Var}\left[\tilde{t}\right] && (8.26)\\
&= 2. && (8.27)
\end{aligned}$$

Consequently,

$$\begin{aligned}
\rho_{\tilde{t},\tilde{d}} &= \frac{\text{Cov}\left[\tilde{t}, \tilde{d}\right]}{\sqrt{\text{Var}\left[\tilde{t}\right]\text{Var}\left[\tilde{d}\right]}} && (8.28)\\
&= \frac{-1}{\sqrt{2}} = -0.707. && (8.29)
\end{aligned}$$

The study finds a strong negative correlation, even though the true and self-evaluated competences are uncorrelated!

8.4 (Cross) The conditional mean $\mu_{\tilde{y} \mid \tilde{x}}(x)$ is zero for all values of x with nonzero density, because all the conditional distributions are uniform and centered at zero. By iterated expectation $\mathbb{E}\left[\tilde{y}\right] = \mathbb{E}\left[\mu_{\tilde{y} \mid \tilde{x}}(\tilde{x})\right] = 0$. Also,

$$\begin{aligned}
\mu_{\tilde{x}\tilde{y} \mid \tilde{x}}(x) &= x\mu_{\tilde{y} \mid \tilde{x}}(x) && (8.30)\\
&= 0 && (8.31)
\end{aligned}$$

for all x with nonzero density. By iterated expectation $E\left[\tilde{x}\tilde{y}\right] = \mathbb{E}\left[\mu_{\tilde{x}\tilde{y} \mid \tilde{x}}(\tilde{x})\right] = 0$. The variables are uncorrelated.

8.5 (Rufus) The mean of $\tilde{x}$ equals

$$\begin{aligned}
\mathbb{E}\left[\tilde{x}\right] = &\int_{y=-50}^{50} \left(\int_{x=-50}^{-20} cx \, \mathrm{d}x + \int_{x=20}^{50} cx \, \mathrm{d}x \right) \mathrm{d}y && (8.32)\\
&+ \int_{x=-20}^{20} cx \, \mathrm{d}x \left(\int_{y=-50}^{-20} \mathrm{d}y + \int_{y=20}^{50} cx \, \mathrm{d}y \right), && (8.33)
\end{aligned}$$

where $c = 1/8400$. Since by a change of variables $t = -x$ $\int_{x=-50}^{-20} cx \, \mathrm{d}x = -\int_{x=20}^{50} cx \, \mathrm{d}x$, and $\int_{x=-20}^{20} cx \, \mathrm{d}x = 0$, we have $\mathbb{E}[\tilde{x}] = 0$.
Since the mean of $\tilde{x}$ is zero, to compute the covariance, we only need to compute $\mathbb{E}[\tilde{x}\tilde{y}]$. By iterated expectation, we can take the mean of $\mu_{\tilde{x}\tilde{y}}(\tilde{y})$, which is computed by first deriving the conditional mean function $\mu_{\tilde{x}\tilde{y}}(y)$ for all possible values of y and plugging in $\tilde{y}$. For y in $[-50, -20]$ or $[20, 50]$, we have

$$\mu_{\tilde{x}\tilde{y}}(y) = \int_{x=-50}^{50} \frac{yx}{100} \, \mathrm{d}x \tag{8.34}$$

$$= \frac{yx}{100} \int_{x=-50}^{50} x \, \mathrm{d}x \tag{8.35}$$

$$= 0. \tag{8.36}$$

For y in $[-20, 20]$, similarly

$$\mu_{\tilde{x}\tilde{y}}(y) = \frac{1}{60} \left(\int_{x=-50}^{-20} x \, \mathrm{d}x + \int_{x=20}^{50} x \, \mathrm{d}x \right) \tag{8.37}$$

$$= 0. \tag{8.38}$$

The conditional mean of $\tilde{x}\tilde{y}$ is always zero for values of $\tilde{y}$ that have nonzero probability. By iterated expectation this implies that the covariance is zero, so the variables are uncorrelated.

8.6 (Bernoulli random variables) Let $\tilde{a}$ and $\tilde{b}$ be two uncorrelated Bernoulli random variables with parameters $\theta_{\tilde{a}}$ and $\theta_{\tilde{b}}$ defined in the same probability space. Their means equal $\theta_{\tilde{a}}$ and $\theta_{\tilde{b}}$ respectively, so

$$\mathrm{Cov}[\tilde{a}, \tilde{b}] = \mathbb{E}[\tilde{a}\tilde{b}] - \theta_{\tilde{a}}\theta_{\tilde{b}}. \tag{8.39}$$

Since

$$\mathbb{E}[\tilde{a}\tilde{b}] = \sum_{a \in \{0,1\}} \sum_{b \in \{0,1\}} ab p_{\tilde{a},\tilde{b}}(a, b) \tag{8.40}$$

$$= p_{\tilde{a},\tilde{b}}(1, 1), \tag{8.41}$$

so

$$p_{\tilde{a},\tilde{b}}(1, 1) = \theta_{\tilde{a}}\theta_{\tilde{b}} = p_{\tilde{a}}(1) p_{\tilde{b}}(1). \tag{8.42}$$

In addition,

$$\theta_{\tilde{a}} = p_{\tilde{a}}(1) = p_{\tilde{a},\tilde{b}}(1, 0) + p_{\tilde{a},\tilde{b}}(1, 1) \tag{8.43}$$

$$= p_{\tilde{a},\tilde{b}}(1, 0) + \theta_{\tilde{a}}\theta_{\tilde{b}}, \tag{8.44}$$

so

$$p_{\tilde{a},\tilde{b}}(1, 0) = \theta_{\tilde{a}}(1 - \theta_{\tilde{b}}) = p_{\tilde{a}}(1) p_{\tilde{b}}(0) \tag{8.45}$$

and by the same argument $p_{\tilde{a},\tilde{b}}(0, 1) = p_{\tilde{a}}(0) p_{\tilde{b}}(1)$. Finally,

$$1 - \theta_{\tilde{a}} = p_{\tilde{a}}(0) = p_{\tilde{a},\tilde{b}}(0, 0) + p_{\tilde{a},\tilde{b}}(0, 1) \tag{8.46}$$

$$= p_{\tilde{a},\tilde{b}}(0, 0) + (1 - \theta_{\tilde{a}})\theta_{\tilde{b}}, \tag{8.47}$$

so

$$p_{\tilde{a},\tilde{b}}(0,0) = (1-\theta_{\tilde{a}})(1-\theta_{\tilde{b}}) = p_{\tilde{a}}(0)p_{\tilde{b}}(0). \tag{8.48}$$

We conclude that uncorrelated Bernoulli random variables are independent.

8.7 (Sample mean and variance of standardized data) We define $X := \{x_1, x_2, \ldots, x_n\}$ and $Y := \{y_1, y_2, \ldots, y_n\}$.

$$\begin{aligned} m(S_X) &= \frac{1}{n}\sum_{i=1}^{n} s(x_i) && (8.49)\\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - m(X)}{\sqrt{v(X)}} && (8.50)\\ &= \frac{1}{\sqrt{v(X)}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} m(X)\right) && (8.51)\\ &= \frac{m(X) - m(X)}{\sqrt{v(X)}} = 0. && (8.52) \end{aligned}$$

By the same argument, $m(Y) = 0$.

$$\begin{aligned} v(S_X) &= \frac{1}{n-1}\sum_{i=1}^{n} \left(s(x_i) - m(S_X)\right)^2 && (8.53)\\ &= \frac{1}{n-1}\sum_{i=1}^{n} s(x_i)^2 && (8.54)\\ &= \frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{x_i - m(X)}{\sqrt{v(X)}}\right)^2 && (8.55)\\ &= \frac{1}{v(X)}\frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - m(X)\right)^2 && (8.56)\\ &= 1. && (8.57) \end{aligned}$$

By the same argument, $v(Y) = 1$.

$$\begin{aligned} c(S_X, S_Y) &= \frac{1}{n-1}\sum_{i=1}^{n} (s(x_i) - m(S_X))(s(x_i) - m(S_Y)) && (8.58)\\ &= \frac{1}{n-1}\sum_{i=1}^{n} s(x_i)s(y_i) && (8.59)\\ &= \frac{1}{n-1}\sum_{i=1}^{n} \frac{\left(x_i - m(X)\right)\left(y_i - m(Y)\right)}{\sqrt{v(X)v(Y)}} && (8.60)\\ &= \rho_{X,Y}. && (8.61) \end{aligned}$$

8.8 (Properties of the sample correlation coefficient)

a) By the results in Exercise 8.7 and the formula for the OLS estimator,

$$\frac{1}{n-1}\sum_{i=1}^{n} r_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \sqrt{v(Y)}\rho_{X,Y}\left(\frac{x-m(X)}{\sqrt{v(X)}}\right) - m(Y)\right)^2 \tag{8.62}$$

$$= \frac{v(Y)}{n-1}\sum_{i=1}^{n}\left(s(y_i) - \rho_{X,Y}s(x_i)\right)^2 \tag{8.63}$$

$$= \frac{v(Y)}{n-1}\left(\sum_{i=1}^{n} s(y_i)^2 + \rho_{X,Y}\sum_{i=1}^{n} s(x_i)^2 - 2\rho_{X,Y}\sum_{i=1}^{n} s(x_i)s(y_i)\right) \tag{8.64}$$

$$= v(Y)\left(1-\rho_{X,Y}^2\right). \tag{8.65}$$

b) By (8.65),

$$\left(1-\rho_{X,Y}^2\right)v(Y) \geq \frac{1}{n-1}\sum_{i=1}^{n} r_i^2 \geq 0, \tag{8.66}$$

which implies that $\rho_{X,Y}^2 \leq 1$, unless $v(Y)$ is zero.

c) By (8.65), if

$$\rho_{X,Y} = \pm 1 \tag{8.67}$$

then

$$\frac{1}{n-1}\sum_{i=1}^{n} r_i^2 = v(Y)\left(1-\rho_{X,Y}^2\right) = 0, \tag{8.68}$$

which implies that $r_i = 0$ for $1 \leq i \leq n$, so that

$$y_i = \ell_{\mathrm{OLS}}(x_i) \tag{8.69}$$

$$= \sqrt{v(Y)}\rho_{X,Y}\left(\frac{x-m(X)}{\sqrt{v(X)}}\right) + m(Y). \tag{8.70}$$

8.9 (Decomposition of sample variance and coefficient of determination)

a) The residual has mean zero,

$$m(R) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sqrt{v(Y)}\rho_{X,Y}\left(\frac{x-m(X)}{\sqrt{v(X)}}\right) - m(Y)\right) \tag{8.71}$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} m(Y) - \frac{\sqrt{v(Y)}\rho_{X,Y}}{\sqrt{v(X)}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} m(X)\right) \tag{8.72}$$

$$= m(Y) - m(Y) - \frac{\sqrt{v(Y)}\rho_{X,Y}}{\sqrt{v(X)}}\left(m(X) - m(X)\right) \tag{8.73}$$

$$= 0. \tag{8.74}$$

Therefore, by the results in Exercise 8.7 and the formula for the OLS estimator,

$$c(X,R) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m(X))\left(y_i - \sqrt{v(Y)}\rho_{X,Y}\left(\frac{x - m(X)}{\sqrt{v(X)}}\right) - m(Y)\right) \tag{8.75}$$

$$= \frac{\sqrt{v(X)v(Y)}}{n-1}\sum_{i=1}^{n} s(x_i)\left(s(y_i) - \rho_{X,Y}s(x_i)\right) \tag{8.76}$$

$$= \sqrt{v(X)v(Y)}\left(\frac{1}{n-1}\sum_{i=1}^{n} s(x_i)s(y_i) - \rho_{X,Y}\frac{1}{n-1}\sum_{i=1}^{n} s(x_i)^2\right) \tag{8.77}$$

$$= \sqrt{v(X)v(Y)}\left(\rho_{X,Y}^2 - \rho_{X,Y}^2\right) = 0. \tag{8.78}$$

b) The sample mean is a linear operation,

$$m(M) = \frac{1}{n}\sum_{i=1}^{n}(a_i + b_i) \tag{8.79}$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_i + \frac{1}{n}\sum_{i=1}^{n} b_i \tag{8.80}$$

$$= m(A) + m(B), \tag{8.81}$$

so

$$v(M) = \frac{1}{n-1}\sum_{i=1}^{n}(a_i + b_i - m(A+B))^2 \tag{8.82}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(a_i - m(A) + b_i - m(B))^2 \tag{8.83}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(a_i - m(A))^2 + \frac{1}{n-1}\sum_{i=1}^{n}(b_i - m(B))^2 \tag{8.84}$$

$$+ \frac{2}{n-1}\sum_{i=1}^{n}(a_i - m(A))(b_i - m(B)) \tag{8.85}$$

$$= v(A) + v(B) + 2c(A,B). \tag{8.86}$$

c) The sample mean of the scaled dataset satisfies,

$$m(A_{\beta,\alpha}) = \frac{1}{n}\sum_{i=1}^{n}(\beta a_i + \alpha) \tag{8.87}$$

$$= \beta m(A) + \alpha, \tag{8.88}$$

so

$$c(A_{\beta,\alpha}, B) = \frac{1}{n-1}\sum_{i=1}^{n}(\beta a_i + \alpha - \beta m(A) - \alpha)(b_i - m(B)) \tag{8.89}$$

$$= \frac{\beta}{n-1}\sum_{i=1}^{n}(a_i - m(A))(b_i - m(B)) \tag{8.90}$$

$$= \beta c(A,B). \tag{8.91}$$

By (8.78) $c(X, R) = 0$, so $c(L, R) = 0$ because $\ell_{\text{OLS}}(x_i)$ is an affine function of x_i. Therefore, by (8.86)

$$v(Y) = v(L) + v(R). \tag{8.92}$$

d) By (8.74) the sample of mean of the residual is zero, so by (8.65)

$$v(R) = \frac{1}{n-1} \sum_{i=1}^{n} r_i^2 \tag{8.93}$$

$$= v(Y) \left(1 - \rho_{X,Y}^2\right). \tag{8.94}$$

By (8.92) we conclude that

$$R^2 := \frac{v(L)}{v(Y)} \tag{8.95}$$

$$= \frac{v(Y) - v(R)}{v(Y)} \tag{8.96}$$

$$= \rho_{X,Y}^2. \tag{8.97}$$

8.10 (Noisy measurement)

a) We have

$$\text{Var}[\tilde{y}] = \text{Var}[\tilde{x}] + \text{Var}[\tilde{z}] \tag{8.98}$$

$$= 1 + \sigma^2, \tag{8.99}$$

$$\mathbb{E}[\tilde{x}\tilde{y}] = \mathbb{E}[\tilde{x}^2] + \mathbb{E}[\tilde{x}\tilde{z}] \tag{8.100}$$

$$= 1, \tag{8.101}$$

$$\rho_{\tilde{x},\tilde{y}} = \frac{\mathbb{E}[\tilde{x}\tilde{y}]}{\sqrt{1+\sigma^2}} \tag{8.102}$$

$$= \frac{1}{\sqrt{1+\sigma^2}}. \tag{8.103}$$

The linear MMSE estimate of $\tilde{x}$ given $\tilde{y} = y$ equals

$$\ell_{\text{MMSE}}(y) = \frac{\rho_{\tilde{x},\tilde{y}} \sigma_{\tilde{x}} y}{\sigma_{\tilde{y}}} \tag{8.104}$$

$$= \frac{y}{1+\sigma^2}. \tag{8.105}$$

b)

$$\mathbb{E}\left[\left(\tilde{x} - \ell_{\text{MMSE}}(\tilde{y})\right)^2\right] = (1 - \rho_{\tilde{x},\tilde{y}}^2)\sigma_{\tilde{x}}^2 \tag{8.106}$$

$$= \frac{\sigma^2}{1+\sigma^2}. \tag{8.107}$$

c) If $\sigma \to 0$ the estimate equals y and the error is zero, which makes sense since in that case the noise is zero and hence $\tilde{y} = \tilde{x}$ with probability one.

d) If $\sigma \to \infty$ the estimate equals zero and the error is 1, which makes sense since in that case we are better off just ignoring the measurement and setting the estimate equal to the mean of $\tilde{x}$. The corresponding error is consequently the variance of $\tilde{x}$.

8.11 (Affine function) Since $\tilde{c}$ is an affine function of $\tilde{b}$, the standardized random variables $s(\tilde{c})$ and $s(\tilde{b})$ are the same, so $\rho_{\tilde{a},\tilde{c}} = \rho_{s(\tilde{a}),s(\tilde{c})} = \rho_{s(\tilde{a}),s(\tilde{b})} = 0.25$. The mean and variance of $\tilde{c}$ equal

$$\mu_{\tilde{c}} = \mathbb{E}(2\tilde{b} + 3) \tag{8.108}$$
$$= 2\mathbb{E}(\tilde{b}) + 3 \tag{8.109}$$
$$= -1, \tag{8.110}$$
$$\sigma_{\tilde{c}}^2 = \mathrm{Var}(2\tilde{b} + 3) \tag{8.111}$$
$$= 4\mathrm{Var}(\tilde{b}) \tag{8.112}$$
$$= 36. \tag{8.113}$$

The linear minimum-mean-squared-error estimate of $\tilde{c}$ given $\tilde{a}$ is therefore

$$\frac{\sigma_{\tilde{c}}\rho_{\tilde{a},\tilde{c}}(a - \mu_{\tilde{a}})}{\sigma_{\tilde{a}}} + \mu_{\tilde{c}} = \frac{6 \cdot 0.25(a-2)}{2} - 1 \tag{8.114}$$
$$= 0.75a - 2.5. \tag{8.115}$$

8.12 (Averaging noisy data)

a) Let $\tilde{w} := \sum_{i=1}^n \tilde{y}_i = n\tilde{x} + \sum_{i=1}^n \tilde{z}_i$. We need to find the linear MMSE estimator of $\tilde{x}$ given $\tilde{w}$. We have

$$\mathrm{Var}[\tilde{w}] = \mathrm{Var}\left[n\tilde{x} + \sum_{i=1}^n \tilde{z}_i\right] \tag{8.116}$$
$$= n^2 + n\sigma^2, \tag{8.117}$$
$$\mathbb{E}[\tilde{x}\tilde{w}] = \mathbb{E}[n\tilde{x}^2] + \sum_{i=1}^n \mathbb{E}[\tilde{x}\tilde{z}_i] \tag{8.118}$$
$$= n, \tag{8.119}$$
$$\rho_{\tilde{x},\tilde{w}} = \frac{\mathbb{E}[\tilde{x}\tilde{w}]}{\sigma_{\tilde{x}}\sigma_{\tilde{w}}} \tag{8.120}$$
$$= \frac{n}{\sqrt{n^2 + n\sigma^2}} \tag{8.121}$$
$$= \sqrt{\frac{n}{n+\sigma^2}}. \tag{8.122}$$

The linear MMSE estimate of $\tilde{x}$ given $\tilde{w} = w$ equals

$$\ell_{\mathrm{MMSE}}(w) = \frac{\rho_{\tilde{x},\tilde{w}}\sigma_{\tilde{x}}w}{\sigma_{\tilde{w}}} \tag{8.123}$$
$$= \frac{w}{n+\sigma^2}. \tag{8.124}$$

Therefore the optimal α equals $\frac{1}{n+\sigma^2}$.

b) As $\sigma^2 \to 0$, the estimator is just the average of the measurements. As $\sigma^2 \to \infty$ the estimator tends to zero: the measurements are ignored because they are just noise.

c) The MSE equals

$$\sigma_{\tilde{x}}^2(1 - \rho_{\tilde{x},\tilde{w}}^2) = 1 - \frac{n}{n+\sigma^2} \tag{8.125}$$
$$= \frac{\sigma^2}{n+\sigma^2}, \tag{8.126}$$

which tends to zero as $n \to \infty$.

8.13 (Interference)

a) Since $\tilde{w}$ and $\tilde{a}$ are independent

$$\mathbb{E}[\tilde{y}] = \mathbb{E}[\tilde{w}\tilde{a}] \tag{8.127}$$
$$= \mathbb{E}[\tilde{w}]\,\mathbb{E}[\tilde{a}] \tag{8.128}$$
$$= 0 \tag{8.129}$$

because $\mathbb{E}[\tilde{w}] = -0.5 + 0.5 = 0$. Similarly,

$$\mathbb{E}[\tilde{a}\tilde{y}] = \mathbb{E}\left[\tilde{w}\tilde{a}^2\right] \tag{8.130}$$
$$= \mathbb{E}[\tilde{w}]\,\mathbb{E}\left[\tilde{a}^2\right] \tag{8.131}$$
$$= 0. \tag{8.132}$$

Therefore,

$$\mathrm{Cov}[\tilde{a}, \tilde{y}] = \mathbb{E}[\tilde{a}\tilde{y}] - \mathbb{E}[\tilde{a}]\,\mathbb{E}[\tilde{y}] \tag{8.133}$$
$$= 0. \tag{8.134}$$

As a result, the linear MMSE estimator of $\tilde{a}$ given $\tilde{y} = y$ is

$$\ell_{\mathrm{MMSE}}(y) = \mathbb{E}[\tilde{a}] \tag{8.135}$$
$$= \mu. \tag{8.136}$$

b) The MSE equals

$$\mathrm{MSE} := \mathbb{E}\left[\left(\tilde{a} - \ell_{\mathrm{MMSE}}(\tilde{y})\right)^2\right] \tag{8.137}$$
$$= \mathbb{E}\left[(\tilde{a} - \mu)^2\right] \tag{8.138}$$
$$= \sigma^2. \tag{8.139}$$

c) Consider the nonlinear estimator $|\tilde{y}|$. Since $\tilde{a}$ is nonnegative, $|y| = \tilde{a}$, so the MSE is zero.

8.14 (Exam)

a) The MMSE estimator is the conditional mean, which we estimate using the sample conditional mean, depicted by the circle markers in the graph below. The advantage of this estimator is that it is very flexible, it indicates that grades increase a lot proportionally to the number of hours up to a certain point, but then there are diminishing returns (if any at all). The main disadvantage is that it is very noisy due to insufficient data, which results in overfitting. Also it does not provide an estimate at 4, 7 or 9.

b) The least-squares simple-linear-regression estimator that approximates the linear MMSE estimator is

$$\ell(h) = \sqrt{v_{\mathrm{grade}}}\rho_{\mathrm{grade,hours}}\left(\frac{h - m_{\mathrm{hours}}}{\sqrt{v_{\mathrm{hours}}}}\right) + m_{\mathrm{grade}} \tag{8.140}$$
$$= \frac{c_{\mathrm{grade,hours}}}{v_{\mathrm{hours}}}h + m_{\mathrm{grade}} - \frac{c_{\mathrm{grade,hours}}}{v_{\mathrm{hours}}}m_{\mathrm{hours}}, \tag{8.141}$$

where the sample means of the hours and grades equal

$$m_{\text{grade}} = \frac{85+90+30+80+90+80+60+40+90+85}{10} \tag{8.142}$$

$$= 73, \tag{8.143}$$

$$m_{\text{hours}} = \frac{5+6+2+6+5+5+3+3+6+8}{10} \tag{8.144}$$

$$= 4.9, \tag{8.145}$$

the sample variance of the hours is

$$\begin{aligned} v_{\text{hours}} &= \frac{(5-4.9)^2+(6-4.9)^2+(2-4.9)^2+(6-4.9)^2+(5-4.9)^2}{9} \\ &\quad + \frac{(5-4.9)^2+(3-4.9)^2+(3-4.9)^2+(6-4.9)^2+(9-4.9)^2}{9} \end{aligned} \tag{8.146}$$

$$= 4.01, \tag{8.147}$$

and the sample covariance is

$$\begin{aligned} &c_{\text{grade,hours}} \\ &= \frac{(85-73)(5-4.9)+(90-73)(6-4.9)+(30-73)(2-4.9)+(80-73)(6-4.9)}{9} \\ &\quad + \frac{(90-73)(5-4.9)+(80-73)(5-4.9)+(60-73)(3-4.9)+(40-73)(3-4.9)}{9} \\ &\quad + \frac{(90-73)(6-4.9)+(85-73)(8-4.9)}{9} \end{aligned} \tag{8.148}$$

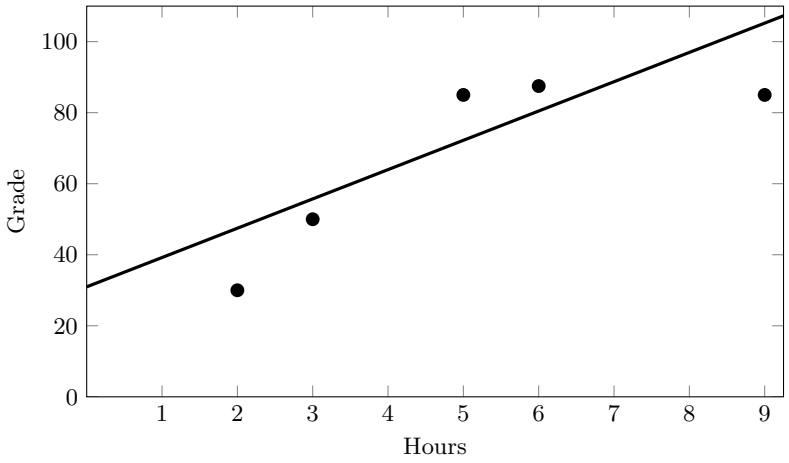
$$= 33.11. \tag{8.149}$$

The least-squares estimator equals

$$\ell(h) = \frac{33.1}{4.01}h + 73 - \frac{34.4}{4.01}4.9 \tag{8.150}$$

$$= 8.25h + 30.97, \tag{8.151}$$

The estimator is depicted as a straight line on the graph below. Its advantage is that it is less noisy than the nonlinear sample conditional mean estimator, and produces estimates for any number of hours. However it underfits the data, as it does not capture the fact that the grade increase eventually saturates when the number of hours exceed 5.



