STAT 511
PROBABILITY

Fall 2020

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

2 Probability

2.1 Introduction

Terminology: Probability is a measure of one's belief in the occurrence of a future event. Probability is called "the mathematics of uncertainty."

Examples: Here are some events we might want to assign a probability to (i.e., to quantify the likelihood of occurrence):

- tomorrow's temperature exceeding 80 degrees
- getting a flat tire on my way home today
- a new policy holder making a claim in the next year
- you being diagnosed with prostate/cervical cancer in the next 20 years
- a new patient developing an addiction to opioids
- President Trump winning re-election in 2020.

Approaches: How do we assign probabilities to events like these and other events?

1. Subjective approach
   - based on prior experience, subject-matter knowledge, feeling, etc.
   - may not be scientific
2. Relative frequency approach
   - requires the ability to observe the occurrence of an event (and its non-occurrence) repeatedly under identical conditions
   - can be carried out using simulation (see Examples 2.1, 2.2, and 2.3)
3. Axiomatic/Model-based approach
   - grounded in set theory/mathematics
   - we will take this approach

Example 2.1. We illustrate how the relative frequency approach works using simulation. Suppose we flip a coin and observe the outcome; the sample space is

$$S = \{\text{H}, \text{T}\}.$$

Let $A = \{\text{T}\}$, the event that a "tail" is observed. How might we assign probability to the event A? Suppose we flip the same coin over and over again and record the fraction of

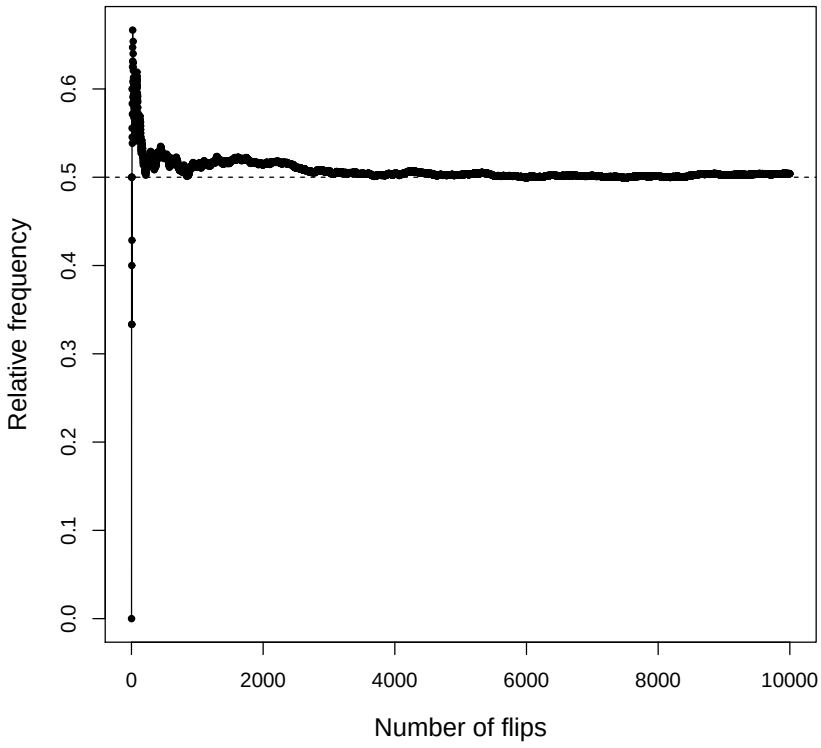

Figure 2.1: Relative frequency of coin flips resulting in "tails" plotted over 10000 flips. A dotted horizontal line at 0.5 has been added.

times A occurs. This fraction is called the **relative frequency**. Specifically, if we flip the coin n times and let $n(A)$ denote the number of times A occurs, then the relative frequency approach to probability says

$$P(A) \approx \frac{n(A)}{n}.$$

The symbol $P(A)$ is shorthand for "the probability that A occurs."

Illustration: I used R to simulate this experiment $n = 10000$ times while assuming the coin is fair; i.e., flipping the same fair coin 10000 times. Figure 2.1 plots the relative frequencies over the 10000 flips. The number of tails observed in this simulation was 5040.

```
> sum(flip)
[1] 5040
```

Therefore, we would assign

$$P(A) = 0.5040$$

on the basis of this simulation. If we repeated the simulation, we would get different answers most likely. In fact, I did this 5 more times and got 5023, 5033, 5016, 5061, and 4976.

Remark: In general, the relative frequency approach to probability says that $n(A)/n$ will "stabilize" around $P(A)$ as n increases. Mathematically,

$$\lim_{n\to\infty} \frac{n(A)}{n} = P(A).$$

Interesting: John Edmund Kerrich (a British mathematician) performed a similar experiment. He flipped an actual fair coin $n = 10000$ times while in an internment camp in Nazi-occupied Denmark in the 1940's (he did not have R!). He observed 5067 heads out of 10000 flips, offering empirical evidence of why the relative frequency approach "works" (as we have just done). □

Example 2.2. *The matching problem.* Suppose M men are at a party, and each man is wearing a hat. Each man throws his hat into the center of the room. Each man then selects a hat at random. What is the probability at least one man selects his own hat; i.e., there is at least one "match"? Define

$$A = \{\text{at least one man selects his own hat}\}.$$

Let's use simulation to estimate $P(A)$ like we did in Example 2.1.

Illustration: I used R to perform the "hat matching" experiment $n = 10000$ times while assuming the party consisted of $M = 10$ men. The event A occurred in 6364 of the simulated parties:

```
> sum(event)
[1] 6364
```

Therefore, we would assign

$$P(A) = 0.6364$$

on the basis of this simulation. The plot of relative frequencies is shown in Figure 2.2 (next page).

Curiosity: What happens if we grow the size of the party? I performed the same simulation with $M = 100$, $M = 1000$, and $M = 10000$ men and obtained the following results:

M	$n(A)$	n	$P(A) = n(A)/n$
10	6364	10000	0.6364
100	6342	10000	0.6342
1000	6300	10000	0.6300
10000	6351	10000	0.6351

Interesting: In general, for a party with M men, the probability of at least one match is

$$P(A) = 1 - \sum_{k=0}^{M} \frac{(-1)^k}{k!}.$$

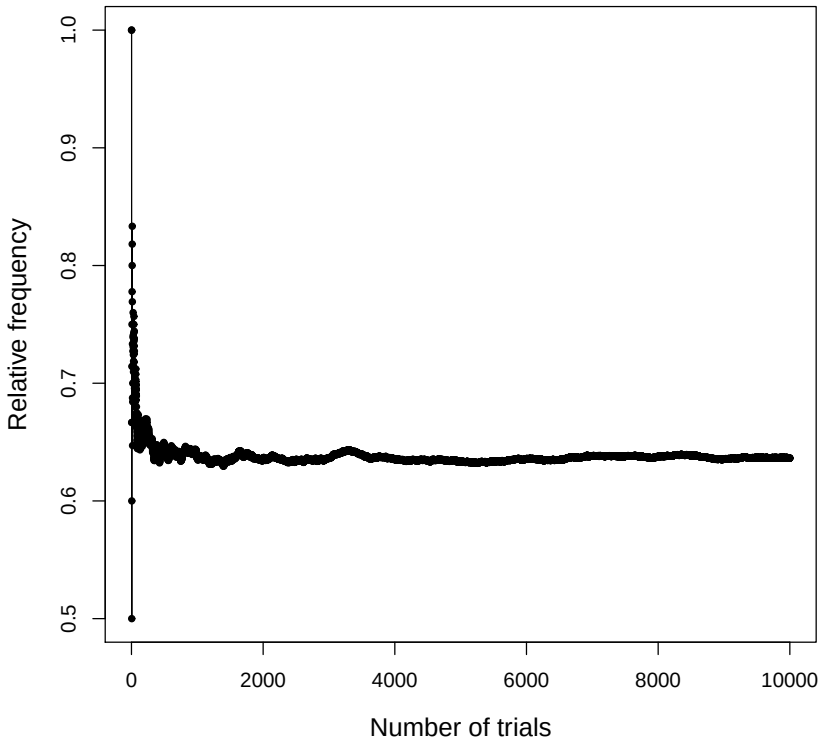

Figure 2.2: The matching problem with $M = 10$ men. Relative frequencies plotted over 10000 trials.

Letting the party grow large without bound is equivalent to letting $M \to \infty$. From calculus,

$$\lim_{M\to\infty}\left[1 - \sum_{k=0}^{M}\frac{(-1)^k}{k!}\right] = 1 - \sum_{k=0}^{\infty}\frac{(-1)^k}{k!} = 1 - e^{-1} \approx 0.6321.$$

Remark: This is an example where intuition usually leads one astray. Some students would reason that as the number of men M increases, the chance of an individual match $(1/M)$ decreases to 0 so $P(A)$ will also approach 0. Other students would reason that because M is large, "hat matching" overall becomes easier so $P(A)$ should approach 1. Neither argument is correct and, in fact, the correct answer lies somewhere in the middle. □

Example 2.3. *The birthday problem.* In a class of M students, what is the probability there will be at least one shared birthday? Define

$$A = \{\text{at least one shared birthday}\}.$$

Let's use simulation to estimate $P(A)$. To make this example concrete, assume that there are 365 days in a year and that there are no siblings (e.g., twins, triplets, etc.) in the class. On July 1, 2018, there were $M = 50$ students enrolled in this class, so we will use this.