8.15 (Ice cream sales) The correlation coefficient equals

$$\rho = \frac{\text{Cov}\left[\text{adv}, \text{sales}\right]}{\sqrt{\text{Var}\left[\text{adv}\right]\text{Var}\left[\text{adv}\right]}} \tag{8.152}$$

$$= \frac{960}{\sqrt{900 \cdot 1600}} \tag{8.153}$$

$$= 0.8. \tag{8.154}$$

Sales and advertising are highly correlated, but this does not necessarily imply that advertising causes higher sales. From the table, it looks like temperature acts as a confounding factor. When it's warmer the company spends more on advertising, and people also buy more ice creams. Consequently, the argument of the marketing department is not very compelling.

8.16 (Nonconfounding variable)

a) By the theorem, the covariance between the observed outcome $\tilde{y}$ and the treatment $\tilde{t}$ equals $\beta_{\text{treat}} + \beta_{\text{conf}}\sigma_{\tilde{t},\tilde{c}}$. The mean of $\tilde{y}$ is zero by linearity of expectation, since $\tilde{t}$, $\tilde{c}$ and $\tilde{z}$ all have zero mean. Since $\tilde{z}$ is independent from $\tilde{t}$ and $\tilde{c}$, again by linearity of expectation, under the assumptions of the theorem, the variance of $\tilde{y}$ equals

$$\text{Var}\left[\tilde{y}\right] \tag{8.155}$$

$$= \mathbb{E}\left[\tilde{y}^2\right] \tag{8.156}$$

$$= \mathbb{E}\left[\left(\beta_{\text{treat}}\tilde{t} + \beta_{\text{conf}}\tilde{c} + \tilde{z}\right)^2\right] \tag{8.157}$$

$$= \mathbb{E}\left[\beta_{\text{treat}}^2\tilde{t}^2 + \beta_{\text{conf}}^2\tilde{c}^2 + \tilde{z}^2 + 2\left(\beta_{\text{treat}}\tilde{t} + \beta_{\text{conf}}\tilde{c}\right)\tilde{z} + 2\beta_{\text{treat}}\beta_{\text{conf}}\tilde{t}\tilde{c}\right] \tag{8.158}$$

$$= \beta_{\text{treat}}^2\mathbb{E}\left[\tilde{t}^2\right] + \beta_{\text{conf}}^2\mathbb{E}\left[\tilde{c}^2\right] + \mathbb{E}\left[\tilde{z}^2\right] + 2\mathbb{E}\left[\beta_{\text{treat}}\tilde{t} + \beta_{\text{conf}}\tilde{c}\right]\mathbb{E}\left[\tilde{z}\right] + 2\beta_{\text{treat}}\beta_{\text{conf}}\mathbb{E}\left[\tilde{t}\tilde{c}\right]$$

$$= \beta_{\text{treat}}^2 + \beta_{\text{conf}}^2 + \sigma_{\tilde{z}}^2 + 2\beta_{\text{treat}}\beta_{\text{conf}}\sigma_{\tilde{t},\tilde{c}}, \tag{8.159}$$

since $\tilde{t}$ and $\tilde{c}$ are standardized, so their mean square is one. Consequently, the correlation

coefficient $\rho_{\tilde{y},\tilde{t}}$ equals

$$\rho_{\tilde{y},\tilde{t}} := \frac{\sigma_{\tilde{t},\tilde{c}}}{\sqrt{\text{Var}\left[\tilde{y}\right]\text{Var}\left[\tilde{t}\right]}} \tag{8.160}$$

$$= \frac{\beta_{\text{treat}} + \beta_{\text{conf}}\sigma_{\tilde{t},\tilde{c}}}{\sqrt{\sigma_{\tilde{z}}^2 + \beta_{\text{treat}}^2 + \beta_{\text{conf}}^2 + 2\beta_{\text{treat}}\beta_{\text{conf}}\sigma_{\tilde{t},\tilde{c}}}}. \tag{8.161}$$

b) The limit of the correlation coefficient as $\sigma_{\tilde{z}} \to \infty$ equals zero. In that case, the nonconfounding variable dominates the observed outcome, rendering it uncorrelated with the treatment, because the treatment variable and the nonconfounding variable are uncorrelated.

c) In Example 8.38, $\beta_{\text{treat}} := 0$ and $\beta_{\text{conf}} := 1$, so setting $\sigma_{\tilde{z}}^2 := 1$,

$$\rho_{\tilde{y},\tilde{t}} = \frac{\sigma_{\tilde{t},\tilde{c}}}{\sqrt{2}}. \tag{8.162}$$

For $\sigma_{\tilde{t},\tilde{c}} := 0.8$, $\rho_{\tilde{y},\tilde{t}} = 0.57$. For $\sigma_{\tilde{t},\tilde{c}} := -0.8$, $\rho_{\tilde{y},\tilde{t}} = -0.57$. For $\sigma_{\tilde{t},\tilde{c}} := 0$, $\rho_{\tilde{y},\tilde{t}} = 0$.

9

Estimation Of Population Parameters

Exercises

9.1 (Markov's and Chebyshev's inequalities are tight)

a) Consider a random variable $\tilde{a}$ that is equal to c with probability θ, and equal to zero otherwise. Then $\mathrm{P}\left(\tilde{a} \geq c\right) = \theta$ and

$$\mathbb{E}\left[\tilde{a}\right] = cP\left(\tilde{a} = c\right) \tag{9.1}$$
$$= c\theta. \tag{9.2}$$

b) Consider a random variable $\tilde{b}$ with pmf $p_{\tilde{b}}$ such that

$$p_{\tilde{b}}(\mu - c) = \frac{\theta}{2}, \tag{9.3}$$
$$p_{\tilde{b}}(\mu) = 1 - \theta, \tag{9.4}$$
$$p_{\tilde{b}}(\mu + c) = \frac{\theta}{2}. \tag{9.5}$$

Then $\mathrm{P}\left(|\tilde{b} - \mu| \geq c\right) = \theta$,

$$\mathbb{E}[\tilde{b}] = \frac{\theta}{2}(\mu - c) + \frac{\theta}{2}(\mu + c) + (1 - \theta)\mu \tag{9.6}$$
$$= \mu, \tag{9.7}$$

and

$$\mathrm{Var}[\tilde{b}] = \mathbb{E}\left[(\tilde{b} - \mu)^2\right] \tag{9.8}$$
$$= c^2 \mathrm{P}\left(\tilde{a} = \mu + c\right) + c^2 \mathrm{P}\left(\tilde{a} = \mu - c\right) \tag{9.9}$$
$$= c^2\theta. \tag{9.10}$$

9.2 (Online poll)

a) The proportion of Democrat voters in the whole population is $\theta_{\mathrm{tot}} = \alpha\theta_{\mathrm{young}} + (1 - \alpha)\theta_{\mathrm{old}}$. The fraction of young voters in the poll that intend to vote Democrat y/n_1 is a reasonable estimate for θ_{young}. The fraction of old voters in the poll that intend to vote Democrat o/n_2 is a reasonable estimate for θ_{old}. The estimator for $\theta_{\mathrm{tot}} = \alpha\theta_{\mathrm{young}} + (1 - \alpha)\theta_{\mathrm{old}}$ therefore equals

$$e(o, y) := \frac{\alpha y}{n_1} + \frac{(1 - \alpha)o}{n_2}. \tag{9.11}$$

b) We have $y = 50$, $n_1 = 70$, $o = 10$, $n_2 = 30$, $\alpha = 0.25$, so

$$e(10, 50) = \frac{1}{4} \cdot \frac{50}{70} + \frac{3}{4} \cdot \frac{10}{30} = 42.9\%. \tag{9.12}$$

c) Let us assume that the young people in the population are sampled independently and uniformly at random with replacement, and so are the old people. Let $\tilde{y}$ denote the number of young voters in the poll that intend to vote Democrat, and let $\tilde{o}$ denote the number of old voters in the poll that intend to vote Democrat. By the notes, $\tilde{y}/n_1$ and $\tilde{o}/n_2$ are unbiased estimators of θ_{young} and θ_{old} respectively. Consequently, by linearity of expectation

$$\mathbb{E}\left[e(\tilde{o},\tilde{y})\right] := \mathbb{E}\left[\frac{\alpha\tilde{y}}{n_1} + \frac{(1-\alpha)\tilde{o}}{n_2}\right] \tag{9.13}$$

$$= \frac{\alpha\mathbb{E}\left[\tilde{y}\right]}{n_1} + \frac{(1-\alpha)\mathbb{E}\left[\tilde{o}\right]}{n_2} \tag{9.14}$$

$$= \alpha\theta_{\text{young}} + (1-\alpha)\theta_{\text{old}} = \theta_{\text{tot}}, \tag{9.15}$$

so the estimator is unbiased (under our assumptions).

d) Assuming independent and uniform random sampling with replacement,

$$\text{Var}\left[\frac{\tilde{y}}{n_1}\right] = \frac{\theta_{\text{young}}(1-\theta_{\text{young}})}{n_1}, \tag{9.16}$$

$$\text{Var}\left[\frac{\tilde{o}}{n_2}\right] = \frac{\theta_{\text{old}}(1-\theta_{\text{old}})}{n_2}. \tag{9.17}$$

Assuming $\tilde{y}$ and $\tilde{o}$ are independent, the variance of our estimator is

$$\text{Var}\left[e(\tilde{o},\tilde{y})\right] = \text{Var}\left[\frac{\alpha\tilde{y}}{n_1} + \frac{(1-\alpha)\tilde{o}}{n_2}\right] \tag{9.18}$$

$$= \alpha^2\text{Var}\left[\frac{\tilde{y}}{n_1}\right] + (1-\alpha)^2\text{Var}\left[\frac{\tilde{o}}{n_2}\right] \tag{9.19}$$

$$= \frac{\alpha^2\theta_{\text{young}}(1-\theta_{\text{young}})}{n_1} + \frac{(1-\alpha)^2\theta_{\text{old}}(1-\theta_{\text{old}})}{n_2}. \tag{9.20}$$

Since the estimator is unbiased, by Chebyshev's inequality, for any ϵ

$$\text{P}\left(|e(\tilde{o},\tilde{y}) - \theta_{\text{tot}}| > \epsilon\right) \leq \frac{\text{Var}\left[e(\tilde{o},\tilde{y})\right]}{\epsilon^2}, \tag{9.21}$$

which converges to zero as $n_1 \to \infty$ and $n_2 \to \infty$, so the estimator converges to θ_{tot} in probability and is therefore consistent.

9.3 (Participation)

a) A reasonable estimator for α_R is the fraction of people in the poll who are certain to vote and will vote Democrat. The estimator is unbiased if the people are selected independently with replacement, because the probability that each of them is in the category of interest is exactly α_R. Following the same reasoning as in Example 9.14, the standard error is $\sqrt{(\alpha_R(1-\alpha_R))/n} = \sqrt{(\alpha_R(1-\alpha_R))/150}$. For the observed data, the estimator equals $30/150 = 0.2$.

b) Let α_D denote the fraction of people who are certain to vote and will vote Democrat, β_R the fraction of people in the population who are likely to vote and will vote Republican, and β_D the fraction of people in the population who are likely to vote and will vote Democrat. According to the study, the fraction of people who will actually vote equals $0.8(\alpha_D + \alpha_R) + 0.1(\beta_D + \beta_R)$. Out of them, the ones who will vote Republican are $0.8\alpha_R + 0.1\beta_R$. As explained above, we can obtain unbiased estimators for α_R, α_D, β_D and β_R by computing the corresponding fractions from the poll, under the assumption

that the poll participants are selected independently with replacement. This yields the following estimate for the fraction of people who will vote Republican:

$$\frac{0.8\frac{30}{150} + 0.1\frac{10}{150}}{0.8\left(\frac{20}{150} + \frac{30}{150}\right) + 0.1\left(\frac{40}{150} + \frac{10}{150}\right)} = \frac{0.8 \cdot 3 + 0.1 \cdot 1}{0.8 \cdot 5 + 0.1 \cdot 5} \tag{9.22}$$

$$= 55.6\%. \tag{9.23}$$

9.4 (The sample covariance is unbiased)

a)

$$\sum_{i=1}^{n}(\tilde{x}_i - \tilde{m}_1)(\tilde{y}_i - \tilde{m}_2) = \sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i - \tilde{m}_1\sum_{i=1}^{n}\tilde{y}_i - \tilde{m}_2\sum_{i=1}^{n}\tilde{x}_i + n\tilde{m}_1\tilde{m}_2 \tag{9.24}$$

$$= \sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i - n\tilde{m}_1\tilde{m}_2. \tag{9.25}$$

b) For any $1 \leq i \leq n$

$$\mathrm{Cov}\left[\tilde{x}_i, \tilde{y}_i\right] := \mathbb{E}\left[(\tilde{x}_i - \mathbb{E}\left[\tilde{x}_i\right])(\tilde{y}_i - \mathbb{E}\left[\tilde{y}_i\right])\right] \tag{9.26}$$

$$= \mathbb{E}\left[(\tilde{x}_i - \mu_a)(\tilde{y}_i - \mu_b)\right] \tag{9.27}$$

$$= \sum_{k=1}^{N}(a_k - \mu_a)(b_k - \mu_b)p_{\tilde{k}_i}(k) \tag{9.28}$$

$$= \frac{1}{N}\sum_{k=1}^{N}(a_k - \mu_a)(b_k - \mu_b) \tag{9.29}$$

$$= c_{\mathrm{pop}}. \tag{9.30}$$

Since the means of $\tilde{m}_1$ and $\tilde{m}_2$ are μ_a and μ_b respectively, the result follows.

c) By linearity of the covariance operator

$$\mathbb{E}\left[\tilde{m}_1\tilde{m}_2\right] = \mathrm{Cov}\left[\tilde{m}_1, \tilde{m}_2\right] + \mu_a\mu_b \tag{9.31}$$

$$= \mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i, \frac{1}{n}\sum_{j=1}^{n}\tilde{y}_j\right] + \mu_a\mu_b \tag{9.32}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left[\tilde{x}_i, \tilde{y}_j\right] + \mu_a\mu_b \tag{9.33}$$

$$= \frac{c_{\mathrm{pop}}}{n} + \mu_a\mu_b. \tag{9.34}$$

d) Combining the above results

$$\mathbb{E}\left[\tilde{c}\right] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \tilde{m}_1)(\tilde{y}_i - \tilde{m}_2)\right] \tag{9.35}$$

$$= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i - n\tilde{m}_1\tilde{m}_2\right)\right] \tag{9.36}$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}\left[\tilde{x}_i\tilde{y}_i\right] - n\mathbb{E}\left[\tilde{m}_1\tilde{m}_2\right]\right) \tag{9.37}$$

$$= \frac{1}{n-1}\left(nc_{\text{pop}} + n\mu_a\mu_b - c_{\text{pop}} - n\mu_a\mu_b\right) \tag{9.38}$$

$$= c_{\text{pop}}. \tag{9.39}$$

9.5 (Herding)

a) Note that

$$\tilde{h}_3 = \frac{\tilde{m}_1}{4} + \frac{\frac{\tilde{m}_1+\tilde{m}_2}{2}}{4} + \frac{\tilde{m}_3}{2} \tag{9.40}$$

$$= \frac{3\tilde{m}_1}{8} + \frac{\tilde{m}_2}{8} + \frac{\tilde{m}_3}{2}. \tag{9.41}$$

Let θ_{pop} denote the population proportion, such that

$$\mathbb{E}\left[\tilde{m}_1\right] = \mathbb{E}\left[\tilde{m}_2\right] = \mathbb{E}\left[\tilde{m}_3\right] = \theta_{\text{pop}}. \tag{9.42}$$

By linearity of expectation and (9.41)

$$\mathbb{E}\left[\tilde{h}_2\right] = \frac{\mathbb{E}\left[\tilde{m}_1\right] + \mathbb{E}\left[\tilde{m}_2\right]}{2} \tag{9.43}$$

$$= \frac{\theta_{\text{pop}}}{2}, \tag{9.44}$$

$$\mathbb{E}\left[\tilde{h}_3\right] = \frac{3\mathbb{E}\left[\tilde{m}_1\right]}{8} + \frac{\mathbb{E}\left[\tilde{m}_2\right]}{8} + \frac{4\mathbb{E}\left[\tilde{m}_3\right]}{8} \tag{9.45}$$

$$= \theta_{\text{pop}}. \tag{9.46}$$

Both herded estimators are unbiased. In this situation, herding does not produce bias.

b) The polls are independent, so by (9.41)

$$\text{Var}\left[\tilde{h}_2\right] = \frac{\text{Var}\left[\tilde{m}_1\right] + \text{Var}\left[\tilde{m}_2\right]}{4} \tag{9.47}$$

$$= \frac{s^2}{2}, \tag{9.48}$$

$$\text{Var}\left[\tilde{h}_3\right] = \frac{9\text{Var}\left[\tilde{m}_1\right]}{64} + \frac{\text{Var}\left[\tilde{m}_2\right]}{64} + \frac{\text{Var}\left[\tilde{m}_3\right]}{4} \tag{9.49}$$

$$= \frac{13s^2}{32}. \tag{9.50}$$

Consequently the standard errors equal

$$\text{se}\left[\tilde{h}_2\right] = \frac{s}{\sqrt{2}} = 0.71s, \tag{9.51}$$

$$\text{se}\left[\tilde{h}_3\right] = \sqrt{\frac{13}{32}}s = 0.64s. \tag{9.52}$$

Herding improves each individual estimator.

c) Note that by (9.41)

$$\tilde{h}_{\text{all}} = \frac{\tilde{m}_1}{3} + \frac{\frac{\tilde{m}_1+\tilde{m}_2}{2}}{3} + \frac{\tilde{h}_3}{3} \tag{9.53}$$

$$= \left(\frac{1}{3} + \frac{1}{6} + \frac{3}{24}\right)\tilde{m}_1 + \left(\frac{1}{6} + \frac{1}{24}\right)\tilde{m}_2 + \frac{\tilde{m}_3}{6} \tag{9.54}$$

$$= \frac{15\tilde{m}_3}{24} + \frac{5\tilde{m}_3}{24} + \frac{\tilde{m}_3}{6}. \tag{9.55}$$

By the independence assumption,

$$\text{Var}\left[\tilde{m}_{\text{all}}\right] = \frac{\text{Var}\left[\tilde{m}_1\right] + \text{Var}\left[\tilde{m}_2\right] + \text{Var}\left[\tilde{m}_2\right]}{9} \tag{9.56}$$

$$= \frac{s^2}{3}, \tag{9.57}$$

$$\text{Var}\left[\tilde{h}_{\text{all}}\right] = \frac{13^2\text{Var}\left[\tilde{m}_1\right]}{24^2} + \frac{5^2\text{Var}\left[\tilde{m}_2\right]}{24^2} + \frac{\text{Var}\left[\tilde{m}_3\right]}{36} \tag{9.58}$$

$$= \frac{133s^2}{288}. \tag{9.59}$$

Consequently the standard errors equal

$$\text{se}\left[\tilde{m}_{\text{all}}\right] = \frac{s}{\sqrt{3}} = 0.58s, \tag{9.60}$$

$$\text{se}\left[\tilde{h}_{\text{all}}\right] = \sqrt{\frac{133}{288}}s = 0.68s. \tag{9.61}$$

The herded estimator $\tilde{h}_{\text{all}}$ has a larger standard error than the non-herded estimator $\tilde{m}_{\text{all}}$, so herding is problematic because it degrades the aggregated estimator. This happens because the first poll (and to a lesser extent the second poll) are overrepresented in the herded estimators.

9.6 (Multiplicative noise)

a) The estimator is

$$\tilde{h} := \frac{\sum_{i=1}^n x_i}{\mu}. \tag{9.62}$$

As long as μ is nonzero, it is unbiased, since by linearity of expectation

$$\mathbb{E}\left[\tilde{h}\right] = \mathbb{E}\left[\frac{\frac{1}{n}\sum_{i=1}^n \tilde{x}_i}{\mu}\right] \tag{9.63}$$

$$= \frac{\gamma}{\mu}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\tilde{a}_i\right] \tag{9.64}$$

$$= \frac{n\mu}{n\mu}\gamma = \gamma. \tag{9.65}$$

b) The variance of the estimator equals

$$\text{Var}\left[\tilde{h}\right] = \text{Var}\left[\frac{\frac{1}{n}\sum_{i=1}^n \tilde{x}_i}{\mu}\right] \tag{9.66}$$

$$= \frac{\gamma^2}{\mu^2}\frac{1}{n^2}\sum_{i=1}^n \text{Var}\left[\tilde{a}_i\right] \tag{9.67}$$

$$= \frac{\gamma^2\sigma^2}{\mu^2 n}, \tag{9.68}$$

so the standard error equals

$$\text{se}\left[\tilde{h}\right] = \frac{\gamma\sigma}{\mu\sqrt{n}}. \tag{9.69}$$

c) Since the estimator is unbiased, by Chebyshev's inequality and the previous question,

$$\text{P}\left(\left|\tilde{h} - \gamma\right| > \epsilon\right) \leq \frac{\text{Var}\left[\tilde{h}\right]}{\epsilon^2} \tag{9.70}$$

$$\leq \frac{\gamma^2\sigma^2}{\mu^2 n\epsilon}, \tag{9.71}$$

which converges to zero as $n \to \infty$, so the estimator is consistent.

9.7 (Consistency of sample median)

a) If n is odd, then for the median to be greater than $\gamma + \epsilon$ at least $(n+1)/2 \geq n/2$ elements have to be greater than $\gamma + \epsilon$. If n is even, then $n/2$ have to be greater than $\gamma + \epsilon$.

b) By the assumption that γ is the only point that satisfies $F_{\tilde{x}_i}(\gamma) = 1/2$, there exists a constant $\epsilon' > 0$ such that for any i the probability that $\tilde{x}_i > \gamma + \epsilon$ is

$$\text{P}\left(\tilde{x}_i > \gamma + \epsilon\right) = 1/2 - \epsilon' := \theta. \tag{9.72}$$

The random variable $\tilde{b}$ is binomial with parameters n and θ. As a result, we have

$$\text{P}\left(\tilde{b} \geq \frac{n}{2}\right) = \text{P}\left(\tilde{b} - n\theta \geq \frac{n}{2} - n\theta\right) \tag{9.73}$$

$$\leq \text{P}\left(\left|\tilde{b} - n\theta\right| \geq n\epsilon'\right) \tag{9.74}$$

$$\leq \frac{\text{Var}[\tilde{b}]}{\left(n\epsilon'\right)^2} \quad \text{by Chebyshev's inequality} \tag{9.75}$$

$$= \frac{n\theta\left(1-\theta\right)}{n^2\left(\epsilon'\right)^2} \tag{9.76}$$

$$= \frac{\theta\left(1-\theta\right)}{n\left(\epsilon'\right)^2}, \tag{9.77}$$

which converges to zero as $n \to \infty$. This completes the proof, combined with the first part.

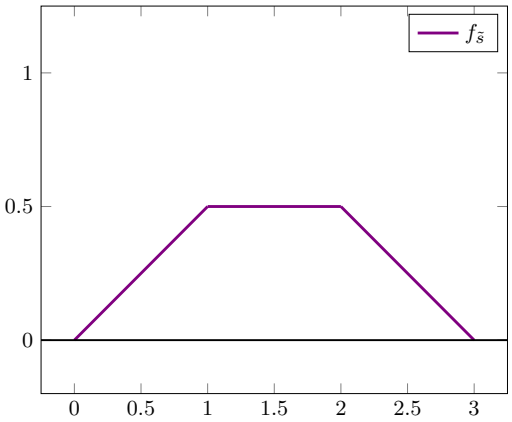
9.8 (Watering a plant) We represent the water poured by the roommates using the random variables $\tilde{a}$ and $\tilde{b}$. We are interested in the pdf of the sum $\tilde{s} := \tilde{a} + \tilde{b}$. It equals

$$f_{\tilde{s}}\left(s\right) = \int_{u=-\infty}^{\infty} f_{\tilde{a}}\left(s-u\right) f_{\tilde{b}}\left(u\right)\, \mathrm{d}u \tag{9.78}$$

$$= \frac{1}{2}\int_{u=0}^{2} f_{\tilde{a}}\left(s-u\right)\, \mathrm{d}u \tag{9.79}$$

$$= \begin{cases} \frac{1}{2}\int_{u=0}^{s} \mathrm{d}u = \frac{s}{2} & \text{if } s \leq 1, \\ \frac{1}{2}\int_{u=s-1}^{s} \mathrm{d}u = \frac{1}{2} & \text{if } 1 \leq s \leq 2, \\ \frac{1}{2}\int_{u=s-1}^{2} \mathrm{d}u = \frac{3-s}{2} & \text{if } 2 \leq s \leq 3. \end{cases} \tag{9.80}$$

The pdf of $\tilde{s}$ is shown below.



9.9 (Confidence intervals)

a) 72 out of 80 intervals contain μ. A reasonable estimate for $1-\alpha$ is therefore 0.9, so the estimate for α is 0.1.

b)
- $X = B$ and $Y = A$. The central limit theorem produces much narrower confidence intervals.
- $X = B$ and $Y = A$. More data results in narrower confidence intervals (see equation in the notes).
- $X = A$ and $Y = B$. Higher variance results in wider confidence intervals (see equation in the notes).

9.10 (Cholesterol)

a) Let $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{200}$ be samples from the men, and $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{100}$ the sample from the women. If each individual is selected independently with replacement, then $\text{Var}\left[\tilde{x}_i\right] = \sigma^2_{\text{men}}$ and $\text{Var}\left[\tilde{y}_i\right] = \sigma^2_{\text{women}}$. As a result, by independence

$$\text{Var}\left[\frac{1}{200}\sum_{i=1}^{200}\tilde{x}_i - \frac{1}{100}\sum_{i=1}^{100}\tilde{y}_i\right] = \text{Var}\left[\frac{1}{200}\sum_{i=1}^{200}\tilde{x}_i\right] + \text{Var}\left[-\frac{1}{100}\sum_{i=1}^{100}\tilde{y}_i\right] \tag{9.81}$$

$$= \text{Var}\left[\frac{1}{200}\sum_{i=1}^{200}\tilde{x}_i\right] + \text{Var}\left[\frac{1}{100}\sum_{i=1}^{100}\tilde{y}_i\right] \tag{9.82}$$

$$= \frac{\sigma^2_{\text{men}}}{200} + \frac{\sigma^2_{\text{women}}}{100}. \tag{9.83}$$

Therefore the standard error is $\sqrt{\frac{\sigma^2_{\text{men}}}{200} + \frac{\sigma^2_{\text{women}}}{100}}$.

b) By the central limit theorem, if μ_{men} and μ_{women} are the population means for men and women respectively, then $\frac{1}{200}\sum_{i=1}^{200}\tilde{x}_i$ is approximately Gaussian with mean μ_{men} and variance $\frac{\sigma^2_{\text{men}}}{200}$, and $\frac{1}{100}\sum_{i=1}^{100}\tilde{y}_i$ is approximately Gaussian with mean μ_{women} and variance $\frac{\sigma^2_{\text{women}}}{100}$. As a result, $\frac{\sigma^2_{\text{men}}}{200} - \frac{1}{100}\sum_{i=1}^{100}\tilde{y}_i$ is approximately Gaussian with mean $\mu_{\text{men}} - \mu_{\text{women}}$ and variance $\frac{\sigma^2_{\text{men}}}{200} + \frac{\sigma^2_{\text{women}}}{100} = 1.5$. By Lemma 9.42 the following is a 0.95 confidence interval for $\mu_{\text{men}} - \mu_{\text{women}}$:

$$\left[20 - 1.96\sqrt{1.5}, 20 + 1.96\sqrt{1.5}\right] = [17.6, 22.4]. \tag{9.84}$$

9.11 (Difference of proportions)

a) The variance of the first proportion equals $\theta_A(1-\theta_A)/n_A$ because it is a binomial with parameters n_A and θ_A multiplied by $1/n_A$. By the same reasoning the variance of the second proportion is $\theta_B(1-\theta_B)/n_B$. Assuming independence between the two sets, the variance of the difference is equal to the sum of the variances (the difference is equal to the sum of the proportions after multiplying one by -1, which does not modify the variance). We conclude that the standard error equals

$$\text{se}_{\text{diff}} = \sqrt{\frac{\theta_A(1-\theta_A)}{n_A} + \frac{\theta_B(1-\theta_B)}{n_B}}. \tag{9.85}$$

b) By the Gaussian approximation to the binomial distribution and the properties of Gaussian random variables, the difference between the proportions is approximately Gaussian with mean $\theta_A - \theta_B$ and variance equal to the standard error se_{diff}. To compute the corresponding confidence interval we apply Lemma 9.42, setting $\theta_A \approx k_A/n_A$ and $\theta_B A \approx k_B/n_B$ to obtain an approximation to the standard error that we denote by $\text{se}_{\text{approx}}$. The confidence interval equals

$$\left[\frac{k_A}{n_A} - \frac{k_B}{n_B} - c_\alpha \,\text{se}_{\text{approx}}, \frac{k_A}{n_A} - \frac{k_B}{n_B} + c_\alpha \,\text{se}_{\text{approx}}\right], \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{9.86}$$

where $F_{\tilde{z}}$ denotes the cdf of a standard Gaussian with zero mean and unit variance.

c) For the first trial, setting $k_A := 52$, $n_A := 100$, $k_B := 30$ and $n_B := 100$ we obtain a standard error $\text{se}_{\text{approx}}$ equal to 0.0678 and a 0.95 confidence interval equal to

$$[0.52 - 0.3 - 1.96 \cdot 0.0678, 0.52 - 0.3 + 1.96 \cdot 0.0678] = [0.0871, 0.353]. \tag{9.87}$$

For the second trial, setting $k_A := 30650$, $n_A := 100000$, $k_B := 30000$ and $n_B := 100000$ we obtain a standard error equal to 0.00206 and a 0.95 confidence interval equal to

$$[0.3065 - 0.3 - 1.96 \cdot 0.00206, 0.3065 - 0.3 + 1.96 \cdot 0.00206] = [0.0025, 0.0105]. \tag{9.88}$$

d) We can approximate the standard error via bootstrapping following Definition 9.50 as follows:

- Generate T batches of n bootstrap samples from Dataset A, $a_j^{[t]}$, $1 \leq j \leq n_A$, $1 \leq t \leq T$, following Definition 9.49, where T is a very large integer. Let $k_A^{[t]}$ denote the number of ones in batch t.
- Generate T batches of n bootstrap samples from Dataset B, $b_j^{[t]}$, $1 \leq j \leq n_B$, $1 \leq t \leq T$, following Definition 9.49. Let $k_A^{[t]}$ denote the number of ones in batch t.
- Compute the bootstrap difference in proportions for the batches:

$$d_t := \frac{k_A^{[t]}}{n_A} - \frac{k_B^{[t]}}{n_B}. \tag{9.89}$$

- Compute the sample standard deviation of the bootstrap differences $d_1, \ldots, d_T$ to obtain the bootstrap standard error se_{bs}.

e) Option 1: We plug the bootstrap standard error se_{bs} into Lemma 9.42 to obtain the $1-\alpha$ bootstrap Gaussian confidence interval,

$$\left[\frac{k_A}{n_A} - \frac{k_B}{n_B} - c_\alpha \,\text{se}_{\text{bs}}, \frac{k_A}{n_A} - \frac{k_B}{n_B} + c_\alpha \,\text{se}_{\text{approx}}\right], \qquad c_\alpha := F_{\tilde{z}}^{-1}\left(1 - \frac{\alpha}{2}\right). \tag{9.90}$$

Option 2: We compute the $\alpha/2$ and $1-\alpha/2$ quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of the bootstrap differences $d_1, \ldots, d_T$. The $1-\alpha$ bootstrap percentile confidence interval is $[q_{\alpha/2}, q_{1-\alpha/2}]$.

By consistency of the sample variance, the bootstrap standard error of the difference is approximately equal to the true standard error of the difference. Similarly, our approximation to the standard error is a consistent estimator of the true standard error (again because of consistency of the sample variance), so the bootstrap Gaussian confidence interval will be very similar to the confidence interval above. The bootstrap percentile confidence interval will also be similar, because the standard errors are approximately the same and the difference is approximately Gaussian as explained above, so we can apply the same argument as in the case of the sample mean (see beginning of Section 9.9.2).

9.12 (Exact confidence interval)

a) Since the variance of $\widetilde{m}$ equals σ^2/n,

$$P\left(\mu \in \left[\widetilde{m} - \frac{b}{\sqrt{\alpha\, n}}, \widetilde{m} + \frac{b}{\sqrt{\alpha\, n}}\right]\right) = 1 - P\left(|\widetilde{m} - \mu| > \frac{b}{\sqrt{\alpha\, n}}\right) \tag{9.91}$$

$$\geq 1 - \frac{\alpha\, n \mathrm{Var}\left[\widetilde{m}\right]}{b^2} \quad \text{by Chebyshev's inequality}$$

$$= 1 - \frac{\alpha\, \sigma^2}{b^2} \tag{9.92}$$

$$\geq 1 - \alpha. \tag{9.93}$$

b) We can set $b = 1$ because the parameter of interest is a ratio bounded by one. If $\alpha = 0.05$, setting the half-width equal to 0.01, yields

$$\frac{1}{\sqrt{0.05\, n}} = 0.01 \tag{9.94}$$

so n must be greater than 200,000, which is 20 times more tests than required by the approximate confidence interval.

9.13 (Modified Gaussian bootstrap confidence interval) Let τ denote the monotonic transformation and $h(X)$ the observed value of the estimator of interest. Since the transformed parameter is approximately Gaussian, we use bootstrap samples to estimate the standard error $\mathrm{se}_{\mathrm{bs}}$ of the transformed estimator (for each bootstrap batch we compute the estimator and apply the transformation, then we compute the sample standard deviation of the transformed values). We can use the bootstrap standard error to construct a confidence interval for the transformed parameter:

$$[\ell, u] = \left[\tau\left(h(X)\right) - c_\alpha\, \mathrm{se}_{\mathrm{bs}}, \tau\left(h(X)\right) + c_\alpha\, \mathrm{se}_{\mathrm{bs}}\right], \tag{9.95}$$

where c_α is set to obtain a 1-α confidence interval under the Gaussian assumption. Now, if the transformed parameter of interest $\tau(\tilde{w})$ is in $[\ell, u]$ with probability $1 - \alpha$, then the parameter of interest $\tilde{w}$ is in the interval $[\tau^{-1}(\ell), \tau^{-1}(u)]$ also with probability $1 - \alpha$, because τ^{-1} is a monotonic function (the inverse of a monotonic function is monotonic). Consequently, the interval

$$\mathcal{I}_{1-\alpha}^{[\tau]} := \left[\tau^{-1}\left(\tau\left(h(X)\right) - c_\alpha\, \mathrm{se}_{\mathrm{bs}}\right), \tau^{-1}\left(\tau\left(h(X)\right) + c_\alpha\, \mathrm{se}_{\mathrm{bs}}\right)\right] \tag{9.96}$$

is a 1-α confidence interval for the parameter of interest.

10

Hypothesis Testing

Exercises

10.1 (Test choice)

a) Test 2. We want to control false positives at all costs, in order to avoid declaring the water safe when it is not and get the kids poisoned. The price to pay is low power, i.e. failure to declare water safe when it actually is.

b) Test 1. Even if this means high significance level (so high probability of false positives), we want as much power as possible to not miss the gold!

10.2 (Scout)

a) The scout can apply a hypothesis test, to ensure that the workout is not consistent with the player having a 3-point percentage smaller or equal to 0.4. The parameter of interest θ is the 3-point percentage of the player. The null hypothesis is that $\theta \leq 0.4$. The test statistic is the number of 3-pointers the player makes before missing. If the p value is below the desired significance level, she will recruit the player.

b) We assume that the probability that the player makes each 3-point shot is θ and that the shots are independent. Let $\tilde{t}_\theta$ be a random variable that represents the number of made 3-pointers if the 3-point percentage is θ. The probability that the player makes t 3-point shots or more is the probability that they make the first t shots, so $\mathrm{P}\left(\tilde{t}_\theta \geq t\right) = \theta^t$. Therefore, under our assumptions, the p-value function equals

$$\begin{aligned} \mathrm{pv}(t) &:= \sup_{\theta \in [0,0.4]} \mathrm{P}\left(\tilde{t}_\theta \geq t\right) && (10.1) \\ &= \sup_{\theta \in [0,0.4]} \theta^t && (10.2) \\ &= 0.4^t. && (10.3) \end{aligned}$$

We need

$$\mathrm{pv}(t) = 0.4^t \leq 0.05, \tag{10.4}$$

so $t \geq \log 0.05 / \log 0.4 = 3.26$. The player needs to make at least 4 3-pointers in a row to be recruited.

c) Since the player is recruited if they make at least 4 3-pointers in a row, the probability that they are not recruited is $1 - 0.5^4 = 0.9375$. The scout is being very conservative and will definitely miss some good shooters.

10.3 (Road renovation)

a) The possible values of the test statistic are the integers between 0 and 4. Since the probability of an accident occurring in the *dangerous* section is 1/4 and the accidents

are independent, the pmf of the test statistic under the null hypothesis t_{null} is binomial with parameters $n := 4$ and $\theta := 1/4$:

$$p_{\tilde{t}_{\text{null}}}(0) = \left(1 - \frac{1}{4}\right)^4 = \frac{81}{256}, \tag{10.5}$$

$$p_{\tilde{t}_{\text{null}}}(1) = 4\left(\frac{1}{4}\right)^3\left(1 - \frac{1}{4}\right) = \frac{108}{256}, \tag{10.6}$$

$$p_{\tilde{t}_{\text{null}}}(2) = 1 - \left(\frac{1}{4}\right)^4 = \frac{54}{256}, \tag{10.7}$$

$$p_{\tilde{t}_{\text{null}}}(3) = 4\left(\frac{1}{4}\right)^3\left(1 - \frac{1}{4}\right) = \frac{12}{256}, \tag{10.8}$$

$$p_{\tilde{t}_{\text{null}}}(4) = \left(\frac{1}{4}\right)^4 = \frac{1}{256}. \tag{10.9}$$

The p value function is

$$\text{pv}(0) := \text{P}\left(\tilde{t}_{\text{null}} \geq 0\right) = 1, \tag{10.10}$$

$$\text{pv}(1) := \text{P}\left(\tilde{t}_{\text{null}} \geq 1\right) = \frac{175}{256}, \tag{10.11}$$

$$\text{pv}(2) := \text{P}\left(\tilde{t}_{\text{null}} \geq 2\right) = \frac{67}{256}, \tag{10.12}$$

$$\text{pv}(3) := \text{P}\left(\tilde{t}_{\text{null}} \geq 3\right) = \frac{13}{256}, \tag{10.13}$$

$$\text{pv}(4) := \text{P}\left(\tilde{t}_{\text{null}} \geq 4\right) = \frac{1}{256}. \tag{10.14}$$

b) The null hypothesis is rejected if $\text{pv}(t) \leq \alpha := 0.05$, which only holds for $t = 4$ ($\text{pv}(3) = 0.0508$). The probability of a false positive is therefore the probability of $\tilde{t}_{\text{null}} = 4$, which equals $1/256 = 3.9 \cdot 10^{-3}$.

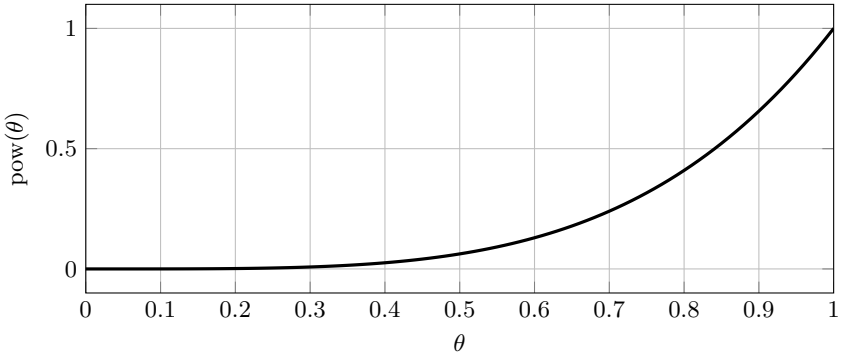
c) The null hypothesis is rejected if the test statistic equals 4. Under the assumption that accidents occur independently, the test statistic $\tilde{t}_\theta$ is binomial with parameters $n := 4$ and θ. The power function consequently is

$$\text{pow}(\theta) := \text{P}\left(\text{pv}\left(\tilde{t}_\theta\right) \leq \alpha\right) \tag{10.15}$$

$$= \text{P}\left(\tilde{t}_{\text{null}} = 4\right) \tag{10.16}$$

$$= \theta^4. \tag{10.17}$$

Here is the graph:



d) We need $\theta^4 \geq 0.5$, so $\theta \geq 0.5^{1/4} = 0.84$.

10.4 (Computer component) We denote the exponential random variables representing the data by $\tilde{x}_1, \ldots, \tilde{x}_n$. The null hypothesis is that the mean of the exponential $1/\lambda$ is smaller than one, or equivalently that $\lambda \geq 1$. For any λ,

$$\begin{aligned} \mathrm{P}\left(\tilde{t}_\lambda \geq t\right) &= \mathrm{P}\left(\min_{1\leq i\leq n} \tilde{x}_i \geq t\right) && (10.18)\\ &= \mathrm{P}\left(\cap_{1\leq i\leq n} \left\{\tilde{x}_i \geq t\right\}\right) && (10.19)\\ &= \prod_{i=1}^{n} \mathrm{P}\left(\tilde{x}_i \geq t\right) && (10.20)\\ &= \prod_{i=1}^{n} \exp\left(-\lambda t\right) && (10.21)\\ &= \exp\left(-\lambda n t\right). && (10.22) \end{aligned}$$

Consequently, the p value function is

$$\begin{aligned} \mathrm{pv}(t) &:= \sup_{0\leq\lambda\leq 1} \mathrm{P}\left(\tilde{t}_\lambda \geq t\right) && (10.23)\\ &= \sup_{0\leq\lambda\leq 1} \exp\left(-\lambda n t\right) && (10.24)\\ &= \exp\left(-nt\right). && (10.25) \end{aligned}$$

The null hypothesis is rejected when $\mathrm{pv}(t) = \exp\left(-nt\right) \leq \alpha$, or equivalently

$$t \geq \frac{1}{n} \log\left(\frac{1}{\alpha}\right). \tag{10.26}$$

By (10.22), the power function then equals

$$\begin{aligned} \mathrm{pow}(\lambda) &:= \mathrm{P}\left(\mathrm{pv}\left(\tilde{t}_\lambda\right) \leq \alpha\right) && (10.27)\\ &= \mathrm{P}\left(\tilde{t}_\lambda \geq \frac{1}{n} \log\left(\frac{1}{\alpha}\right)\right) && (10.28)\\ &= \exp\left(-\lambda n \cdot \frac{1}{n} \log\left(\frac{1}{\alpha}\right)\right) && (10.29)\\ &= \alpha^\lambda. && (10.30) \end{aligned}$$

The power at $\lambda = 1$ is α. As $\lambda \to 0$ it approaches one.

10.5 (Test statistic in two-sample test) Recall that $\tilde{h}_i$ and $\tilde{a}_i$ are independent Bernoulli random variables representing the ith free throw attempted at home and away respectively, for $1 \leq i \leq n$, so their parameters equal θ_{home} and θ_{away}, respectively. Consequently, $\sum_{i=1}^{n_{\mathrm{home}}} \tilde{h}_i$ is binomial with parameters n_{home} and θ_{home}, as explained in the proof of Lemma 7.19. By the Gaussian approximation to the binomial distribution (Definition 9.39) and Theorem 3.31, this implies that $\frac{1}{n_{\mathrm{home}}} \sum_{i=1}^{n_{\mathrm{home}}} \tilde{h}_i$ is approximately Gaussian with mean θ_{home} and variance $\frac{\theta_{\mathrm{home}}(1-\theta_{\mathrm{home}})}{n_{\mathrm{home}}}$. By the same argument, $-\frac{1}{n_{\mathrm{away}}} \sum_{i=1}^{n_{\mathrm{away}}} \tilde{a}_i$ is approximately Gaussian with mean θ_{away} and variance $\frac{\theta_{\mathrm{away}}(1-\theta_{\mathrm{away}})}{n_{\mathrm{away}}}$. By Theorem 9.36,

$$\tilde{t}_\theta := \frac{1}{n_{\mathrm{home}}} \sum_{i=1}^{n_{\mathrm{home}}} \tilde{h}_i - \frac{1}{n_{\mathrm{away}}} \sum_{i=1}^{n_{\mathrm{away}}} \tilde{a}_i, \tag{10.31}$$

is therefore approximately Gaussian with mean $\theta_{\mathrm{home}} - \theta_{\mathrm{away}}$ and variance

$$\frac{\theta_{\mathrm{home}}(1-\theta_{\mathrm{home}})}{n_{\mathrm{home}}} + \frac{\theta_{\mathrm{away}}(1-\theta_{\mathrm{away}})}{n_{\mathrm{away}}}. \tag{10.32}$$

10.6 (Tom Brady and hurricanes)

a) In the $n_W := 7$ years that Brady won, there were $k_W := 4$ hurricanes. In the $n_L := 13$ years that Brady didn't win, there were also $k_L := 4$ hurricanes. The test statistic of the one-tailed test is

$$t_{\text{data}} := \frac{k_W}{n_W} - \frac{k_L}{n_L} \tag{10.33}$$

$$= \frac{4}{7} - \frac{4}{13} \tag{10.34}$$

$$= 0.264. \tag{10.35}$$

Let $k := k_W + k_L = 8$ and $n := n_W + n_L = 20$. The variance under the null hypothesis is

$$\sigma^2_{\text{null}} := \frac{k(n-k)}{n^2}\left(\frac{1}{n_W} + \frac{1}{n_L}\right) \tag{10.36}$$

$$= \frac{8 \cdot 12}{20^2}\left(\frac{1}{7} + \frac{1}{13}\right) \tag{10.37}$$

$$= 0.0257. \tag{10.38}$$

The p value equals

$$\text{pv}(t_{\text{data}}) = 1 - F_{\tilde{z}}\left(\frac{t_{\text{data}}}{\sigma_{\text{null}}}\right) \tag{10.39}$$

$$= 0.125. \tag{10.40}$$

b) No, because the null hypothesis and test statistic were not decided before looking at the data.

10.7 (Unemployment in Spain)

a) Our null hypothesis is that the population correlation coefficient ρ is zero. Let $\tilde{r}$ be a random variable that represents the sample correlation coefficient computed from n samples under the null hypothesis. After applying Fisher's transformation, which we denote by M, $M(\tilde{r})$ is approximately Gaussian with mean $M(\rho) = M(0) = 0$ and standard deviation $1/\sqrt{n-3}$.

We select the absolute value of the observed sample correlation coefficient $|r|$ as the test statistic. Let $\tilde{z}$ be a standard Gaussian random variable with mean 0 and unit variance, then by monotonicity of M the p-value function is

$$\text{P}\left(|\tilde{r}| \geq |r|\right) = \text{P}\left(|M(\tilde{r})| \geq |M(r)|\right) \tag{10.41}$$

$$\approx \text{P}\left(\frac{|\tilde{z}|}{\sqrt{n-3}} \geq |M(r)|\right) \tag{10.42}$$

$$= 2\left(1 - F_{\tilde{z}}\left(\sqrt{n-3}\,|M(r)|\right)\right) \tag{10.43}$$

$$= 2\left(1 - F_{\tilde{z}}\left(\left|\frac{\sqrt{n-3}}{2}\log\left(\frac{1+t}{1-t}\right)\right|\right)\right). \tag{10.44}$$

b) Setting $t := -0.21$ and $n := 92$, the p value equals

$$2\left(1 - F_{\tilde{z}}\left(\left|\frac{\sqrt{89}}{2}\log\left(\frac{0.79}{1.21}\right)\right|\right)\right) = 2\left(1 - F_{\tilde{z}}\left(2.01\right)\right) \tag{10.45}$$

$$= 0.044. \tag{10.46}$$

The correlation between unemployment and temperature in Spain is statistically significant at a significance level of 0.05.

c) The data are not sampled i.i.d. from a population, they just include all the months over a certain period (with some exclusions due to the COVID-19 pandemic). The hypothesis test indicates that if they were i.i.d. samples, then it would be very unlikely for them to be uncorrelated. This is convincing evidence that the correlation between temperature and unemployment is not due to small sample size, but we can definitely *not* conclude that higher temperatures cause a decrease in unemployment, as explained in detail in Section 8.8.

10.8 (Median cholesterol) The test statistic is the median cholesterol of the ill people minus the median cholesterol of the healthy people. The value of the test statistic for each permutation is:
HIHIIH: 250-240=10
HIIIHH: 260-240 = 20
IIHHHI: 250-260=-10
IHHIHI: 300-250=50
HHIIIH: 260-250=10.
The actual test statistic equals 300-250=50. The p value is the fraction of the time the test statistic is larger or equal to this for the different permutations. It equals 1/5.

10.9 (False positives of a permutation test)

a) By Theorem 10.22, the p-value function equals

$$\text{pv}(t) = \frac{\sum_{v \in \Pi_{x_{\text{data}}}} 1\left(T(v) \geq t\right)}{|\Pi_{x_{\text{data}}}|}. \tag{10.47}$$

In our case, there are 24 permutations. For each possible value of the test statistic t (see Figure 10.7), we compute what fraction of permutations correspond to a test statistic greater or equal to t. The resulting values of the p value function are listed below, to the right of the corresponding value of t:

t	-4.5	-1.5	-0.5	0.5	1.5	4.5
$\text{pv}(t)$	1	$\frac{5}{6}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

b) Under the null hypothesis that the entries of $\tilde{x}_{\text{null}}$ are exchangeable, every permutation is equally likely: each occurs with probability 1/24. Since each value of the test statistic corresponds to 4 permutations (see Figure 10.7), the conditional probability that it occurs is $4/24 = 1/6$. Consequently, that is also the conditional probability of each of the possible values of $\tilde{u}$, because for any t and u such that $\text{pv}(t) = u$,

$$\text{P}\left(\tilde{u} = u \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right) = \text{P}\left(\tilde{t}_{\text{null}} = t \,|\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}\right). \tag{10.48}$$

To compute the conditional cdf of $\tilde{u}$ given $\tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}}$, we sum the corresponding probabilities, obtaining:

u	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1
$F_{\tilde{u}}(u \,\vert\, \tilde{x}_{\text{null}} \in \Pi_{x_{\text{data}}})$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1

Notice that this is consistent with Theorem 10.27.

c) A false positive happens if $\tilde{u} \leq 0.2$, which can only happen if $\tilde{u} = \frac{1}{6}$. The conditional probability of a false positive is therefore equal to $\frac{1}{6}$, which is smaller or equal to the significance level in accordance with Theorem 10.27.

d) Yes, as established in Corollary 10.28. Intuitively, the marginal probability of false positives is controlled by the significance level α, because the conditional probability is bounded by α for every possible value of the data.

10.10 (Sign test)

a) Since $\tilde{a}$ and $\tilde{b}$ are continuous, $\mathrm{P}\left(\tilde{a} = \tilde{b}\right) = 0$, so any outcome in the probability space associated to the random variables must belong to one of the two disjoint events $\tilde{a} > \tilde{b}$ and $\tilde{b} > \tilde{a}$. Consequently the sum of their probabilities must equal one

$$1 = \mathrm{P}\left(\{\tilde{a} > \tilde{b}\} \cup \{\tilde{b} > \tilde{a}\}\right) = \mathrm{P}(\tilde{a} > \tilde{b}) + \mathrm{P}(\tilde{b} > \tilde{a}). \tag{10.49}$$

Under the assumptions, these two probabilities are the same. Since $f_{\tilde{a},\tilde{b}}(a,b) = f_{\tilde{a}}(a) f_{\tilde{a}}(b)$:

$$\begin{aligned}
\mathrm{P}(\tilde{a} > \tilde{b}) &= \int_{b=-\infty}^{\infty} \int_{a=b}^{\infty} f_{\tilde{a},\tilde{b}}(a,b) \, \mathrm{d}a \, \mathrm{d}b && (10.50)\\
&= \int_{b=-\infty}^{\infty} \int_{a=b}^{\infty} f_{\tilde{a}}(a) f_{\tilde{a}}(b) \, \mathrm{d}a \, \mathrm{d}b && (10.51)\\
&= \int_{a=-\infty}^{\infty} \int_{b=a}^{\infty} f_{\tilde{a}}(b) f_{\tilde{a}}(a) \, \mathrm{d}a \, \mathrm{d}b && (10.52)\\
&= \int_{a=-\infty}^{\infty} \int_{b=a}^{\infty} f_{\tilde{a},\tilde{b}}(a,b) \, \mathrm{d}a \, \mathrm{d}b && (10.53)\\
&= \mathrm{P}(\tilde{b} > \tilde{a}). && (10.54)
\end{aligned}$$

We conclude that $\mathrm{P}(\tilde{a} > \tilde{b}) = 1/2$.

b) Let $\tilde{x}_i$ and $\tilde{y}_i$ denote the first and second entry for each data point. By the previous question, $1\left(\tilde{y}_i > \tilde{x}_i\right)$ is a Bernoulli random variable with parameter $1/2$. Consequently, under the null hypothesis the test statistic is a binomial random variable with parameters n and $\theta := 1/2$.

c) The test statistic equals 7. Let $\tilde{t}_{\text{null}}$ denote its distribution under the null hypothesis. The p value equals

$$\begin{aligned}
\mathrm{pv}\left(7\right) &:= \mathrm{P}\left(\tilde{t}_{\text{null}} \geq 7\right) && (10.55)\\
&= \frac{1}{2^{10}} \sum_{i=7}^{10} \binom{10}{i} && (10.56)\\
&= 0.172. && (10.57)
\end{aligned}$$

d) The sign test is still valid because under these assumptions $1\left(\tilde{y}_i > \tilde{x}_i\right)$ is still a Bernoulli random variable with parameter $1/2$

10.11 (Drug design)

a) Let $\tilde{t}_{\text{null}}$ denote the test statistic under the null hypothesis. The p value function equals

$$\begin{aligned}
\mathrm{pv}(t) &:= \mathrm{P}\left(\tilde{t}_{\text{null}} \geq t\right) && (10.58)\\
&= \int_{a=t}^{\infty} \exp(-a) \, \mathrm{d}a && (10.59)\\
&= \exp(-t). && (10.60)
\end{aligned}$$

If $\tilde{t}_\theta$ is the test statistic when the drug is effective, the corresponding p value equals

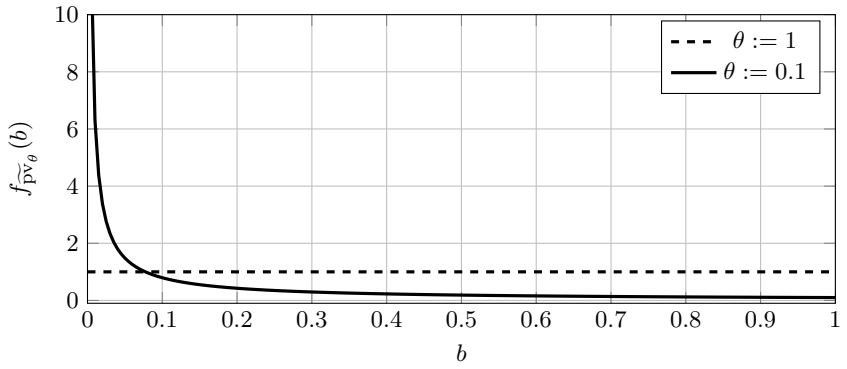
$\widetilde{\text{pv}}_\theta := \text{pv}(\tilde{t}_\theta) = \exp(-\tilde{t}_\theta)$. For $0 \leq b \leq 1$, the cdf of this random variable is

$$\begin{aligned} F_{\widetilde{\text{pv}}_\theta}(b) &= \text{P}\left(\widetilde{\text{pv}}_\theta \leq b\right) && (10.61)\\ &= \text{P}\left(\exp(-\tilde{t}_\theta) \leq b\right) && (10.62)\\ &= \text{P}\left(\tilde{t}_\theta \geq \log\left(\frac{1}{b}\right)\right) && (10.63)\\ &= \int_{a=\log\left(\frac{1}{b}\right)}^{\infty} \theta \exp(-\theta a)\ \text{d}a && (10.64)\\ &= \exp(-\theta \log\left(\frac{1}{b}\right)) && (10.65)\\ &= b^\theta. && (10.66) \end{aligned}$$

The corresponding pdf equals

$$f_{\widetilde{\text{pv}}_\theta}(b) = \theta b^{\theta-1}, \tag{10.67}$$

for $0 \leq b \leq 1$ and zero otherwise. For $\theta := 1$ unsurprisingly, the distribution is uniform, since this corresponds to the null hypothesis. The plot for $\theta := 0.1$ is shown below.



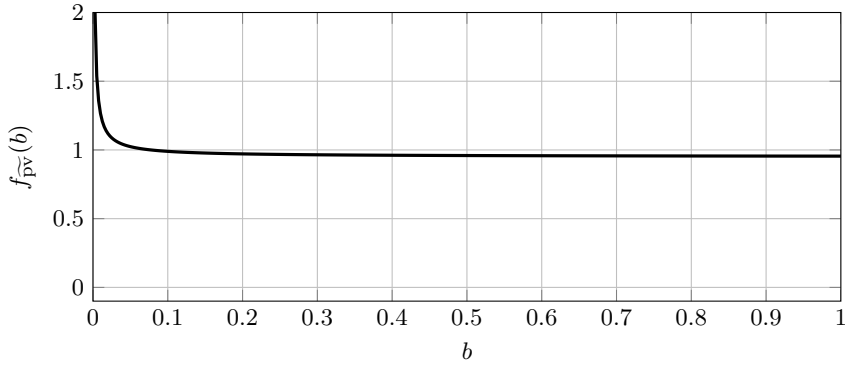
b) Let $\widetilde{\text{pv}}$ represent the p value and let $\tilde{d}$ be a Bernoulli random variable that equals one if the drug is effective and zero otherwise. From the previous question we know that:

$$\begin{aligned} f_{\widetilde{\text{pv}}\,|\,\tilde{d}}(b\,|\,0) &= f_{\widetilde{\text{pv}}_0}(b) = 1, && (10.68)\\ f_{\widetilde{\text{pv}}\,|\,\tilde{d}}(b\,|\,1) &= f_{\widetilde{\text{pv}}_\theta}(b) = \theta b^{\theta-1}, && (10.69) \end{aligned}$$

for $0 \leq b \leq 1$ and zero otherwise. The pdf of the p value therefore equals

$$\begin{aligned} f_{\widetilde{\text{pv}}}(b) &= \sum_{b=0}^{1} p_{\tilde{b}}(b) f_{\widetilde{\text{pv}}\,|\,\tilde{d}}(b\,|\,0) && (10.70)\\ &= 0.95 + 0.05\theta b^{\theta-1}, && (10.71) \end{aligned}$$

for $0 \leq b \leq 1$ and zero otherwise. Here is the pdf for $\theta := 0.1$:



c) We reject the null hypothesis if the p value $\widetilde{\text{pv}} \leq \alpha$. A false positive occurs if we reject when $\tilde{d} = 0$. The conditional probability of a false positive given that we reject the null hypothesis is therefore equal to

$$\text{P}\left(\tilde{d} = 0 \,|\, \widetilde{\text{pv}} \leq \alpha\right) = \frac{\text{P}\left(\tilde{d} = 0, \widetilde{\text{pv}} \leq \alpha\right)}{\text{P}\left(\widetilde{\text{pv}} \leq \alpha\right)}. \tag{10.72}$$

We have

$$\begin{aligned} \text{P}\left(\tilde{d} = 0, \widetilde{\text{pv}} \leq \alpha\right) &= \text{P}\left(\tilde{d} = 0\right) \text{P}\left(\widetilde{\text{pv}} \leq \alpha \,|\, \tilde{d} = 0\right) && (10.73)\\ &= 0.95 F_{\widetilde{\text{pv}}_0}(\alpha) && (10.74)\\ &= 0.95\alpha. && (10.75) \end{aligned}$$

From (b) we know that the conditional cdf of $\widetilde{\text{pv}}$ given $\tilde{d} = 1$ is

$$F_{\widetilde{\text{pv}} \,|\, \tilde{d}}(b \,|\, 1) = b^\theta, \tag{10.76}$$

so

$$\begin{aligned} \text{P}\left(\widetilde{\text{pv}} \leq \alpha\right) &= \sum_{d=0}^{1} \text{P}\left(\tilde{d} = d\right) \text{P}\left(\widetilde{\text{pv}} \leq \alpha \,|\, \tilde{d} = d\right) && (10.77)\\ &= 0.95\alpha + 0.05 F_{\widetilde{\text{pv}} \,|\, \tilde{d}}(\alpha \,|\, 1) && (10.78)\\ &= 0.95\alpha + 0.05\alpha^\theta. && (10.79) \end{aligned}$$

We conclude that

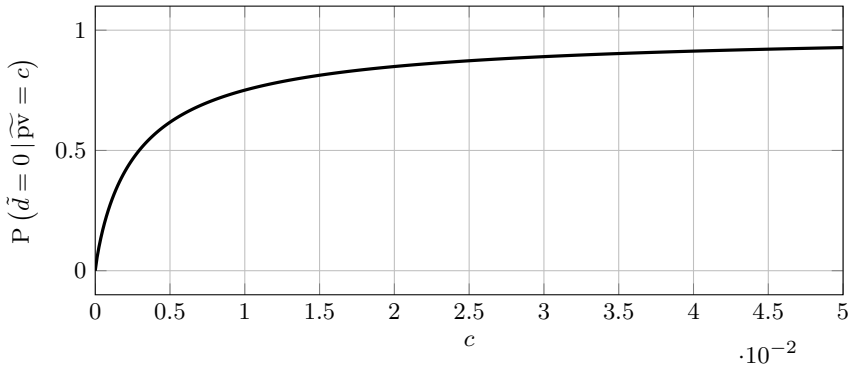
$$\text{P}\left(\tilde{d} = 0 \,|\, \widetilde{\text{pv}} \leq \alpha\right) = \frac{0.95\alpha}{0.95\alpha + 0.05\alpha^\theta}. \tag{10.80}$$

For $\alpha = 0.05$ and $\theta := 0.1$ the probability equals 0.562. More than half of the findings are false positives.

d) By Bayes rule for discrete and continuous random variables,

$$\begin{aligned} \text{P}\left(\tilde{d} = 0 \,|\, \widetilde{\text{pv}} = c\right) &= \frac{p_{\tilde{d}}(0) f_{\widetilde{\text{pv}} \,|\, \tilde{d}}(c \,|\, 0)}{f_{\widetilde{\text{pv}}}(c)} && (10.81)\\ &= \frac{0.95 \cdot 1}{0.95 + 0.05\theta c^{\theta - 1}}. && (10.82) \end{aligned}$$

Here is the plot for $\theta := 0.1$:



The probability increases monotonically with the p value, which suggests doing follow-up testing based on the associated p value, starting with the drug that has the smallest p value.

10.12 (P-hacking)

a) For $u \leq \alpha$, the conditional cdf equals

$$F_{\tilde{u} \,|\, \tilde{u} \leq \alpha}(u) = \mathrm{P}\left(\tilde{u} \leq u \,|\, \tilde{u} \leq \alpha\right) \tag{10.83}$$

$$= \frac{\mathrm{P}\left(\tilde{u} \leq u, \tilde{u} \leq \alpha\right)}{\mathrm{P}\left(\tilde{u} \leq \alpha\right)} \tag{10.84}$$

$$= \frac{\mathrm{P}\left(\tilde{u} \leq u\right)}{\mathrm{P}\left(\tilde{u} \leq \alpha\right)} \tag{10.85}$$

$$= \frac{F_{\tilde{u}}(u)}{F_{\tilde{u}}(\alpha)} \tag{10.86}$$

$$= \frac{u}{\alpha}. \tag{10.87}$$

For $u < 0$, $F_{\tilde{u} \,|\, \tilde{u} \leq \alpha}(u) = 0$ and for $u > \alpha$, $F_{\tilde{u} \,|\, \tilde{u} \leq \alpha}(u) = 1$. The pdf is therefore equal to $1/\alpha$ between 0 and α. The conditional distribution is uniform between 0 and α.

b) The histogram is exactly equal to the continuous pdf derived in the previous question for $\alpha := 0.05$. Since for continuous test statistics, the p value is uniformly distributed between 0 and 1, this is the distribution we would see if the null hypothesis always holds, but only p values smaller than α are reported. This suggests that the publications could be based on meaningless results that are only statistically significant by chance.

c) If our conjecture holds, the group is only reporting 5% of their results, which means that there would be approximately 2,000 total results, and therefore 1,900 unpublished ones.

11

Principal Component Analysis And Low-Rank Models

Exercises

11.1 (Random vector)

a) The covariance matrix equals

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.25 & 0.25 \\ 0 & 0.25 & 0.25 \end{bmatrix}, \tag{11.1}$$

so $\text{Var}[\tilde{x}_1] = 1$, $\text{Var}[\tilde{x}_2] = 0.25$, $\text{Var}[\tilde{x}_3] = 0.25$.

b) The maximum variance in any direction is given by the largest eigenvalue, which is equal to 1. There cannot be another direction with higher variance.

c) Let $\tilde{y} := a_1\tilde{x}_1 + a_2\tilde{x}_2 + a_3\tilde{x}_3$. By linearity of expectation, $\mathbb{E}[\tilde{y}] = 0$. By Chebyshev's inequality, if $\text{Var}[\tilde{y}] = 0$ then $\text{P}(Y \neq 0) = 1$, which is exactly what we want. According to the eigendecomposition, the variance is zero in the direction of the third eigenvector, so setting $a_1 = 0$, $a_2 = 1/\sqrt{2}$, and $a_3 = -1/\sqrt{2}$ does the trick.
Geometrically, the random vector lies in the plane orthogonal to the vector

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \tag{11.2}$$

with probability one.

11.2 (Basketball team) Let $\tilde{x}_A$, $\tilde{x}_B$ and $\tilde{x}_C$ denote the points scored by each player. We have

$$\begin{aligned}
\text{Var}[\tilde{x}_A + \tilde{x}_B] &= \text{Var}[\tilde{x}_A] + \text{Var}[\tilde{x}_B] + 2\text{Cov}[\tilde{x}_A, \tilde{x}_B] && (11.3)\\
&= 100 + 81 - 160 && (11.4)\\
&= 21, && (11.5)\\
\text{Var}[\tilde{x}_A + \tilde{x}_C] &= \text{Var}[\tilde{x}_A] + \text{Var}[\tilde{x}_C] + 2\text{Cov}[\tilde{x}_A, \tilde{x}_C] && (11.6)\\
&= 100 + 100 + 20 && (11.7)\\
&= 220, && (11.8)\\
\text{Var}[\tilde{x}_B + \tilde{x}_C] &= \text{Var}[\tilde{x}_B] + \text{Var}[\tilde{x}_C] + 2\text{Cov}[\tilde{x}_B, \tilde{x}_C] && (11.9)\\
&= 81 + 100 + 100 && (11.10)\\
&= 281. && (11.11)
\end{aligned}$$

If the team is winning by a lot, it is probably a good idea to minimize the variance, so we would recommend playing A and B. If they are losing by a lot, they should maximize the variance to try to win, so we would recommend playing B and C (but note that this could result in an even more humiliating loss!).

11.3 (Correlation and PCA) If the correlation coefficient is one, then the covariance equals the product of the standard deviations $\text{Cov}(\tilde{x}[1], \tilde{x}[2]) = \sigma_{\tilde{x}[1]}\sigma_{\tilde{x}[2]}$. The covariance matrix equals

$$\Sigma_{\tilde{x}} = \begin{bmatrix} \sigma^2_{\tilde{x}[1]} & \sigma_{\tilde{x}[1]}\sigma_{\tilde{x}[2]} \\ \sigma_{\tilde{x}[1]}\sigma_{\tilde{x}[2]} & \sigma^2_{\tilde{x}[2]} \end{bmatrix} \tag{11.12}$$

$$= \begin{bmatrix} \sigma_{\tilde{x}[1]} \\ \sigma_{\tilde{x}[2]} \end{bmatrix} \begin{bmatrix} \sigma_{\tilde{x}[1]} & \sigma_{\tilde{x}[2]} \end{bmatrix}. \tag{11.13}$$

The rank of the matrix is one. The first eigenvalue equals $\sigma^2_{\tilde{x}[1]} + \sigma^2_{\tilde{x}[2]}$ and the second equals zero. The variance of the second principal component is therefore zero. Intuitively, both components are linearly dependent, so the data lie along a line. The variance along the line is the variance of the first principal component. The variance orthogonal to the line, which is the variance of the second principal component, is zero because all the points are on the line.

11.4 (Principal components are uncorrelated) Let X be a dataset containing n vectors x_1, x_2, ..., $x_n \in \mathbb{R}^d$ with d features each. The sample mean of each principal component is zero, for $1 \leq j \leq d$,

$$\frac{1}{n}\sum_{i=1}^{n} w_j[i] = \frac{1}{n}\sum_{i=1}^{n} u_j^T \,\text{ct}\,(x_i) \tag{11.14}$$

$$= u_j^T \left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{l=1}^{n} x_l \right) = 0, \tag{11.15}$$

where $u_1, \ldots, u_d$ are the eigenvectors of the sample covariance matrix Σ_X. Consequently, the sample covariance between two principal components ($j \neq k$) equals

$$\frac{1}{n-1}\sum_{i=1}^{n} w_j[i]w_k[i] = \frac{1}{n-1}\sum_{i=1}^{n} u_j^T \,\text{ct}\,(x_i)\,\text{ct}\,(x_i)^T\, u_k \tag{11.16}$$

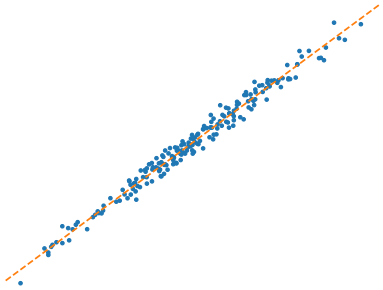
$$= u_j^T \Sigma_X u_k \tag{11.17}$$

$$= \lambda_k u_j^T u_k = 0, \tag{11.18}$$

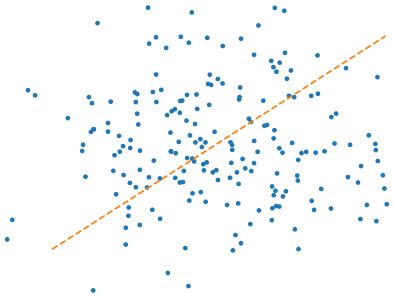
because u_j and u_k are orthogonal by the spectral theorem.

11.5 (Estimating a direction)

a) To produce the samples we can first sample from $\tilde{x}$ and multiply the value by v. This will produce a point on the line collinear with v at an average distance from the origin of σ_{signal}. This is then added to a sample from a Gaussian with independent entries centered at the origin with standard deviation σ_{noise}. If σ_{signal} is much larger than σ_{noise} then the samples will lie close to v. They look like this (the dashed line is collinear with v):



b) If σ_{noise} is much larger than σ_{signal} then the noise component dominates, and the samples will not lie as close to v. They look like this (the dashed line is collinear with v):



c) No. By linearity of expectation, the mean of the random variable that represents the data equals

$$\begin{aligned}\mathbb{E}\left[\tilde{y}\right] &= \mathbb{E}\left[\tilde{x}v + \tilde{z}\right] && (11.19)\\ &= \mathbb{E}\left[\tilde{x}\right]v + \mathbb{E}\left[\tilde{z}\right] && (11.20)\\ &= 0, && (11.21)\end{aligned}$$

so we won't be able to estimate v by averaging.

d) By linearity of expectation and the fact that the mean of $\tilde{x}$ and $\tilde{z}$ is zero,

$$\begin{aligned}\mathbb{E}\left[\tilde{y}\tilde{y}^T\right] &= \mathbb{E}\left[(\tilde{x}v + \tilde{z})(\tilde{x}v + \tilde{z})^T\right] && (11.22)\\ &= \mathbb{E}\left[\tilde{x}^2\right]vv^T + \mathbb{E}\left[\tilde{z}\tilde{z}^T\right] && (11.23)\\ &= \sigma^2_{\text{signal}}vv^T + \sigma^2_{\text{noise}}I. && (11.24)\end{aligned}$$

e) Setting $U := \begin{bmatrix} v & u_2 & \cdots & u_d \end{bmatrix}$, we have $UU^T = I$ by the orthonormality assumption. We can therefore rewrite the covariance matrix as

$$\begin{aligned}\mathbb{E}\left[\tilde{y}\tilde{y}^T\right] &= \sigma^2_{\text{signal}}vv^T + \sigma^2_{\text{noise}}UU^T && (11.25)\\ &= U\begin{bmatrix}\sigma^2_{\text{signal}} + \sigma^2_{\text{noise}} & 0_{d-1} \\ 0_{d-1} & \sigma^2_{\text{noise}}I_{d-1}\end{bmatrix}U^T, && (11.26)\end{aligned}$$

where 0_{d-1} is a vector of $d-1$ zeros and I_{d-1} is the $d-1 \times d-1$ identity matrix.

f) Apply principal component analysis (i.e. compute the eigendecomposition of the sample covariance matrix) and use the first principal direction (i.e. the eigenvector with the largest eigenvalue) as an estimate of the direction of v.

11.6 (Normalization)

a) We have $\tilde{y} = A\tilde{x}$, where

$$A := \begin{bmatrix} \frac{1}{10} & 0 & 0 \\ 0 & \frac{1}{20} & 0 \\ 0 & 0 & \frac{1}{0.4} \end{bmatrix}, \tag{11.27}$$

so

$$\Sigma_{\tilde{y}} := \mathbb{E}\left[\tilde{y}\tilde{y}^T\right] \tag{11.28}$$

$$= \mathbb{E}\left[A\tilde{x}\tilde{x}^T A^T\right] \tag{11.29}$$

$$= A\Sigma_{\tilde{x}} A^T \tag{11.30}$$

$$= \begin{bmatrix} 1 & 0.125 & 0 \\ 0.125 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{11.31}$$

b) No. For example, in the direction of the vector

$$v := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} \tag{11.32}$$

the directional variance of $\tilde{y}$ equals

$$\text{Var}\left[v^T\tilde{y}\right] = v^T\Sigma_{\tilde{y}} v \tag{11.33}$$

$$= 1.125. \tag{11.34}$$

c) The eigendecomposition of $\Sigma_{\tilde{x}}$ yields eigenvalues equal to 402, 97.9 and 0.16, corresponding to the eigenvectors

$$u_1(\tilde{x}) = \begin{bmatrix} -0.997 \\ 0.082 \\ 0 \end{bmatrix}, \quad u_2(\tilde{x}) = \begin{bmatrix} -0.082 \\ -0.997 \\ 0 \end{bmatrix}, \quad u_3(\tilde{x}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{11.35}$$

respectively. To reduce dimensionality we compute the inner product with $u_1(\tilde{x})$ and $u_2(\tilde{x})$.

The eigendecomposition of $\Sigma_{\tilde{y}}$ yields eigenvalues equal to 1.125, 1 and 0.875, corresponding to the eigenvectors

$$u_1(\tilde{y}) = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}, \quad u_2(\tilde{y}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad u_3(\tilde{y}) = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \tag{11.36}$$

respectively. To reduce dimensionality we compute the inner product with $u_1(\tilde{y})$ and $u_2(\tilde{y})$.

d) (1) Using $\tilde{y}$ is probably better because the variances and covariances correspond to different units that are not directly comparable. Normalizing makes the analysis independent of the units. The third feature is height in meters so it is normal that its variance is smaller than one. This does not mean that it is negligible.

(2) Using $\tilde{x}$ is probably better. The three features are dimensions of cars in the same unit. The variance of the height is very small compared to the other two, which indicates that the height of the cars does not vary at all. It therefore makes sense to ignore it while reducing the dimensionality.

11.7 (PCA and sample variance) Let $X := \{x_1, \ldots, x_n\}$ denote the dataset of interest. We want to establish that the first k eigenvectors of the sample covariance matrix Σ_X of X, associated to the k largest eigenvalues, satisfy

$$\{u_1, ..., u_k\} = \arg \max_{\substack{\{b_1,\ldots,b_k\} \\ ||b_j||_2=1 \; 1\leq j\leq k \\ b_j \perp b_k \text{ for } j\neq k}} \sum_{j=1}^{k} v(X_{b_j}), \tag{11.37}$$

where $v(X_{b_j})$ denotes the sample variance of the set $X_{b_j} := \left\{b_j^T x_1, ..., b_j^T x_n\right\}$. We prove this mimicking the proof of the theorem that establishes the analogous result for random vectors, which leverages induction on k. The base case $k := 1$ follows immediately from the fact that the first eigenvector maximizes the sample variance,

$$u_1 = \arg \max_{||b||_2=1} v(X_b). \tag{11.38}$$

To complete the proof we establish that if the induction hypothesis

$$\{u_1, ..., u_{k-1}\} = \arg \max_{\substack{\{b_1,\ldots,b_{k-1}\} \\ ||b_j||_2=1 \; 1\leq j\leq k-1 \\ b_j \perp b_{k-1} \text{ for } j\neq k-1}} \sum_{j=1}^{k-1} v(X_{b_j}) \tag{11.39}$$

holds, then (11.37) holds. We now set $b_1, \ldots, b_k$ to be an arbitrary fixed set of k orthonormal vectors, and show that they cannot capture more variance than $u_1, \ldots, u_k$ if the induction hypothesis holds. Consider the subspace $\mathcal{S} := \text{span}(b_1, \ldots, b_k)$ spanned by $b_1, \ldots, b_k$. We are interested in the projection of the data onto this subspace, because

$$\sum_{j=1}^{k} v(X_{b_j}) = \sum_{j=1}^{k} \frac{1}{n-1} \sum_{i=1}^{n} \left(b_i^T x_i\right)^2 \tag{11.40}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left|\left| \sum_{j=1}^{k} b_j^T x_i b_j \right|\right|_2^2 \tag{11.41}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} ||\mathcal{P}_{\mathcal{S}}\, x_i||_2^2 \,. \tag{11.42}$$

We can represent this projection using an arbitrary orthonormal basis of $\mathcal{S}$. Indeed, any set of k orthonormal vectors $a_1, \ldots, a_k$ spanning $\mathcal{S}$ satisfy

$$\mathcal{P}_{\mathcal{S}}\, x_i := \sum_{j=1}^{k} b_j^T x_i b_j = \sum_{j=1}^{k} a_j^T x_i a_j. \tag{11.43}$$

The key is to choose the basis wisely. $\mathcal{S}$ has dimension k, so it must contain at least one vector $a_\perp$ that is orthogonal to $u_1, u_2, \ldots, u_{k-1}$. By the properties of the principal directions, the sample variance in that direction cannot be higher than in the kth principal direction

$$v\left(X_{u_k}\right) \geq v\left(X_{a_\perp}\right). \tag{11.44}$$

We build our wisely-chosen orthonormal basis $a_1, a_2, \ldots, a_k$ for $\mathcal{S}$ setting $a_k := a_\perp$ (we can construct such a basis via the Gram-Schmidt process, starting with $a_\perp$). By the induction hypothesis,

$$\sum_{j=1}^{k-1} v\left(X_{u_j}\right) \geq \sum_{j=1}^{k-1} v\left(X_{a_j}\right). \tag{11.45}$$

Combining (11.45) and (11.44) yields

$$\sum_{j=1}^{k} v\left(X_{u_j}\right) \geq \sum_{j=1}^{k} v\left(X_{a_j}\right) \tag{11.46}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} ||\mathcal{P}_{\mathcal{S}}\, x_i||_2^2 \tag{11.47}$$

$$= \sum_{j=1}^{k} v\left(X_{b_j}\right). \tag{11.48}$$

Since this holds for any choice of b_1, ..., b_k the proof is complete.

11.8 (Whitening) If the entries of $\tilde{y}$ are all uncorrelated and have unit variance then $\tilde{y}$ has a covariance matrix equal to the identity. The covariance matrix equals

$$\Sigma_{\tilde{y}} := \mathbb{E}\left[\tilde{y}\tilde{y}^T\right] = \mathbb{E}\left[A\tilde{x}\tilde{x}^T A^T\right] \tag{11.49}$$

$$= A\Sigma_{\tilde{x}} A^T. \tag{11.50}$$

Now in order for this product to equal the identity we build A based on the eigendecomposition of $\Sigma_{\tilde{x}} = U\Lambda U^T$. Since covariance matrices are positive semidefinite, and $\Sigma_{\tilde{x}}$ is full rank, the entries of the diagonal matrix Λ are positive, so we can define $\sqrt{\Lambda}$, which is a diagonal matrix that contains the square root of the entries of Λ. Setting $A := \sqrt{\Lambda}^{-1} U^T$, since U is an orthogonal matrix by the spectral theorem so that $U^T U = I$,

$$\Sigma_{\tilde{y}} = A\Sigma_{\tilde{x}} A^T \tag{11.51}$$

$$= \sqrt{\Lambda}^{-1} U^T U \Lambda U^T U \sqrt{\Lambda}^{-1} \tag{11.52}$$

$$= \sqrt{\Lambda}^{-1} \Lambda \sqrt{\Lambda}^{-1} = I. \tag{11.53}$$

11.9 (Numerical rank) Let us assume that M has dimensions $n_1 \times n_2$ and $n_1 \leq n_2$ (this is without loss of generality, because we can always consider the transpose of M if $n_2 < n_1$). Let $s_1 > s_2 > ... > s_{n_1}$ be the singular values of M. The numerical rank r is the smallest positive integer such that

$$\sum_{i=r+1}^{n} s_i^2 \leq \epsilon^2. \tag{11.54}$$

To compute it, we sum the squared singular values starting from the smallest one until the sum reaches ϵ^2. Consider the rank-r matrix

$$L_{\text{SVD}} := \sum_{l=1}^{r} s_l u_l v_l^T, \tag{11.55}$$

obtained by truncating the SVD of M. As shown in the notes (eq. 11.162):

$$||M - L_{\text{SVD}}||_{\text{F}} = \sqrt{\sum_{l=r+1}^{n_1} s_i^2}, \tag{11.56}$$

which is smaller or equal to ϵ by the definition of r. Consequently, there is a rank r matrix such that the distance to M is smaller or equal to ϵ.
Now we show that there cannot be a rank-r' matrix L' with $r' < r$, such that $\left|\left|L' - M\right|\right|_{\text{F}} \leq \epsilon$. Consider the r' low-rank approximation to M,

$$L_{\text{SVD}} := \sum_{l=1}^{r'} s_l u_l v_l^T. \tag{11.57}$$

This is the optimal rank r' approximation, so

$$||L - M||_{\text{F}}^2 \geq ||L_{\text{SVD}} - M||_{\text{F}}^2 \tag{11.58}$$

$$= \sum_{i=r'+1}^{n_1} s_i^2. \tag{11.59}$$

If $\left|\left|L' - M\right|\right|_{\text{F}} \leq \epsilon$, then $\sum_{i=r'+1}^{n_1} s_i^2 \leq \epsilon^2$. However, this is impossible because r is defined to be the smallest positive integer for which $\sum_{i=r+1}^{n} s_i^2 \leq \epsilon^2$.

11.10 (Double centering)

a) The average of the entries of the column sample mean is the grand sample mean,

$$\frac{1}{n_2}\sum_{j=1}^{n_2} \mu_{\text{row}}[j] = \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{n_1}\sum_{i=1}^{n_1} D[i,j] \tag{11.60}$$

$$= \mu_{\text{all}}. \tag{11.61}$$

The average of the entries of the row sample mean is also the grand sample mean,

$$\frac{1}{n_1}\sum_{i=1}^{n_1} \mu_{\text{col}}[i] = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{n_2}\sum_{j=1}^{n_2} D[i,j] \tag{11.62}$$

$$= \mu_{\text{all}}. \tag{11.63}$$

b) We define a matrix S obtained by subtracting the row sample mean and the row column mean,

$$S[i,j] := D[i,j] - \mu_{\text{col}}[i] - \mu_{\text{row}}[j], \quad 1 \leq i \leq n_1, 1 \leq j \leq n_2. \tag{11.64}$$

The jth entry of the row sample mean of S equals,

$$\frac{1}{n_1}\sum_{i=1}^{n_1} S[i,j] = \frac{1}{n_1}\sum_{i=1}^{n_1}(D[i,j] - \mu_{\text{col}}[i] - \mu_{\text{row}}[j]) \tag{11.65}$$

$$= \mu_{\text{row}}[j] - \frac{1}{n_1}\sum_{i=1}^{n_1}\mu_{\text{col}}[i] - \mu_{\text{row}}[j] \tag{11.66}$$

$$= -\mu_{\text{all}}, \tag{11.67}$$

where the last equality follows from (11.63). Consequently, the row sample mean of the matrix is constant and equal to the additive inverse of the grand sample mean of the

original matrix. This is also the case for the column sample mean. The ith entry of the column sample mean equals

$$\frac{1}{n_2}\sum_{j=1}^{n_2} S[i,j] = \frac{1}{n_2}\sum_{j=1}^{n_2} \left(D[i,j] - \mu_{\text{col}}[i] - \mu_{\text{row}}[j]\right) \tag{11.68}$$

$$= \mu_{\text{col}}[i] - \mu_{\text{col}}[i] - \frac{1}{n_2}\sum_{j=1}^{n_2} \mu_{\text{row}}[j] \tag{11.69}$$

$$= -\mu_{\text{all}}, \tag{11.70}$$

where the last equality follows from (11.61).

c) Above we see that subtracting the row and column sample means results in a matrix where the row and column sample means equal the additive inverse of the grand sample mean. This suggests adding it to each entry to cancel out both means. The entries of the resulting doubly centered matrix are defined by

$$C[i,j] := D[i,j] - \mu_{\text{col}}[i] - \mu_{\text{row}}[j] + \mu_{\text{all}}, \quad 1 \leq i \leq n_1, 1 \leq j \leq n_2. \tag{11.71}$$

We check that the matrix indeed has zero column and row sample means. The jth entry of the row sample mean of C equals,

$$\frac{1}{n_1}\sum_{i=1}^{n_1} C[i,j] = \frac{1}{n_1}\sum_{i=1}^{n_1} (D[i,j] - \mu_{\text{col}}[i] - \mu_{\text{row}}[j] + \mu_{\text{all}}) \tag{11.72}$$

$$= \mu_{\text{row}}[j] - \frac{1}{n_1}\sum_{i=1}^{n_1} \mu_{\text{col}}[i] - \mu_{\text{row}}[j] + \mu_{\text{all}} \tag{11.73}$$

$$= 0, \tag{11.74}$$

where the last equality follows from (11.63). The argument for the column sample mean is essentially the same.

11.11 (Mutant mosquito)

a) Based on the SVD analysis, which yields the rank-one model

$$L[i,j] := m(D) + s_i u_1[i] v_1[j], \tag{11.75}$$

we can set $c := m(D) = 3$ and $b[\text{user}]$ equal to the corresponding entry of v_1, so

$$b = \begin{matrix} \text{Bob} & \text{Molly} & \text{Mary} & \text{Larry} \end{matrix} \\ \begin{pmatrix} 0.48 & 0.52 & -0.48 & -0.52 \end{pmatrix}. \tag{11.76}$$

b) The residual sum of squared errors is

$$\text{error} = \sum_{\text{user} \in \{\text{Bob,Mary,Larry}\}} (\text{estimate} - \text{observed})^2 \tag{11.77}$$

$$= (3 + 0.48a[\text{MM}] - 2)^2 + (3 - 0.48a[\text{MM}] - 4)^2 + (3 - 0.052a[\text{MM}] - 4)^2$$

$$= 3 + 2.024a[\text{MM}] + 0.464a[\text{MM}]^2. \tag{11.78}$$

This is a convex quadratic (as the quadratic coefficient is positive). Setting its derivative

to zero yields the coefficient that minimizes the sum of squared errors:

$$a_{\text{ls}}[\text{MM}] = -\frac{2.024}{2 \cdot 0.464} \tag{11.79}$$

$$= -2.18. \tag{11.80}$$

c) The estimate for Molly is

$$\text{Molly} := c - a_{\text{ls}}[\text{MM}]b[\text{Molly}] \tag{11.81}$$

$$= 3 - 2.18 \cdot 0.52 = 1.87. \tag{11.82}$$

11.12 (Exact matrix completion) The matrix cannot be rank 1, because the first and second columns are clearly linearly independent. We will try to find a rank-2 matrix that is consistent with the revealed entry. If the matrix is rank 2, then the third column is a linear combination of the first and second columns with coefficients that satisfy

$$2\alpha + 2\beta = 3, \tag{11.83}$$

$$2\alpha + 3\beta = 4. \tag{11.84}$$

so $\alpha = 0.5$ and $\beta = 1$. As a result, $D[4, 3] = 5$.
Similarly, the fourth column is a linear combination of the first and second columns with coefficients that satisfy

$$2\gamma + 3\theta = -2, \tag{11.85}$$

$$2\gamma + 4\theta = -3. \tag{11.86}$$

so $\gamma = 0.5$ and $\theta = -1$. As a result, $D[2, 4] = -1$.
Finally, the first row must be a linear combination of the second and third rows with coefficients that satisfy

$$3\delta + 4\psi = 2, \tag{11.87}$$

$$-\delta - 2\psi = 0. \tag{11.88}$$

so $\delta = 2$ and $\psi = -1$. As a result, $D[1, 1] = 2$ and $D[1, 2] = 1$.
We have succeeded in completing the entries with a rank-2 matrix, which equals

$$D = \begin{bmatrix} 2 & 1 & 2 & 0 \\ 2 & 2 & 3 & -1 \\ 2 & 3 & 4 & -2 \\ 2 & 4 & 5 & -3 \end{bmatrix}. \tag{11.89}$$

11.13 (Topic modeling)

a) The singular values equal

$$D = USV^T = U \begin{bmatrix} 23.64 & 0 & 0 & 0 & & \\ 0 & 18.82 & 0 & 0 & 0 & 0 \\ 0 & 0 & 14.23 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.63 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.03 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.36 \end{bmatrix} V^T \tag{11.90}$$

Three of the singular values are much larger than the rest, which suggests fitting a

model with three topics. The singular vectors do not have an intuitive interpretation. In particular, they do not allow to cluster the words

$$\begin{array}{cccccccc} & & \text{a} & \text{b} & \text{c} & \text{d} & \text{e} & \text{f} \\ u_1 & = & \begin{pmatrix}-0.24 & -0.47 & -0.24 & -0.32 & -0.58 & -0.47\end{pmatrix} \\ u_2 & = & \begin{pmatrix} 0.64 & -0.23 & 0.67 & -0.03 & -0.18 & -0.21\end{pmatrix} \\ u_3 & = & \begin{pmatrix}-0.08 & -0.39 & -0.08 & 0.77 & 0.28 & -0.40\end{pmatrix} \end{array} \tag{11.91}$$

or the articles

$$\begin{array}{cccccccccc} & & \text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{concert} & \text{market} & \text{band} \\ v_1 & = & \begin{pmatrix}-0.18 & -0.24 & -0.51 & -0.38 & -0.46 & -0.34 & -0.2 & -0.3 & -0.22\end{pmatrix} \\ v_2 & = & \begin{pmatrix} 0.47 & 0.01 & -0.22 & -0.15 & -0.25 & -0.07 & 0.63 & -0.05 & 0.49\end{pmatrix} \\ v_3 & = & \begin{pmatrix}-0.13 & 0.47 & -0.3 & -0.14 & -0.37 & 0.52 & -0.04 & 0.49 & -0.07\end{pmatrix} \end{array}$$

b) The left and right singular vectors must be orthogonal, so the only way they can all have nonnegative entries is if their support is completely disjoint.

c) We apply NMF to obtain two matrices $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{k\times n}$ such that

$$M[i,j] \approx \sum_{l=1}^{3} w_l[i]h_l[j], \quad w_l[i] \geq 0, \qquad 1 \leq l \leq r,\ 1 \leq i \leq 6, \tag{11.92}$$

$$h_l[j] \geq 0, \qquad 1 \leq l \leq r,\ 1 \leq j \leq 9. \tag{11.93}$$

The vectors h_1, h_2 and h_3 cluster the words into three topics:

$$\begin{array}{cccccccccc} & & \text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{concert} & \text{market} & \text{band} \\ h_1 & = & \begin{pmatrix} 0.34 & 0 & 3.73 & 2.54 & 3.67 & 0.52 & 0 & 0.35 & 0.35\end{pmatrix} \\ h_2 & = & \begin{pmatrix} 0 & 2.21 & 0.21 & 0.45 & 0 & 2.64 & 0.21 & 2.43 & 0.22\end{pmatrix} \\ h_3 & = & \begin{pmatrix} 3.22 & 0.37 & 0.19 & 0.2 & 0 & 0.12 & 4.13 & 0.13 & 3.43\end{pmatrix} \end{array}$$

The first topic corresponds to the entries that are not zero (or very small) in h_1: senate, election and vote. The second corresponds to h_2: GDP, stock and market. The third corresponds to h_3: singer, concert and band.
The entries of w_1, w_2 and w_3 allow us to assign the topics to articles:

$$\begin{array}{cccccccc} & & \text{a} & \text{b} & \text{c} & \text{d} & \text{e} & \text{f} \\ w_1 & = & \begin{pmatrix}0.03 & 2.23 & 0 & 0 & 1.59 & 2.24\end{pmatrix} \\ w_2 & = & \begin{pmatrix} 0.1 & 0 & 0.08 & 3.13 & 2.32 & 0\end{pmatrix} \\ w_3 & = & \begin{pmatrix}2.13 & 0 & 2.22 & 0 & 0 & 0.03\end{pmatrix} \end{array} \tag{11.94}$$

Articles b, e and f are about politics (topic 1), d and e about economics (topic 3) and a and c about music (topic 3).

d) The Frobenius-norm error of the SVD-based approximation is smaller (4.38 compared to 4.44). This makes sense because it is optimal, and nonnegative matrix factorization has an additional constraint that can only make the fit worse. Nevertheless, the NMF model is clearly better for the purpose of topic modeling, because it yields a reasonable choice of models that fit the data well and allow us to cluster the words and articles.

12

Regression And Classification

Exercises

12.1 (Heart beat)

a) The simple-regression linear MMSE estimator of $\tilde{b}$ given $\tilde{x}[1]$ is

$$\ell_{\text{MMSE}}(\tilde{x}[1]) = \frac{\text{Cov}[\tilde{x}[1], \tilde{b}]\tilde{x}[1]}{\text{Var}\left[\tilde{x}[1]\right]}. \tag{12.1}$$

By the independence assumptions,

$$\begin{aligned}
\text{Var}[\tilde{x}[1]] &= \text{Var}[\tilde{b} + \tilde{m} + \tilde{z}_1] && (12.2)\\
&= \text{Var}[\tilde{b}] + \text{Var}[\tilde{m}] + \text{Var}[\tilde{z}_1] && (12.3)\\
&= 12, && (12.4)\\
\text{Cov}[\tilde{b}, \tilde{x}[1]] &= \mathbb{E}\left[\tilde{b}(\tilde{b} + \tilde{m} + \tilde{z}_1)\right] && (12.5)\\
&= \mathbb{E}[\tilde{b}^2] + \mathbb{E}\left[\tilde{b}\tilde{m}\right] + \mathbb{E}\left[\tilde{b}\tilde{z}_1\right] && (12.6)\\
&= \mathbb{E}[\tilde{b}^2] + \mathbb{E}[\tilde{b}]\mathbb{E}\left[\tilde{m}\right] + \mathbb{E}[\tilde{b}]\mathbb{E}\left[\tilde{z}_1\right] && (12.7)\\
&= 1. && (12.8)
\end{aligned}$$

Therefore,

$$\ell_{\text{MMSE}}(\tilde{x}[1]) = \frac{\tilde{x}[1]}{12}. \tag{12.9}$$

The estimator just shrinks the measurement.

The coefficient of determination equals

$$\begin{aligned}
R^2 &:= \frac{\text{Var}\left[\ell_{\text{MMSE}}(\tilde{x}[1])\right]}{\text{Var}[\tilde{b}]} && (12.10)\\
&= \frac{12}{12^2} = \frac{1}{12}, && (12.11)
\end{aligned}$$

which indicates that the estimate is not great.

b) The linear MMSE estimator equals

$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{b}}^T \Sigma_{\tilde{x}}^{-1} \tilde{x}. \tag{12.12}$$

By the independence assumptions, $\text{Var}\left[\tilde{x}[1]\right] = 12$,

$$\begin{aligned}
\text{Var}\left[\tilde{x}[2]\right] &= \text{Var}\left[\tilde{m}\right] + \text{Var}\left[\tilde{z}_2\right] && (12.13)\\
&= 11, && (12.14)\\
\text{Cov}\left[\tilde{x}[1]\tilde{x}[2]\right] &= \mathbb{E}\left[(\tilde{b} + \tilde{m} + \tilde{z}_1)(\tilde{m} + \tilde{z}_2)\right] && (12.15)\\
&= \mathbb{E}\left[\tilde{m}^2\right] && (12.16)\\
&= 10, && (12.17)
\end{aligned}$$

so

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 12 & 10 \\ 10 & 11 \end{bmatrix} \tag{12.18}$$

Again by the independence assumptions, $\text{Cov}\left[\tilde{b}, \tilde{x}[1]\right] = 1$, and

$$\begin{aligned}
\text{Cov}\left[\tilde{b}\tilde{x}[2]\right] &= \mathbb{E}\left[\tilde{b}(\tilde{m} + \tilde{z}_2)\right] && (12.19)\\
&= 0, && (12.20)
\end{aligned}$$

so

$$\Sigma_{\tilde{x}\tilde{b}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{12.21}$$

Therefore

$$\begin{aligned}
\ell_{\text{MMSE}}(\tilde{x}) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 12 & 10 \\ 10 & 11 \end{bmatrix}^{-1} \tilde{x} && (12.22)\\
&= \frac{11\tilde{x}[1] - 10\tilde{x}[2]}{32}. && (12.23)
\end{aligned}$$

The estimate approximately cancels out the signal from the mother.
The linear MMSE estimator can be expressed as

$$\begin{aligned}
\ell_{\text{MMSE}}(\tilde{x}) &= \frac{11(\tilde{b} + \tilde{m} + \tilde{z_1}) - 10(\tilde{m} + \tilde{z}_2)}{32} && (12.24)\\
&= \frac{11\tilde{b} + \tilde{m} + 11\tilde{z_1} - 10\tilde{z}_2}{32}. && (12.25)
\end{aligned}$$

By the independence assumption,

$$\begin{aligned}
\text{Var}\left[\ell_{\text{MMSE}}(\tilde{x})\right] &= \frac{11^2\text{Var}\left[\tilde{b}\right] + \text{Var}\left[\tilde{m}\right] + 11^2\text{Var}\left[\tilde{z}_1\right] + 10^2\text{Var}\left[\tilde{z}_2\right]}{32^2} && (12.26)\\
&= \frac{11^2 + 10 + 11^2 + 10^2}{32^2} = 0.344. && (12.27)
\end{aligned}$$

The coefficient of determination equals

$$\begin{aligned}
R^2 &:= \frac{\text{Var}\left[\ell_{\text{MMSE}}(\tilde{x})\right]}{\text{Var}[\tilde{b}]} && (12.28)\\
&= 0.344, && (12.29)
\end{aligned}$$

which is much higher than $1/12 = 0.083$.

12.2 (PCA and OLS) For PCA, we have

$$u_1 = \arg\max_{||v||_2=1} v^T \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right) v \tag{12.30}$$

$$= \arg\max_{||v||_2=1} \sum_{i=1}^{n} (v^T x_i)^2. \tag{12.31}$$

Let us define the ℓ_2-norm distance between x_i and the line collinear with a unit ℓ_2-norm vector v, as the distance between x_i and its nearest point on the line:

$$d_v(x_i) := \arg\min_{\gamma \in \mathbb{R}} ||x_i - \gamma v||_2^2 . \tag{12.32}$$

The nearest point to x_i on the line is the projection $(x_i^T v)v$ (you can check this by computing the derivative with respect to γ), which results in a residual that is orthogonal to the projection. Consequently, by Pythagoras' theorem,

$$d_v(x_i) = \left|\left| x_i - (x_i^T v)v \right|\right|_2^2 \tag{12.33}$$

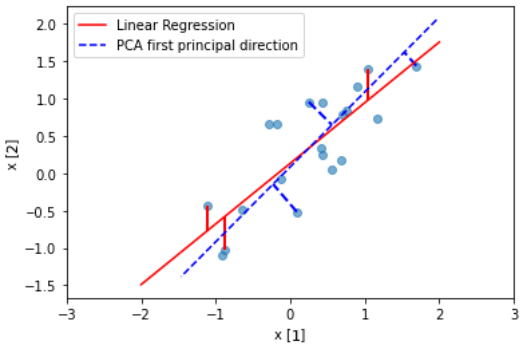
$$= ||x_i||_2^2 - \left|\left| (x_i^T v)v \right|\right|_2^2 \tag{12.34}$$

$$= ||x_i||_2^2 - (x_i^T v)^2. \tag{12.35}$$

We conclude that, by Eq. (12.31), the line collinear with u_1 minimizes the sum of squared ℓ_2-norm distances to the data.
The line corresponding to the OLS estimator has slope β_{OLS}, satisfying

$$\beta_{\text{OLS}} := \arg\min_{\beta} \sum_{i=1}^{n} (x_i[2] - \beta x_i[1])^2. \tag{12.36}$$

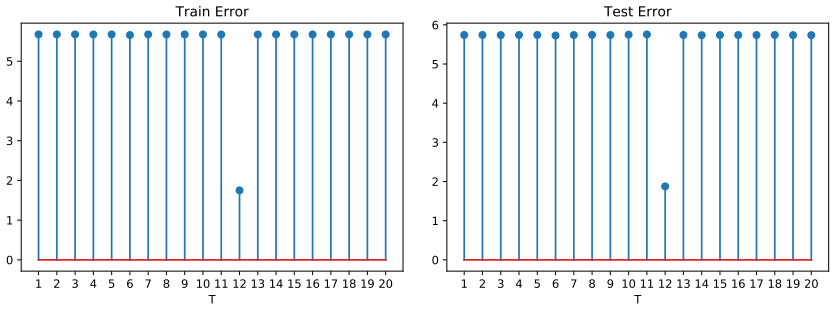
The OLS cost function can be interpreted as the sum of squared differences between the second component of the data point x_i and the second component of the point $(x_i[1], \beta x[2])$, which is on the line with slope β. On a 2D plane, if the second component is on the vertical axis, this exactly equal to the sum of squared vertical distances between the data and the line. The line obtained from linear regression minimizes this sum. In contrast PCA minimizes the sum of square Euclidean distances as explained above, so they are different.



12.3 (Global warming)

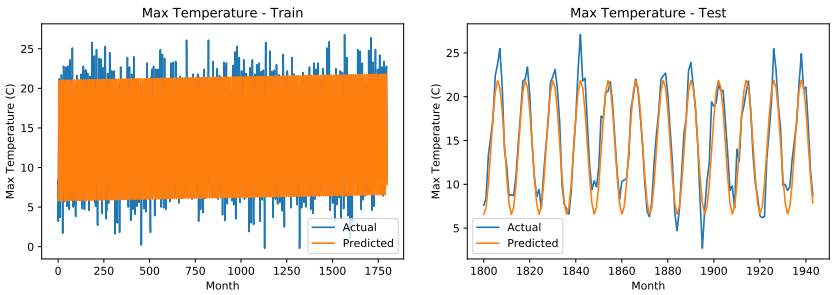
a) We have four parameters in the model and 1800 data points to fit the model. The model is very unlikely to overfit.

b) Below we show a plot of the test and train error as a function of the period T:



The best value is $T^* = 12$, which is not surprising, since the weather is cyclical with a 12 month period.

c) Here is the fit of the model to the training and test data



d) The OLS coefficients are:

Term	OLS coefficients
1	$1.34 \cdot 10^{1}$
t	$4.38 \cdot 10^{-4}$
$\cos(2\pi t/12)$	$-7.61 \cdot 10^{0}$
$\sin(2\pi t/12)$	$-4.83 \cdot 10^{-1}$

The constant term provides an offset for all temperatures. The linear term linear trends in temperature. The cosine and sine terms account for the yearly periodic pattern.

e) The coefficient of the linear term indicates the increase or decrease rate in temperatures. According to the model, the temperature in Oxford is increasing at a rate of $4.38 \cdot 10^{-4}$ per month, which translates to 0.526 degrees per century.

12.4 (OLS estimator) We denote the residual sum of squares, as a function of the vector of coefficients β and the additive constant α, by

$$\mathrm{RSS}(\beta, \alpha) := \sum_{i=1}^{n} \left(y_i - \beta^T x_i - \alpha \right)^2 . \tag{12.37}$$

We denote by $\alpha^*(\beta)$ the optimal value of α for a fixed value of β. Equivalently, $\alpha^*(\beta)$ is the best constant estimate of $y_i - \beta x_i$, $1 \leq i \leq n$, in terms of the residual sum of squares.

By Theorem 7.31,

$$\alpha_*(\beta) := \arg\min_{\alpha} \sum_{i=1}^{n} \left(y_i - \beta^T x_i - \alpha\right)^2 \tag{12.38}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta^T x_i\right) \tag{12.39}$$

$$= m(Y) - \beta^T m(X). \tag{12.40}$$

Consequently, for any β and any α $\mathrm{RSS}(\beta, \alpha) \geq \mathrm{RSS}(\beta, \alpha^*(\beta))$. To obtain β_{OLS} we fix α to $\alpha^*(\beta)$ and minimize the sum of squared errors,

$$\beta_{\mathrm{OLS}} = \arg\min_{\beta} \mathrm{RSS}(\beta, \alpha^*(\beta)). \tag{12.41}$$

Let $\mathrm{ct}\,(x_i) := x_i - m(X)$ and $y_i - m(Y)$ be the centered features and response. We have

$$\mathrm{RSS}(\beta, \alpha^*(\beta)) = \sum_{i=1}^{n} \left(y_i - \beta^T x_i - m(Y) + \beta^T m(X)\right)^2 \tag{12.42}$$

$$= \sum_{i=1}^{n} \mathrm{ct}\,(y_i)^2 + \beta^T \left(\sum_{i=1}^{n} \mathrm{ct}\,(x_i)\right) \beta - 2\beta^T \sum_{i=1}^{n} \mathrm{ct}\,(x_i)\,\mathrm{ct}\,(y_i) \tag{12.43}$$

$$= (n-1)\left(v(Y) + \beta^T \Sigma_X \beta - 2\beta^T \Sigma_{XY}\right). \tag{12.44}$$

As a function of β, the residual sum of squares is a quadratic form. Its gradient and Hessian with respect to β equal

$$\nabla_\beta \mathrm{RSS}(\beta, \alpha^*(\beta)) = 2(n-1)\Sigma_X \beta - 2\Sigma_{XY}, \tag{12.45}$$

$$\nabla^2_\beta \mathrm{RSS}(\beta, \alpha^*(\beta)) = 2(n-1)\Sigma_X. \tag{12.46}$$

Covariance matrices are positive semidefinite, because by the properties of sample covariance matrices, for any vector $a \in \mathbb{R}^d$

$$a^T \Sigma_X a = a^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \mathrm{ct}\,(x_i)\,\mathrm{ct}\,(x_i)^T\right) a \tag{12.47}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(a^T \mathrm{ct}\,(x_i)\right)^2 \geq 0. \tag{12.48}$$

Since Σ_X is invertible, there cannot be a nonzero vector such that $\Sigma_X a$ equals the zero vector, so the inequality is strict as long as a is not the zero vector. This means that the quadratic function is strictly convex and we can set its gradient to zero to find the value of β that achieves the unique minimum:

$$\beta_{\mathrm{OLS}} = \Sigma_X^{-1} \Sigma_{XY}. \tag{12.49}$$

The corresponding value of $\alpha^*(\beta)$ is

$$\alpha_{\mathrm{OLS}} = \alpha^*(\beta_{\mathrm{OLS}}) \tag{12.50}$$

$$= m(Y) - \Sigma_{XY}^T \Sigma_X^{-1} m(X). \tag{12.51}$$

12.5 (Ice cream sales) To adjust for the effect of temperature, we fit a long regression model where the response is the sales and the features are advertising and temperature. Let X denote the set of features (i.e. advertising revenue and temperature pairs), and Y the

response. We can extract the sample covariance matrix of X and the cross-covariance between X and Y directly from the sample covariance matrix of the data. The OLS coefficients equal

$$\beta_{\text{OLS}} = \Sigma_X^{-1}\Sigma_{X,Y} \tag{12.52}$$

$$= \begin{bmatrix} 900 & 540 \\ 540 & 400 \end{bmatrix}^{-1} \begin{bmatrix} 960 \\ 720 \end{bmatrix} \tag{12.53}$$

$$= \begin{bmatrix} -0.07 \\ 1.89 \end{bmatrix}. \tag{12.54}$$

Consequently, according to the model, for a fixed temperature the sales are on average *inversely proportional* to advertising! The correlation between advertising and sales is due to the confounding effect of temperature. We definitely don't agree with the marketing department; there is no evidence that advertising is beneficial.

12.6 (Linear regression and maximum likelihood estimation) To alleviate notation, we define the design matrix

$$X_{\text{train}} := \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}, \tag{12.55}$$

the observed response vector

$$y_{\text{train}} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \tag{12.56}$$

and the random vector

$$\tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \cdots \\ \tilde{y}_n \end{bmatrix}. \tag{12.57}$$

Under the modeling assumptions, the entries of $\tilde{y}$ are independent Gaussians with mean $x_i^T\beta$ and variance σ^2, so its joint pdf evaluated at the observed data equals

$$f_{\tilde{y}}(y_{\text{train}}) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - x_i^T\beta\right)^2\right) \tag{12.58}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\,||y_{\text{train}} - X_{\text{train}}\beta||_2^2\right), \tag{12.59}$$

which is equal to the likelihood, if we interpret it as a function of β. Consequently the log likelihood equals

$$\log \mathcal{L}_y(\beta) = -\frac{1}{2\sigma^2}\,||y_{\text{train}} - X_{\text{train}}\beta||_2^2 - \frac{n}{2}\log\left(2\pi\sigma^2\right), \tag{12.60}$$

so the maximum-likelihood estimator is

$$\beta_{\text{ML}} = \arg\max_{\beta} \log \mathcal{L}_y(\beta) \tag{12.61}$$

$$= \arg\max_{\beta} -\frac{1}{2\sigma^2} ||y_{\text{train}} - X_{\text{train}}\beta||_2^2 - \frac{n}{2} \log\left(2\pi\sigma^2\right) \tag{12.62}$$

$$= \arg\min_{\beta} ||y_{\text{train}} - X_{\text{train}}\beta||_2^2 , \tag{12.63}$$

which is *the same as the OLS estimator*.

12.7 (Best linear unbiased estimator) To alleviate notation, we set $\tilde{y} := \tilde{y}_{\text{train}}$, $X := X_{\text{train}}$, $\tilde{z} := \tilde{z}_{\text{train}}$ and $\beta := \beta_{\text{true}}$.

a) By linearity of expectation

$$\mathbb{E}[C\tilde{y}] = \mathbb{E}\left[C(X\beta + \tilde{z})\right] = CX\beta + C\mathbb{E}[\tilde{z}] = CX\beta. \tag{12.64}$$

b) Let $\mu := \mathbb{E}[C\tilde{y}] = CX\beta$. The covariance matrix equals

$$\Sigma_{C\tilde{y}} = \mathbb{E}\left[(C\tilde{y})(C\tilde{y})^T\right] - \mu\mu^T = C\mathbb{E}\left[\tilde{y}\tilde{y}^T\right] C^T - \mu\mu^T \tag{12.65}$$

$$= C\mathbb{E}\left[X\beta\beta^T X^T + zz^T\right] C^T - \mu\mu^T \tag{12.66}$$

$$= CX\beta\beta^T X^T C^T + C(\sigma^2 I)C^T - \mu\mu^T \tag{12.67}$$

$$= \mu\mu^T + \sigma^2 CC^T - \mu\mu^T \tag{12.68}$$

$$= \sigma^2 CC^T. \tag{12.69}$$

c) Since $\mathbb{E}[C\tilde{y}] = CX\beta$,

$$\mathbb{E}[C\tilde{y}] = CX\beta = \left((X^TX)^{-1}X^T + D\right) X\beta = (I + DX)\beta = \beta + DX\beta. \tag{12.70}$$

Thus we require $DX = 0$.

d) We have

$$\text{Var}\left[v^T C\tilde{y}\right] = v^T \Sigma_{C\tilde{y}} v \geq v^T \Sigma_{\tilde{\beta}_{\text{OLS}}} v = \text{Var}\left[v^T \tilde{\beta}_{\text{OLS}}\right] \tag{12.71}$$

for any $v \in \mathbb{R}^d$, if and only if the matrix $\Sigma_{C\tilde{y}} - \Sigma_{\tilde{\beta}_{\text{OLS}}}$ is positive semidefinite. By (12.69), since $D := C - (X^TX)^{-1}X^T$ and $DX = 0$ (which must hold if $C\tilde{y}$ is unbiased),

$$\Sigma_{C\tilde{y}} = \sigma^2 CC^T \tag{12.72}$$

$$= \sigma^2 \left((X^TX)^{-1}X^T + D\right)\left((X^TX)^{-1}X^T + D\right)^T \tag{12.73}$$

$$= \sigma^2 \left((X^TX)^{-1}X^TX(X^TX)^{-1} + DD^T + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1}\right)$$

$$= \sigma^2((X^TX)^{-1} + DD^T) \tag{12.74}$$

$$= \Sigma_{\tilde{\beta}_{\text{OLS}}} + \sigma^2 DD^T, \tag{12.75}$$

where we have used the fact that by (12.69)

$$\Sigma_{\tilde{\beta}_{\text{OLS}}} = \Sigma_{(X^TX)^{-1}X^T\tilde{y}} \tag{12.76}$$

$$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \tag{12.77}$$

$$= \sigma^2(X^TX)^{-1}. \tag{12.78}$$

DD^T is positive semidefinite since for any v $v^T DD^T v = \left|\left|D^T v\right|\right|_2^2 \geq 0$, so $\Sigma_{C\tilde{y}} - \Sigma_{\tilde{\beta}_{\text{OLS}}} = \sigma^2 DD^T$ is indeed positive semidefinite and the proof is complete.

12.8 (Augmented dataset) The ridge-regression cost function can be reformulated as the following least-squares problem:

$$\beta_{\text{RR}} := \arg\min_{\beta} \left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} \beta \right|\right|_2^2. \tag{12.79}$$

There are d additional examples. For the ith example the feature vector equals a one-hot vector where the ith entry equals λ and the rest equal zero. All the response values for the additional examples equal zero. The additional examples force the linear coefficients to fit zeros, preventing them from becoming too large.

12.9 (Correlated features)

a) Let us define $\bar{\alpha} := \sqrt{1-\alpha^2}$ to alleviate notation. We have

$$\begin{aligned}
\beta_{\text{OLS}} &= (X_{\text{train}}^T X_{\text{train}})^{-1} X_{\text{train}}^T y && (12.80)\\
&= \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1^T \\ \alpha x_1^T + \bar{\alpha} x_\perp^T \end{bmatrix} (\beta_{\text{true}} x_1 + z) && (12.81)\\
&= \frac{1}{1-\alpha^2} \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \alpha(\beta_{\text{true}} + 0.1) + 0.1\bar{\alpha} \end{bmatrix} && (12.82)\\
&= \frac{1}{1-\alpha^2} \begin{bmatrix} (1-\alpha^2)(\beta_{\text{true}} + 0.1) - 0.1\alpha\bar{\alpha} \\ 0.1\bar{\alpha} \end{bmatrix} && (12.83)\\
&= \begin{bmatrix} \beta_{\text{true}} + 0.1 - \frac{0.1\alpha}{\bar{\alpha}} \\ \frac{0.1}{\bar{\alpha}} \end{bmatrix}. && (12.84)
\end{aligned}$$

The estimated coefficients tend to

$$\lim_{\alpha \to 1} \beta_{\text{OLS}} = \begin{bmatrix} -\infty \\ \infty \end{bmatrix}. \tag{12.85}$$

The OLS estimator overfits the response vector, which causes the entries of the coefficient estimate to explode.

b) The estimate of the response equals

$$\begin{aligned}
y_{\text{OLS}} &:= X_{\text{train}} \beta_{\text{OLS}} && (12.86)\\
&= \begin{bmatrix} x_1 & \alpha x_1 + \bar{\alpha} x_\perp \end{bmatrix} \begin{bmatrix} \beta_{\text{true}} + 0.1 - \frac{0.1\alpha}{\bar{\alpha}} \\ \frac{0.1}{\bar{\alpha}} \end{bmatrix} && (12.87)\\
&= \left(\beta_{\text{true}} + 0.1 - \frac{0.1\alpha}{\bar{\alpha}}\right) x_1 + \frac{0.1}{\bar{\alpha}} (\alpha x_1 + \bar{\alpha} x_\perp) && (12.88)\\
&= (\beta_{\text{true}} + 0.1) x_1 + 0.1 x_\perp, && (12.89)
\end{aligned}$$

which does not change as $\alpha \to 1$. The estimate of the response is not collinear with x_1 even when $\alpha \to 1$, although the second feature tends to be collinear with the first one in this limit! This is because the noise in the training response vector has a component in that direction, and the linear regression estimate fits it, even if the corresponding component in the second feature vector is tiny.

c)

$$\beta_{\text{RR}} = (X_{\text{train}}^T X_{\text{train}} + \lambda I)^{-1} X_{\text{train}}^T y \tag{12.90}$$

$$= \begin{bmatrix} 1+\lambda & \alpha \\ \alpha & 1+\lambda \end{bmatrix}^{-1} \begin{bmatrix} x_1^T \\ \alpha x_1^T + \bar{\alpha} x_\perp^T \end{bmatrix} (\beta_{\text{true}} x_1 + z) \tag{12.91}$$

$$= \frac{1}{(1+\lambda)^2 - \alpha^2} \begin{bmatrix} 1+\lambda & -\alpha \\ -\alpha & 1+\lambda \end{bmatrix} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \alpha(\beta_{\text{true}} + 0.1) + 0.1\bar{\alpha} \end{bmatrix} \tag{12.92}$$

$$= \frac{1}{(1+\lambda)^2 - \alpha^2} \begin{bmatrix} (1+\lambda-\alpha^2)(\beta_{\text{true}} + 0.1) - 0.1\alpha\bar{\alpha} \\ \lambda\alpha(\beta_{\text{true}} + 0.1) + 0.1\bar{\alpha}(1+\lambda) \end{bmatrix}. \tag{12.93}$$

The estimated coefficients tend to

$$\lim_{\alpha \to 1} \beta_{\text{RR}} = \frac{1}{2\lambda + \lambda^2} \begin{bmatrix} \lambda(\beta_{\text{true}} + 0.1) \\ \lambda(\beta_{\text{true}} + 0.1) \end{bmatrix} \tag{12.94}$$

$$= \frac{1}{2+\lambda} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \beta_{\text{true}} + 0.1 \end{bmatrix}. \tag{12.95}$$

The coefficients no longer explode as in OLS. The two coefficients are the same, which makes sense because when $\alpha \to 1$ both features are equal to x_1.

d)

$$\lim_{\alpha \to 1} y_{\text{RR}} := (\lim_{\alpha \to 1} X_{\text{train}}) \lim_{\alpha \to 1} \beta_{\text{RR}} \tag{12.96}$$

$$= \begin{bmatrix} x_1 & x_1 \end{bmatrix} \frac{1}{2+\lambda} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \beta_{\text{true}} + 0.1 \end{bmatrix} \tag{12.97}$$

$$= \frac{2(\beta_{\text{true}} + 0.1) x_1}{2+\lambda}. \tag{12.98}$$

In contrast to the OLS response estimate, the ridge-regression response estimate is collinear with the true feature vector in the limit where $\alpha \to 1$.

12.10 (Prior knowledge)

a) A natural way to modify the ridge-regression cost function is to incorporate β_{prior} in the regularization term, to promote solutions that are close to it:

$$\beta_{\text{RRP}} := \arg\min_{\beta} \|y_{\text{train}} - X_{\text{train}}\beta\|_2^2 + \lambda \|\beta_{\text{prior}} - \beta\|_2^2. \tag{12.99}$$

b) The cost function can be reformulated to equal a modified least-squares problem

$$\beta_{\text{RRP}} := \arg\min_{\beta} \left\| \begin{bmatrix} \tilde{y}_{\text{train}} \\ \sqrt{\lambda}\beta_{\text{prior}} \end{bmatrix} - \begin{bmatrix} X_{\text{train}} \\ \sqrt{\lambda} I \end{bmatrix} \beta \right\|_2^2. \tag{12.100}$$

By Lemma 12.8, the solution to the modified least-squares problem equals

$$\tilde{\beta}_{\text{RRP}} = \left(\begin{bmatrix} X_{\text{train}}^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X_{\text{train}} \\ \sqrt{\lambda} I \end{bmatrix} \right)^{-1} \begin{bmatrix} X_{\text{train}}^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} \tilde{y}_{\text{train}} \\ \sqrt{\lambda}\beta_{\text{prior}} \end{bmatrix} \tag{12.101}$$

$$= \left(X_{\text{train}}^T X_{\text{train}} + \lambda I \right)^{-1} (X_{\text{train}}^T \tilde{y}_{\text{train}} + \lambda \beta_{\text{prior}}) \tag{12.102}$$

$$= \tilde{\beta}_{\text{RR}} + \left(\Sigma_X + \frac{\lambda}{n-1} I \right)^{-1} \lambda \beta_{\text{prior}}, \tag{12.103}$$

because by (12.254) in the main text, the ridge-regression coefficients equal

$$\tilde{\beta}_{\text{RR}} = \left(X_{\text{train}}^T X_{\text{train}} + \lambda I\right)^{-1} X_{\text{train}}^T \tilde{y}_{\text{train}}. \tag{12.104}$$

Consequently, the estimator is the same as the ridge-regression estimator except for a deterministic constant. As a result, the covariance matrix is unchanged and the mean equals $\mathbb{E}\left[\tilde{\beta}_{\text{RR}}\right] + \lambda \left(\Sigma_X + \frac{\lambda}{n-1} I\right)^{-1} \beta_{\text{prior}}$, which approaches β_{prior} as λ increases (since $\mathbb{E}\left[\tilde{\beta}_{\text{RR}}\right]$ approaches zero by Theorem 12.33).

12.11 (Lasso estimation with two features)

a) If $c_{\text{true,noise}} < -a_{\text{true}}$

$$\frac{x_{\text{true}}^T y_{\text{train}}}{n-1} = \frac{a_{\text{true}} x_{\text{true}}^T x_{\text{true}} + x_{\text{true}}^T z}{n-1} \tag{12.105}$$

$$= a_{\text{true}} + c_{\text{true,noise}} < 0. \tag{12.106}$$

b) The derivative of $\mathcal{L}_{\text{lasso}}(a, 0)$ under our assumptions is

$$\frac{\mathrm{d}\mathcal{L}_{\text{lasso}}(a,0)}{\mathrm{d}a} = \frac{\mathrm{d}\left(y_{\text{train}} - x_{\text{true}} a\right)^T \left(y_{\text{train}} - x_{\text{true}} a\right) + \lambda a}{\mathrm{d}a} \tag{12.107}$$

$$= -2 x_{\text{true}}^T \left(y_{\text{train}} - x_{\text{true}} a\right) + \lambda \tag{12.108}$$

$$= -2 x_{\text{true}}^T \left(x_{\text{true}}(a_{\text{true}} - a) + z\right) + \lambda \tag{12.109}$$

$$= -2 x_{\text{true}}^T \left(x_{\text{true}}(a_{\text{true}} - a) + z\right) + \lambda \tag{12.110}$$

$$= 2(n-1)\left(a - a_{\text{true}} - c_{\text{true,noise}}\right) + \lambda. \tag{12.111}$$

The second derivative equals

$$\frac{\mathrm{d}^2 \mathcal{L}_{\text{lasso}}(a,0)}{\mathrm{d}a^2} = 2(n-1), \tag{12.112}$$

which is positive. The function is therefore convex and we can find the minimum by setting the first derivative to zero. Dividing the resulting equation by $n-1$, yields

$$2(a - a_{\text{true}}) - 2c_{\text{true,noise}} + \lambda_n = 0. \tag{12.113}$$

The derivative is zero at

$$a_{\text{lasso}} := a_{\text{true}} + c_{\text{true,noise}} - \frac{\lambda_n}{2}. \tag{12.114}$$

c) Since $a_{\text{true}} \geq 0$ and $c_{\text{true,noise}} \geq -a_{\text{true}}$, we have $a_{\text{lasso}} \geq 0$ as long as

$$\lambda_n \leq 2\left(a_{\text{true}} + c_{\text{true,noise}}\right) := \lambda_{\max}. \tag{12.115}$$

d) The derivative equals

$$\frac{\mathrm{d}\mathcal{L}_{\text{lasso}}(a_{\text{lasso}}, b)}{\mathrm{d}b} \tag{12.116}$$

$$= \frac{\mathrm{d}\left(y_{\text{train}} - x_{\text{true}} a_{\text{lasso}} - x_{\text{other}} b\right)^T \left(y_{\text{train}} - x_{\text{true}} a - x_{\text{other}} b\right) + \lambda a_{\text{lasso}} + \lambda b}{\mathrm{d}b}$$

$$= -2 x_{\text{other}}^T \left(y_{\text{train}} - x_{\text{true}} a_{\text{lasso}} - x_{\text{other}} b\right) + \lambda \tag{12.117}$$

$$= -2 x_{\text{other}}^T \left(x_{\text{true}}(a_{\text{true}} - a_{\text{lasso}}) + z - x_{\text{other}} b\right) + \lambda. \tag{12.118}$$

Dividing this expression by $n-1$ and using the fact that $x_{\text{other}}^T x_{\text{other}} = n-1$ yields

$$2\rho\left(c_{\text{true,noise}} - \frac{\lambda_n}{2}\right) - 2c_{\text{other,noise}} + v(x_{\text{other}})b + \lambda_n \tag{12.119}$$

$$= (1-\rho)\lambda_n - 2c_{\text{other,noise}} + 2\rho c_{\text{true,noise}} + b, \tag{12.120}$$

which is positive for small enough b if

$$\frac{\lambda_n}{2} > \frac{c_{\text{other,noise}} - \rho\, c_{\text{true,noise}}}{1-\rho}. \tag{12.121}$$

e) By the same argument used to derive (12.120), the derivative $\frac{\mathrm{d}\mathcal{L}_{\text{lasso}}(a_{\text{lasso}},b)}{\mathrm{d}b}$ for $b<0$ is negative if

$$-(1-\rho)\lambda_n - 2c_{\text{other,noise}} + 2\rho c_{\text{true,noise}} + b < 0, \tag{12.122}$$

which is the case when the magnitude of b is small enough, as long as

$$\frac{\lambda_n}{2} > \frac{\rho\, c_{\text{true,noise}} - c_{\text{other,noise}}}{1-\rho}. \tag{12.123}$$

12.12 (Ridge regression and sparsity) Under the assumptions, the sample covariance matrix of the features equals

$$\Sigma_X = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}. \tag{12.124}$$

Since everything has zero sample mean, the sample covariance between the response and the true feature is

$$c_{\text{true,response}} = \frac{x_{\text{true}}^T y}{n-1} \tag{12.125}$$

$$= 1 + c_{\text{true,noise}} = 0.85, \tag{12.126}$$

whereas the sample covariance between the response and the spurious feature is

$$c_{\text{other,response}} = \frac{x_{\text{other}}^T y}{n-1} \tag{12.127}$$

$$= \rho + c_{\text{other,noise}} = 0.355. \tag{12.128}$$

Consequently the OLS coefficient estimate equals

$$\beta_{\text{OLS}} = \Sigma_X^{-1}\Sigma_{XY} \tag{12.129}$$

$$= \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.85 \\ 0.355 \end{bmatrix} \tag{12.130}$$

$$= \begin{bmatrix} 0.81 \\ 0.19 \end{bmatrix}. \tag{12.131}$$

The eigenvectors of the sample covariance matrix (normalized to have unit ℓ_2 norm) equal

$$u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad u_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{12.132}$$

and the corresponding eigenvalues equal $\xi_1 = 1.2$ and $\xi_2 = 0.8$. In the basis of eigenvectors, the OLS coefficients equal

$$c_{\text{OLS}} := \begin{bmatrix} u_1^T \\ u_2^T \end{bmatrix} \beta_{\text{OLS}} = \begin{bmatrix} 0.71 \\ -0.44 \end{bmatrix}, \tag{12.133}$$

so the ridge regression coefficients equal

$$\beta_{\mathrm{RR}} = \frac{c_{\mathrm{OLS}}[1]}{1+\lambda_n/\xi_1}u_1 + \frac{c_{\mathrm{OLS}}[2]}{1+\lambda_n/\xi_2}u_2 \tag{12.134}$$

$$= \frac{0.71}{1+\lambda_n/1.2}\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} + \frac{0.44}{1+\lambda_n/0.8}\frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}, \tag{12.135}$$

where $\lambda_n := \frac{\lambda}{n-1}$. The first entry only cancels out for $\lambda_n = 0.518$, so the ridge-regression coefficient estimate is never sparse.

12.13 (Derivative analysis of logistic regression)

a) By basic calculus

$$\frac{\mathrm{d}\,\mathrm{lgf}(\ell)}{\mathrm{d}\ell} = \frac{\mathrm{d}}{\mathrm{d}\ell}\frac{1}{1+\exp(-\ell)} \tag{12.136}$$

$$= \frac{\exp(-\ell)}{(1+\exp(-\ell))^2} \tag{12.137}$$

$$= \frac{\mathrm{lgf}(\ell)\exp(-\ell)}{1+\exp(-\ell)} \tag{12.138}$$

$$= \mathrm{lgf}(\ell)\left(1-\frac{1}{1+\exp(-\ell)}\right) \tag{12.139}$$

$$= \mathrm{lgf}(\ell)(1-\mathrm{lgf}(\ell)). \tag{12.140}$$

b) By the chain rule for derivatives and (12.140), since $p_{\alpha,\beta}(x) = \mathrm{lgf}(\beta^T x + \alpha)$,

$$\frac{\partial p_{\alpha,\beta}(x)}{\partial x[j]} = \beta[j]p_{\alpha,\beta}(x)(1-p_{\alpha,\beta}(x)), \quad 1 \le j \le d. \tag{12.141}$$

As the probability approaches 0 or 1, the derivative tends to zero, which indicates that the model becomes less sensitive to changes in the features.

c) By the chain rule for derivatives and (12.140), since $p_{\alpha,\beta}(x) = \mathrm{lgf}(\beta^T x + \alpha)$,

$$\frac{\partial p_{\alpha,\beta}(x)}{\partial \beta[j]}(\alpha,\beta) = x[j]p_{\alpha,\beta}(x)(1-p_{\alpha,\beta}(x)). \tag{12.142}$$

Consequently, by basic calculus

$$\frac{\partial \log \mathcal{L}_{XY}}{\partial \beta[j]}(\alpha,\beta) \tag{12.143}$$

$$= \frac{\partial}{\partial \beta[j]}\left(\sum_{\{i:y_i=0\}}\log\left(1-p_{\alpha,\beta}(x_i)\right) + \sum_{\{l:y_l=1\}}\log p_{\alpha,\beta}(x_l)\right) \tag{12.144}$$

$$= \sum_{\{i:y_i=0\}} x_i\frac{-p_{\alpha,\beta}(x_i)(1-p_{\alpha,\beta}(x_i))}{\left(1-p_{\alpha,\beta}(x_i)\right)} + \sum_{\{l:y_l=1\}} x_l\frac{p_{\alpha,\beta}(x_l[j])(1-p_{\alpha,\beta}(x_l[j]))}{p_{\alpha,\beta}(x_l)}$$

$$= \sum_{\{i:y_i=0\}} -x_i[j]p_{\alpha,\beta}(x_i) + \sum_{\{l:y_l=1\}} x_l[j](1-p_{\alpha,\beta}(x_l)) \tag{12.145}$$

$$= \sum_{m=1}^{n} x_m[j]\left(y_m - p_{\alpha,\beta}(x_m)\right), \tag{12.146}$$

where y_m is the mth label. The contribution of the mth data point to the gradient-ascent update is therefore $\eta x_m[j]\left(y_m - p_{\alpha^{(t-1)},\beta^{(t-1)}}(x_m)\right)$. It consists of the jth feature scaled by a small constant and by the error in the probability estimate for that

data point. When the label equals one, adding the contribution increases the logit $(\beta^{(t-1)})^T x_j$ for the data point, bringing the estimated probability closer to one. Conversely, when the label equals zero, it decreases the logit, bringing it closer to zero.

12.14 (Building regression trees)

a) We use induction to prove that the leaves of a regression tree form a partition of the feature space. The induction is applied on the number of leaves L. The base of induction is a tree with a single leaf, $L := 1$, which represents the whole feature space and is clearly a partition (it's a single set that is equal to the whole space).
For the induction step, we assume that for a tree with L leaves, the corresponding regions R_1, ..., R_L form a partition (their intersections are all empty and their union equals their whole space). To add a new leaf, we choose one of the regions in the partition and split it by thresholding one of the features. Let R_l denote the chosen region, j the feature used for thresholding and τ the threshold. The two new regions are $R_{l1} := R_l \cap \{x : x_j \leq \tau\}$ and $R_{l2} := R_l \cap \{x : x_j > \tau\}$. If we replace R_l by these two regions, then the resulting set of regions is still a partition. All the regions are disjoint because R_{l1} and R_{l2} are disjoint, and they cannot have any points in common with any of the other regions, because the other regions are disjoint with R_l. In addition, the union of R_{l1} and R_{l2} is R_l, so the union of all the regions still equals the whole feature space. We conclude that the $L+1$ leaves in the new tree correspond to regions of the feature space that form a partition, so the proof is complete.

b) Since $\mathcal{A}_r(s,t) \cup \mathcal{B}_r(s,t) = R_r$, we can separate the first term in the expression, as follows

$$\begin{aligned}
\triangle \mathrm{RSS}(r,s,t) = &\sum_{\{i:x_i \in R_r\}}^{n} \left(y_i - \alpha_r^{\mathrm{tree}}\right)^2 - \sum_{\{i:x_i \in \mathcal{A}_r(s,t)\}}^{n} \left(y_i - \alpha_{\mathcal{A}_r(s,t)}\right)^2 \\
&- \sum_{\{i:x_i \in \mathcal{B}_r(s,t)\}}^{n} \left(y_i - \alpha_{\mathcal{B}_r(s,t)}\right)^2 \qquad (12.147) \\
= &\sum_{\{i:x_i \in \mathcal{A}_r(s,t)\}}^{n} \left(y_i - \alpha_r^{\mathrm{tree}}\right)^2 + \sum_{\{i:x_i \in \mathcal{B}_r(s,t)\}}^{n} \left(y_i - \alpha_r^{\mathrm{tree}}\right)^2 \\
&- \sum_{\{i:x_i \in \mathcal{A}_r(s,t)\}}^{n} \left(y_i - \alpha_{\mathcal{A}_r(s,t)}\right)^2 - \sum_{\{i:x_i \in \mathcal{B}_r(s,t)\}}^{n} \left(y_i - \alpha_{\mathcal{B}_r(s,t)}\right)^2 . \qquad (12.148)
\end{aligned}$$

By 7.31

$$\alpha_{\mathcal{A}_r(s,t)} = \arg\min_c \sum_{x_i \in \mathcal{A}_r(s,t)} (x_i - c)^2 , \qquad (12.149)$$

so in particular

$$\sum_{\{i:x_i \in \mathcal{A}_r(s,t)\}}^{n} \left(y_i - \alpha_{\mathcal{A}_r(s,t)}\right)^2 \leq \sum_{\{i:x_i \in \mathcal{A}_r(s,t)\}}^{n} \left(y_i - \alpha_r^{\mathrm{tree}}\right)^2 . \qquad (12.150)$$

For the same reason,

$$\sum_{\{i:x_i \in \mathcal{B}_r(s,t)\}}^{n} \left(y_i - \alpha_{\mathcal{B}_r(s,t)}\right)^2 \leq \sum_{\{i:x_i \in \mathcal{B}_r(s,t)\}}^{n} \left(y_i - \alpha_r^{\mathrm{tree}}\right)^2 . \qquad (12.151)$$

Consequently, $\triangle \mathrm{RSS}(r,s,t) \geq 0$.

12.15 (Tree with categorical features)

Leaf 1:

The mean duration is

$$\alpha_{\text{all}} := \frac{30+20+50+15+40+40+40+30}{8} = 33.125. \tag{12.152}$$

The RSS of the single-leaf tree is

$$\begin{aligned}
\triangle \text{RSS} &:= \sum_{i=1}^{8} (y_i - \alpha_{\text{all}})^2 && (12.153)\\
&= (30-33.125)^2 + (20-33.125)^2 + (50-33.125)^2 + (15-33.125)^2 && (12.154)\\
&\quad + (40-33.125)^2 + (40-33.125)^2 + (40-33.125)^2 + (30-33.125)^2 && (12.155)\\
&= 946.9. && (12.156)
\end{aligned}$$

Leaf 2:

The mean of the durations for humans is

$$\alpha_h := \frac{20+15+40+30}{4} = 26.25. \tag{12.157}$$

The mean of the durations for animals is

$$\alpha_a := \frac{30+50+40+40}{4} = 40. \tag{12.158}$$

If we threshold according to the categorical variable, then the reduction in RSS is

$$\begin{aligned}
\triangle \text{RSS} &= \sum_{i=1}^{8} (y_i - \alpha_{\text{all}})^2 - \sum_{\{i:\text{human}\}} (y_i - \alpha_h)^2 - \sum_{\{i:\text{animal}\}} (y_i - \alpha_a)^2 && (12.159)\\
&= 946.9 - \Big((20-26.25)^2 + (15-26.25)^2 + (40-26.25)^2 + (30-26.25)^2\Big)\\
&\quad - \Big((30-40)^2 + (50-40)^2 + (40-40)^2 + (40-40)^2\Big)\\
&= 378.1. && (12.160)
\end{aligned}$$

Now we consider thresholding the number of scans using a threshold τ.
If $\tau := 1$, the mean of the durations when the number of scans is 1 or less is

$$\alpha_{\leq 1} := \frac{30+15+40}{3} = 28.33. \tag{12.161}$$

The mean of the durations when the number of scans is more than 1 is

$$\alpha_{>1} := \frac{20+50+40+40+30}{5} = 36. \tag{12.162}$$

The reduction in RSS is

$$\triangle \text{RSS} = \sum_{i=1}^{8} (y_i - \alpha_{\text{all}})^2 - \sum_{\{i:\text{duration}\leq 1\}} \left(y_i - \alpha_{\leq 1}\right)^2 - \sum_{\{i:\text{duration}>1\}} (y_i - \alpha_{>1})^2 \tag{12.163}$$

$$\begin{aligned} &= 946.9 - \left((30 - 28.33)^2 + (15 - 28.33)^2 + (40 - 28.33)^2\right) \\ &\quad - \left((20 - 36)^2 + (50 - 36)^2 + (40 - 40)^2 + (40 - 36)^2 + (30 - 36)^2\right) \end{aligned} \tag{12.164}$$

$$= 126.2. \tag{12.164}$$

If $\tau := 2$, the mean of the durations when the number of scans is 2 or less is

$$\alpha_{\leq 2} := \frac{30 + 20 + 15 + 40 + 40 + 30}{6} = 29.17. \tag{12.165}$$

The mean of the durations when the number of scans is more than 2 is

$$\alpha_{>2} := \frac{50 + 40}{2} = 45. \tag{12.166}$$

The reduction in RSS is

$$\triangle \text{RSS} = \sum_{i=1}^{8} (y_i - \alpha_{\text{all}})^2 - \sum_{\{i:\text{duration}\leq 1\}} \left(y_i - \alpha_{\leq 2}\right)^2 - \sum_{\{i:\text{duration}>1\}} (y_i - \alpha_{>2})^2 \tag{12.167}$$

$$\begin{aligned} &= 946.9 - \Big((30 - 29.17)^2 + (20 - 29.17)^2 + (15 - 29.17)^2 \\ &\quad + (40 - 29.17)^2 + (40 - 29.17)^2 + (30 - 29.17)^2\Big) \\ &\quad - \left((50 - 45)^2 + (40 - 45)^2\right) \end{aligned} \tag{12.168}$$

$$= 376.0. \tag{12.169}$$

The highest reduction in RSS is achieved by thresholding according to the categorical variable. The first leaf corresponds to humans, and the second to animals.

Leaf 3:

Now we compare the reduction in RSS achieved by thresholding the examples associated with each leaf according to the number of scans using different values of the threshold τ. The RSS for the first leaf is:

$$\text{RSS}_h = \sum_{\{i:\text{human}\}} (y_i - \alpha_h)^2 \tag{12.170}$$

$$= (20 - 26.25)^2 + (15 - 26.25)^2 + (40 - 26.25)^2 + (30 - 26.25)^2 \tag{12.171}$$

$$= 368.8 \tag{12.172}$$

If $\tau := 1$, the mean of the durations when the number of scans is 1 or less is

$$\alpha_{\leq 1} := 15. \tag{12.173}$$

The mean of the durations when the number of scans is more than 1 is

$$\alpha_{>1} := \frac{20 + 40 + 30}{3} = 30. \tag{12.174}$$

The reduction in RSS is

$$\triangle \text{RSS} = \text{RSS}_h - \sum_{\{i:\text{duration}\leq 1\}} \left(y_i - \alpha_{\leq 1}\right)^2 - \sum_{\{i:\text{duration}>1\}} \left(y_i - \alpha_{>1}\right)^2 \tag{12.175}$$

$$= 368.8 - (15-15)^2 - \left((20-30)^2 + (40-30)^2 + (30-30)^2\right)$$

$$= 168.8. \tag{12.176}$$

If $\tau := 2$, the mean of the durations when the number of scans is 2 or less is

$$\alpha_{\leq 2} := \frac{15+20+30}{3} = 21.67. \tag{12.177}$$

The mean of the durations when the number of scans is more than 2 is

$$\alpha_{>2} := 40. \tag{12.178}$$

The reduction in RSS is

$$\triangle \text{RSS} = \text{RSS}_h - \sum_{\{i:\text{duration}\leq 2\}} \left(y_i - \alpha_{\leq 2}\right)^2 - \sum_{\{i:\text{duration}>2\}} \left(y_i - \alpha_{>2}\right)^2 \tag{12.179}$$

$$= 368.8 - \left((15-21.67)^2 + (20-21.67)^2 + (30-21.67)^2\right) - (40-40)^2$$

$$= 252.1. \tag{12.180}$$

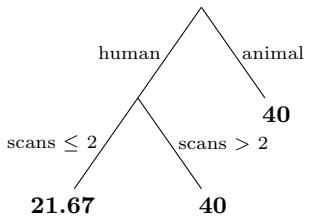
The RSS for the second leaf is:

$$\text{RSS}_a = \sum_{\{i:\text{animal}\}} \left(y_i - \alpha_h\right)^2 \tag{12.181}$$

$$= (30-40)^2 + (50-40)^2 + (40-40)^2 + (40-40)^2 \tag{12.182}$$

$$= 200. \tag{12.183}$$

We don't need to compute the reductions in RSS for the second leaf, as they cannot be greater than 252.1. Therefore, the split has to be associated with the first leaf, and the optimal threshold is $\tau := 2$.
We end up with the following regression tree, which estimates a duration of 40 minutes unless the subject is a human and the number of scans is 2 or less, in which case the estimate is 21.67 minutes.



12.16 (Early stopping in linear regression)

a)

$$\nabla \operatorname{RSS}(\beta) = \frac{1}{2(n-1)} \sum_{i=1}^{n} \nabla \left(y_i - \beta^T x_i \right)^2 \tag{12.184}$$

$$= \frac{1}{2(n-1)} \sum_{i=1}^{n} \nabla \left(y_i^2 + \beta^T x_i x_i^T \beta - 2 y_i x_i^T \beta \right) \tag{12.185}$$

$$= \frac{1}{2(n-1)} \sum_{i=1}^{n} \left(2 x_i x_i^T \beta - 2 y_i x_i \right) \tag{12.186}$$

$$= \Sigma_X \beta - \Sigma_{XY}. \tag{12.187}$$

Consequently

$$\beta^{(t)} := \beta^{(t-1)} - \eta \Sigma_X \beta^{(t-1)} + \eta \Sigma_{XY}. \tag{12.188}$$

b) If we set $\beta^{(0)} := 0$, then

$$\beta^{(1)} = \eta \Sigma_{XY}, \tag{12.189}$$

$$\beta^{(2)} = \left(I - \eta \Sigma_X \right) \eta \Sigma_{XY} + \eta \Sigma_{XY}, \tag{12.190}$$

and after t iterations

$$\beta^{(t)} = \sum_{l=0}^{t-1} \left(I - \eta \Sigma_X \right)^l \eta \Sigma_{XY}. \tag{12.191}$$

To alleviate notation we denote diagonal matrices as $\operatorname{diag}_{j=1}^{d} (a_i)$, where $a_1, \ldots, a_d$ are the entries of the diagonal. Setting $\Sigma_X = U \Lambda U^T$, where U contains the eigenvectors and Λ the eigenvalues of Σ_X, and $I = U U^T$ (which holds since U is an orthogonal matrix):

$$\beta^{(t)} = \eta \sum_{l=0}^{t-1} \left(U U^T - \eta U \Lambda U^T \right)^l \Sigma_{XY} \tag{12.192}$$

$$= \eta \sum_{l=0}^{t-1} U \left(I - \eta \Lambda \right)^l U^T \Sigma_{XY} \tag{12.193}$$

$$= \eta \sum_{l=0}^{t-1} U \operatorname{diag}_{j=1}^{d} \left(1 - \eta \xi_j \right)^l U^T \Sigma_{XY} \tag{12.194}$$

$$= \eta U \operatorname{diag}_{j=1}^{d} \left(\sum_{l=0}^{t-1} \left(1 - \eta \xi_j \right)^l \right) U^T \Sigma_{XY} \tag{12.195}$$

$$= \eta U \operatorname{diag}_{j=1}^{d} \left(\frac{1 - \left(1 - \eta \xi_j \right)^t}{\eta \xi_j} \right) U^T \Sigma_{XY}, \tag{12.196}$$

where ξ_j is the jth eigenvalue of Σ_X and the last step follows by the geometric sum

formula. Since $\beta_{\text{OLS}} = \Sigma_X^{-1}\Sigma_{XY}$ and $\Sigma_X^{-1} = U\Lambda^{-1}U^T$, we conclude that

$$\beta^{(t)} = U \operatorname{diag}_{j=1}^{d} \left(\frac{1 - \left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \Sigma_{XY} \tag{12.197}$$

$$= U\Lambda^{-1}U^T\Sigma_{XY} + U \operatorname{diag}_{j=1}^{d} \left(\frac{\left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \Sigma_{XY} \tag{12.198}$$

$$= \Sigma_X^{-1}\Sigma_{XY} + U \operatorname{diag}_{j=1}^{d} \left(\frac{\left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \Sigma_{XY} \tag{12.199}$$

$$= \beta_{\text{OLS}} + U \operatorname{diag}_{j=1}^{d} \left(\frac{\left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \Sigma_{XY}. \tag{12.200}$$

Assuming $\eta < 1/\xi_{\min} < 1/\xi_j$, $\eta\xi_j < 1$ so that $\left(1 - \eta\xi_j\right)^t \to 0$ as $t \to \infty$ for all j. Consequently $\beta^{(t)} \to \beta_{\text{OLS}}$ as $t \to \infty$.

c) Under the assumptions $\widetilde{\Sigma}_{XY} = \Sigma_X\beta_{\text{true}} + \widetilde{\Sigma}_{XZ}$ (see (12.154) in the main text), so by (12.197)

$$\tilde{\beta}^{(t)} = U \operatorname{diag}_{j=1}^{d} \left(\frac{1 - \left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \left(\Sigma_X\beta_{\text{true}} + \widetilde{\Sigma}_{XZ} \right), \tag{12.201}$$

which implies

$$\tilde{c}_{\text{GD}} = \operatorname{diag}_{j=1}^{d} \left(\frac{1 - \left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \left(\Sigma_X\beta_{\text{true}} + \widetilde{\Sigma}_{XZ} \right). \tag{12.202}$$

Since $\mathbb{E}\left[\widetilde{\Sigma}_{XZ}\right] = 0$ (see (12.167) in the main text), by linearity of expectation,

$$\mathbb{E}\left[\tilde{c}_{\text{GD}}\right] = \operatorname{diag}_{j=1}^{d} \left(\frac{1 - \left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T \Sigma_X\beta_{\text{true}} \tag{12.203}$$

$$= \operatorname{diag}_{j=1}^{d} \left(\frac{1 - \left(1 - \eta\xi_j\right)^t}{\xi_j} \right) U^T U\Lambda U^T\beta_{\text{true}} \tag{12.204}$$

$$= \operatorname{diag}_{j=1}^{d} \left(1 - \left(1 - \eta\xi_j\right)^t \right) c_{\text{true}}, \tag{12.205}$$

so that

$$\mathbb{E}\left[\tilde{c}_{\text{GD}}[j]\right] = \left(1 - \left(1 - \eta\xi_j\right)^t \right) c_{\text{true}}[j]. \tag{12.206}$$

d) By (12.202), $\mathbb{E}\left[\widetilde{\Sigma}_{XZ}\right] = 0$ and $\mathbb{E}\left[\widetilde{\Sigma}_{XZ}\widetilde{\Sigma}_{XZ}^T\right] = \frac{\sigma^2}{n-1}\Sigma_X = \frac{\sigma^2}{n-1}U\Lambda U^T$ (see (12.177)

in the main paper),

$$\Sigma_{\tilde{c}_{\mathrm{GD}}} = \mathbb{E}\left[\mathrm{ct}\left(\tilde{c}_{\mathrm{GD}}\right)\mathrm{ct}\left(\tilde{c}_{\mathrm{GD}}\right)^T\right] \tag{12.207}$$

$$= \mathbb{E}\left[\mathrm{diag}_{j=1}^{d}\left(\frac{1-\left(1-\eta\xi_j\right)^t}{\xi_j}\right)U^T\widetilde{\Sigma}_{XZ}\widetilde{\Sigma}_{XZ}^T U\,\mathrm{diag}_{j=1}^{d}\left(\frac{1-\left(1-\eta\xi_j\right)^t}{\xi_j}\right)\right]$$

$$= \mathrm{diag}_{j=1}^{d}\left(\frac{1-\left(1-\eta\xi_j\right)^t}{\xi_j}\right)U^T\mathbb{E}\left[\widetilde{\Sigma}_{XZ}\widetilde{\Sigma}_{XZ}^T\right]U\,\mathrm{diag}_{j=1}^{d}\left(\frac{1-\left(1-\eta\xi_j\right)^t}{\xi_j}\right)$$

$$= \frac{\sigma^2}{n-1}\,\mathrm{diag}_{j=1}^{d}\left(\frac{1-\left(1-\eta\xi_j\right)^t}{\xi_j}\right)U^T U\Lambda U^T U\,\mathrm{diag}_{j=1}^{d}\left(\frac{1-\left(1-\eta\xi_j\right)^t}{\xi_j}\right)$$

$$= \frac{\sigma^2}{n-1}\,\mathrm{diag}_{j=1}^{d}\left(\frac{\left(1-\left(1-\eta\xi_j\right)^t\right)^2}{\xi_j}\right). \tag{12.208}$$

We conclude that

$$\mathrm{Var}\left[\tilde{c}_{\mathrm{GD}}[j]\right] = \frac{\sigma^2}{n-1}\left(\frac{\left(1-\left(1-\eta\xi_j\right)^t\right)^2}{\xi_j}\right). \tag{12.209}$$

e) We pick t small enough to ensure that for the small eigenvalue

$$\left(1-\eta\xi_{\mathrm{min}}\right)^t \approx 1, \tag{12.210}$$

but for the large eigenvalues ξ_{large}

$$\left(1-\eta\xi_{\mathrm{large}}\right)^t \ll 1, \tag{12.211}$$

which is possible because $\xi_{\mathrm{large}} \gg \xi_{\mathrm{min}}$. For the index j_{min} corresponding to the small eigenvalue, by (12.206) and (12.209),

$$\mathbb{E}\left[\tilde{c}_{\mathrm{GD}}[j_{\mathrm{min}}]\right] \approx 0, \tag{12.212}$$

$$\mathrm{Var}\left[\tilde{c}_{\mathrm{GD}}[j_{\mathrm{min}}]\right] \approx 0, \tag{12.213}$$

which introduces some bias, but avoids noise amplification due to the small eigenvalue. For any of the remaining indices, which we denote by j_{large},

$$\mathbb{E}\left[\tilde{c}_{\mathrm{GD}}[j_{\mathrm{large}}]\right] \approx c_{\mathrm{true}}[j], \tag{12.214}$$

$$\mathrm{Var}\left[\tilde{c}_{\mathrm{GD}}[j_{\mathrm{large}}]\right] \approx \frac{\sigma^2}{(n-1)\xi_{\mathrm{large}}}, \tag{12.215}$$

which does not produce noise amplification, because ξ_{large} is large.

12.17 (Product)

a) The mean of the features is equal to zero, because they equal 1 or -1 with the same probability. By the independence assumption, the mean of the response equals

$$\mathbb{E}\left[\tilde{y}\right] = \sum_{x[1]\in\{-1,1\}}\sum_{x[2]\in\{-1,1\}} x[1]x[2]p_{\tilde{x}}(x[1],x[2]) \tag{12.216}$$

$$= p_{\tilde{x}}(-1,-1) - p_{\tilde{x}}(-1,1) - p_{\tilde{x}}(1,-1) + p_{\tilde{x}}(1,1) \tag{12.217}$$

$$= p_{\tilde{x}[1]}(-1)p_{\tilde{x}[2]}(-1) - p_{\tilde{x}[1]}(-1)p_{\tilde{x}[2]}(1) - p_{\tilde{x}[1]}(1)p_{\tilde{x}[2]}(-1) + p_{\tilde{x}[1]}(1)p_{\tilde{x}[2]}(1)$$

$$= 0. \tag{12.218}$$

The covariance between the response and each of the features equals

$$\begin{aligned}\text{Cov}\left[\tilde{x}[1], \tilde{y}\right] &= \mathbb{E}\left[\tilde{x}[1]\tilde{y}\right] && (12.219)\\ &= \mathbb{E}\left[\tilde{x}[1]^2\tilde{x}[2]\right] && (12.220)\\ &= \mathbb{E}\left[\tilde{x}[2]\right] && (12.221)\\ &= 0, && (12.222)\end{aligned}$$

because $\tilde{x}[1]^2 = 1$ with probability one. By the same reasoning, $\tilde{x}[2]$ and $\tilde{y}$ are also uncorrelated. Consequently, the linear MMSE estimator is

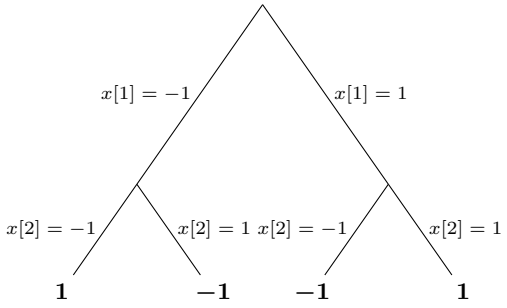
$$\ell_{\text{MMSE}}(\tilde{x}) = \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \tilde{x} = 0, \tag{12.223}$$

for all values of the features. The corresponding MSE is

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{y} - \ell_{\text{MMSE}}(\tilde{x})\right)^2\right] &= \mathbb{E}\left[\tilde{y}^2\right] && (12.224)\\ &= \mathbb{E}\left[\tilde{x}[1]^2\tilde{x}[2]^2\right] && (12.225)\\ &= 1. && (12.226)\end{aligned}$$

b) The minimum MSE estimator of $\tilde{y}$ given $\tilde{x}[1] = x_1$ and $\tilde{x}[2] = x_2$, is the conditional mean function which equals $x_1 x_2$ and is exactly equal to $\tilde{y}$. Therefore the corresponding MSE is zero. The estimator is perfect!

c) The regression tree is shown below.



d) There are many different neural networks that implement the minimum MSE estimator. A possibility is to have a first linear layer with four hidden variables that equal

$$\begin{aligned}h[1] &:= \frac{x[1] + x[2]}{2}, && (12.227)\\ h[2] &:= -\frac{x[1] + x[2]}{2}, && (12.228)\\ h[3] &:= \frac{x[1] - x[2]}{2}, && (12.229)\\ h[4] &:= \frac{-x[1] + x[2]}{2}, && (12.230)\end{aligned}$$

followed by a ReLU to produce the rectified hidden variables h_+. Due to the ReLU, if $x[1] = 1$ and $x[2] = 1$, then $h_+[1] = 1$ and the rest of rectified hidden variables are all zero. If $x[1] = -1$ and $x[2] = -1$, $h_+[2] = 1$ and the rest are zero. If $x[1] = 1$ and $x[2] = -1$, $h_+[3] = 1$ and the rest are zero. If $x[1] = -1$ and $x[2] = 1$, $h_+[4] = 1$ and

the rest are zero. Consequently, we can implement the minimum MSE estimator by weighting the rectified hidden variables as follows in the last layer:

$$n\ell_{\text{nnet}}\left(x\right) := h_+[1] + h_+[2] - h_+[3] - h_+[4]. \tag{12.231}$$

12.18 (Diagnosing cancer)

a) The model is probably learning features that enable it to identify each patient and memorize its corresponding label, but have nothing to do with cancer recurrence. This explains the perfect accuracy on the test set, which includes images from the same patients as the training set, and its bad performance on data from new patients.

b) Laura can apply the model setting the additional input to *treatment*, and also to *no treatment*. If *treatment* leads to a lower predicted probability of recurrence than *no treatment*, then she concludes that the patient would benefit from the treatment. However, this only works if the treatment in the training set is randomized. Otherwise, the conditional probabilities estimated by the model cannot be interpreted causally, because the treatment and the potential outcomes associated with recurrence are not necessarily independent (there could be confounders).

12.19 (ROC curve of random classifier)

a) The probability that we select each of the P positive examples is $p_{\tilde{x}_+}(x_i) = 1/\,\mathrm{P}$ for any positive example x_i, so

$$\begin{aligned} \mathrm{P}\left(\tilde{s}_+ > \tau\right) &= \mathrm{P}\left(p_{\text{pred}}(\tilde{x}_+) > \tau\right) && (12.232)\\ &= \sum_{\{i:y_i=+,p_{\text{pred}}(x_i)>\tau\}} p_{\tilde{x}_+}(x_i) && (12.233)\\ &= \frac{\mathrm{TP}}{\mathrm{P}} = \mathrm{TPR}(\tau), && (12.234) \end{aligned}$$

since the positive examples classified as positive are exactly the true positives (TP). Similarly, since $p_{\tilde{x}_-}(x_i) = 1/\,\mathrm{N}$ for any negative example x_i,

$$\begin{aligned} \mathrm{P}\left(\tilde{s}_- > \tau\right) &= \mathrm{P}\left(p_{\text{pred}}(\tilde{x}_-) > \tau\right) && (12.235)\\ &= \sum_{\{i:y_i=-,p_{\text{pred}}(x_i)>\tau\}} p_{\tilde{x}_-}(x_i) && (12.236)\\ &= \frac{\mathrm{FP}}{\mathrm{N}} = \mathrm{FPR}(\tau). && (12.237) \end{aligned}$$

b) By definition of the random classifier, the predicted probabilities are independent samples from a uniform distribution between 0 and 1. Consequently, for a large number of examples, the histograms of the probabilities associated with positive and negative labels will both approximate a uniform distribution between 0 and 1. In that case, $\tilde{s}_+ := p_{\text{pred}}(\tilde{x}_+)$ and $\tilde{s}_- := p_{\text{pred}}(\tilde{x}_-)$ are both approximately uniformly distributed between 0 and 1. As a result, by (12.234)

$$\begin{aligned} \mathrm{TPR}(\tau) &= \mathrm{P}\left(\tilde{s}_+ > \tau\right) && (12.238)\\ &\approx \int_\tau^1 \mathrm{d}u && (12.239)\\ &= 1 - \tau, && (12.240) \end{aligned}$$

and by (12.237) $\mathrm{FPR}(\tau) \approx 1 - \tau$. We conclude that $\mathrm{TPR}(\tau) = \mathrm{FPR}(\tau)$ for all τ, so that the ROC curve of the classifier approximates a diagonal line from the origin ($\tau := 1$) to $(1, 1)$ ($\tau := 0$).

12.20 (Probabilistic interpretation of AUC)

a) The area under the ROC curve can be computed by integrating the TPR as a function of the FPR. To compute the integral, we express both quantities as a function of the threshold τ that we apply to the predicted probability. For $\tau = 1$, FPR = 0 and TPR = 0. Then as τ decreases, the two quantities never decrease and eventually reach FPR = 1 and TPR = 1. Consequently, we can express the area under the ROC curve as

$$\text{AUC} = \int_{\tau=1}^{0} \text{TPR}(\tau) \, \text{dFPR}(\tau). \tag{12.241}$$

By (12.237) and basic calculus, the differential of FPR equals

$$\begin{aligned} \text{dFPR}(\tau) &= \frac{\text{d}\,\text{FPR}(\tau)}{\text{d}\tau} \, \text{d}\tau & (12.242) \\ &= \frac{\text{d}\left(1 - F_{\tilde{s}_-}(\tau)\right)}{\text{d}\tau} \, \text{d}\tau & (12.243) \\ &= -f_{\tilde{s}_-}(\tau) \, \text{d}\tau. & (12.244) \end{aligned}$$

Combining this with (12.234) and performing the change of variable $\tau = -t$, we obtain

$$\text{AUC} = \int_{t=0}^{1} (1 - F_{\tilde{s}_+}(t)) f_{\tilde{s}_-}(t) \, \text{d}t. \tag{12.245}$$

b) Since $\tilde{x}_+$ and $\tilde{x}_-$ are independent, then so are $\tilde{s}_+$ and $\tilde{s}_-$, so

$$\begin{aligned} \text{P}\left(\tilde{s}_+ > \tilde{s}_-\right) &= \int_{t=0}^{1} \int_{a=t}^{1} f_{\tilde{s}_+, \tilde{s}_-}(a, t) \, \text{d}t \, \text{d}a & (12.246) \\ &= \int_{t=0}^{1} \int_{a=t}^{1} f_{\tilde{s}_+}(a) f_{\tilde{s}_-}(t) \, \text{d}t \, \text{d}a & (12.247) \\ &= \int_{t=0}^{1} (1 - F_{\tilde{s}_+}(t)) f_{\tilde{s}_-}(t) \, \text{d}t & (12.248) \\ &= \text{AUC}, & (12.249) \end{aligned}$$

where the last step follows from (12.245).

c) For the random classifier, the ROC is a diagonal line from $(0, 0)$ to (1,1), so the area under the curve is 1/2. Since $\tilde{s}_+ := p_{\text{pred}}(\tilde{x}_+)$ and $\tilde{s}_- := p_{\text{pred}}(\tilde{x}_-)$ are both approximately uniformly distributed between 0 and 1, the probability that the score for the positive example is larger than that of the negative example, is 1/2,

$$\begin{aligned} \text{P}\left(\tilde{s}_+ > \tilde{s}_-\right) &= \int_{t=0}^{1} \int_{a=t}^{1} \text{d}t \, \text{d}a & (12.250) \\ &= \int_{t=0}^{1} (1 - t) \, \text{d}t & (12.251) \\ &= \frac{1}{2}. & (12.252) \end{aligned}$$

d) We can approximate $\text{P}\left(\tilde{s}_+ > \tilde{s}_-\right)$ via the Monte Carlo method, by randomly sampling a large number of pairs, each consisting of a positive and a negative example, from the data and then computing the fraction of them for which the score associated with the positive example is larger. For the example considered in Section 12.9, using 10^5 pairs produces an estimate equal to 0.8470, which is very close to the value of the AUC obtained by computing the area under the ROC curve (0.8472).

12.21 (F1 score)

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{TPR} \cdot \mathrm{Precision}}{\mathrm{TPR} + \mathrm{Precision}} \tag{12.253}$$

$$= \frac{1}{\frac{1}{2} \cdot \frac{1}{\mathrm{TPR}} + \frac{1}{2} \cdot \frac{1}{\mathrm{Precision}}} \tag{12.254}$$

$$= \frac{1}{\frac{1}{2} \cdot \frac{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}} + \frac{1}{2} \cdot \frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}}} \tag{12.255}$$

$$= \frac{\mathrm{TP}}{\mathrm{TP} + \frac{\mathrm{FN}}{2} + \frac{\mathrm{FP}}{2}} \tag{12.256}$$

$$= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{AFE}}. \tag{12.257}$$

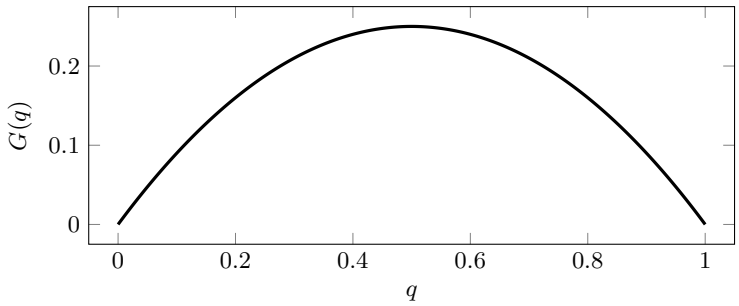
12.22 (Decomposition of the Brier score)

a) As explained in Section 12.9.2, we evaluate calibration by dividing the unit interval in bins, assigning the probability estimates to each bin, and then comparing the estimated probabilities in each bin to the empirical probability (fraction of data with label 1) in each bin. This is exactly what the calibration component does, it sums the squared differences between the estimated probabilities (p_b) and the empirical probabilities (q_b) weighted by the fraction of the total data that belongs to each bin (n_b/n). Consequently, a smaller calibration component corresponds to a better calibrated classifier.

b) The graph below shows the Gini index

$$G(q) := q(1-q) \tag{12.258}$$

for $q \in [0, 1]$.



Similar to the entropy, the Gini index is highest when q and $1-q$ are the same, and lowest when one equals zero and the other equals one. The discrimination component in the Brier score is equal to the Gini index of the empirical probability of one in each bin. The component is small if the empirical probability in each bin is either close to zero or close to one, which means that the classifier is able to separate the data according to the labels. Consequently, a smaller discrimination component corresponds to a more discriminative classifier.

c) We begin by reformulating the Brier score in terms of the bins:

$$\text{Brier score} := \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i)^2 \tag{12.259}$$

$$= \frac{1}{n} \sum_{b=1}^{m} \sum_{\{p_i \in \mathcal{B}_b\}} (y_i - p_i)^2 . \tag{12.260}$$

By the assumption that the probability estimate in the bth bin is approximately equal to p_b,

$$\sum_{\{p_i \in \mathcal{B}_b\}} (y_i - p_i)^2 \tag{12.261}$$

$$\approx \sum_{\{p_i \in \mathcal{B}_b\}} (y_i - p_b)^2 \tag{12.262}$$

$$= \sum_{\{p_i \in \mathcal{B}_b\}} (y_i - p_b + q_b - q_b)^2 \tag{12.263}$$

$$= \sum_{\{p_i \in \mathcal{B}_b\}} (q_b - p_b)^2 + \sum_{\{p_i \in \mathcal{B}_b\}} (y_i - q_b)^2 + 2\,(q_b - p_b) \sum_{\{p_i \in \mathcal{B}_b\}} (y_i - q_b)\,. \tag{12.264}$$

The first term in (12.264) is equal to the calibration component,

$$\sum_{\{p_i \in \mathcal{B}_b\}} (q_b - p_b)^2 = n_b\,(q_b - p_b)^2\,. \tag{12.265}$$

Since the labels are equal to zero or one, $y_i^2 = y_i$ for $1 \leq i \leq n$, so

$$\frac{1}{n_b} \sum_{\{p_i \in \mathcal{B}_b\}} y_i^2 = \frac{1}{n_b} \sum_{\{p_i \in \mathcal{B}_b\}} y_i \tag{12.266}$$

$$:= q_b. \tag{12.267}$$

Consequently, the second term in (12.264) is equal to the discrimination component,

$$\sum_{\{p_i \in \mathcal{B}_b\}} (y_i - q_b)^2 = \sum_{\{p_i \in \mathcal{B}_b\}} y_i^2 - 2q_b \sum_{\{p_i \in \mathcal{B}_b\}} y_i + n_b q_b^2 \tag{12.268}$$

$$= \sum_{\{p_i \in \mathcal{B}_b\}} \left(y_i^2 - 2q_b y_i + q_b^2\right) \tag{12.269}$$

$$= n_b q_b - 2n_b q_b^2 + n_b q_b^2 \tag{12.270}$$

$$= n_b q_b (1 - q_b). \tag{12.271}$$

Finally, the third term in (12.264) is zero by the definition of q_b,

$$\sum_{\{p_i \in \mathcal{B}_b\}} (y_i - q_b) = n_b q_b - n_b q_b = 0. \tag{12.272}$$

d) The perfectly discriminative classifier only produces two different probability estimates, so we can divide them into two bins ($m = 2$). Bin 1 contains the probability estimates equal to 0.8, all of which have label zero, and bin 2 the estimates equal to 0.9, all of which have label one. Consequently, the empirical probabilities in each bin are zero and one, respectively:

$$q_1 := \frac{1}{n_1} \sum_{\{p_i \in \mathcal{B}_1\}} y_i = 0, \tag{12.273}$$

$$q_2 := \frac{1}{n_2} \sum_{\{p_i \in \mathcal{B}_2\}} y_i = \frac{n_2}{n_2} = 1. \tag{12.274}$$

As a result, the discrimination term is equal to zero. This means that the calibration term must equal the Brier score (0.504).

e) The perfectly calibrated classifier assigns every data point to the same estimate, so we only need a single bin ($m = 1$). The empirical probability q_1 of label 1 in the bin is equal to the fraction of ones by definition, which is equal to the probability estimate by assumption. Consequently, $p_1 = q_1$ and the calibration component is zero. This means that the discrimination component must equal 0.169.