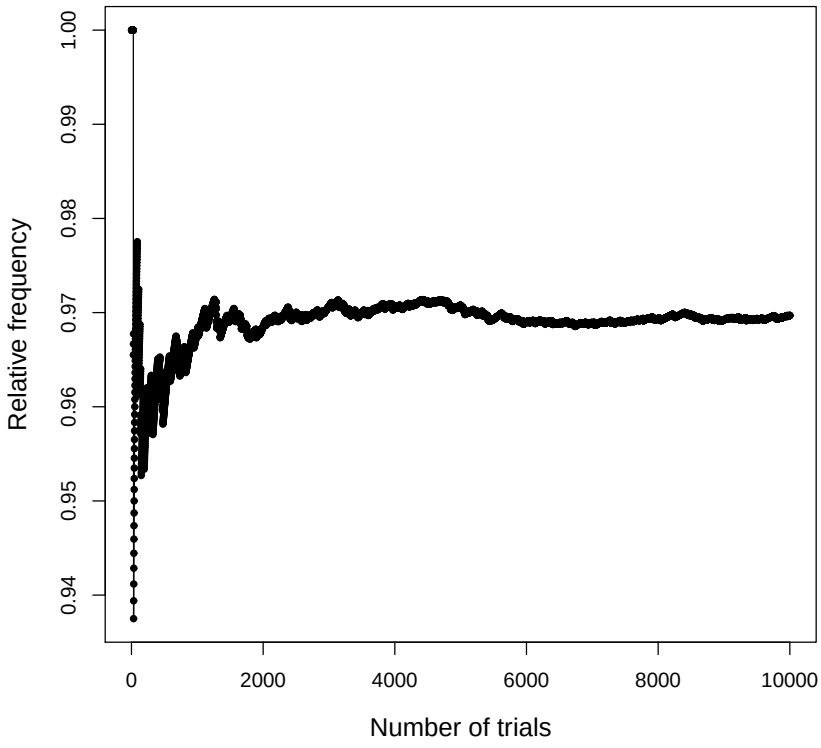

Figure 2.3: The birthday problem with $M = 50$ students. Relative frequencies plotted over 10000 trials.

Illustration: I used R to perform the "shared birthday" experiment $n = 10000$ times while assuming the class consists of $M = 50$ students. The event A occurred in 9697 of the simulated parties:

```
> sum(event)
[1] 9697
```

Therefore, we would assign

$$P(A) = 0.9697$$

on the basis of this simulation. The plot of relative frequencies is shown in Figure 2.3.

Interesting: In general, for a class with M students, the correct answer is

$$P(A) = 1 - \frac{M!\binom{365}{M}}{365^M}.$$

When $M = 50$, this probability is 0.9704 (to 4 dp), so our simulation was accurate. Interestingly, M need only be 23 for $P(A)$ to exceed 1/2. Intuition might suggest that you would need many more students than this. □

2.2 Sample spaces and sets

Terminology: A **random experiment** is an experiment that produces outcomes which are not predictable with certainty in advance. The **sample space** S for a random experiment is the set of all possible outcomes.

Example 2.4. Consider the following random experiments and their associated sample spaces. Let ω denote a generic outcome.

(a) Observe the high temperature for today:

$$S = \{\omega : -\infty < \omega < \infty\} = \mathbb{R}$$

(b) Record the number of planes landing at CAE:

$$S = \{\omega : \omega = 0, 1, 2, ..., \} = \mathbb{Z}^+$$

(c) Toss a coin three times:

$$S = \{\text{(HHH), (HHT), (HTH), (THH), (HTT), (THT), (TTH), (TTT)}\}$$

(d) Measure the length of a female subject's largest uterine fibroid:

$$S = \{\omega : \omega \geq 0\} = \mathbb{R}^+ \ \square$$

Definitions: We say that a set (e.g., A, B, S, etc.) is **countable** if its elements can be put into a 1:1 correspondence with the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, ..., \}.$$

If a set is not countable, we say it is **uncountable**. In Example 2.4,

(a) $S = \mathbb{R}$ is uncountable

(b) $S = \mathbb{Z}^+$ is countable (i.e., countably infinite); $|S| = +\infty$

(c) $S = \{\text{(HHH), (HHT), ..., (TTT)}\}$ is countable (i.e., countably finite); $|S| = 8$

(d) $S = \mathbb{R}^+$ is uncountable

Note: Any **finite set** is countable. By "finite," we mean that $|A| < \infty$, that is, "the process of counting the elements in A comes to an end." An infinite set A can be countable or uncountable. By "infinite," we mean that $|A| = +\infty$. For example,

- $\mathbb{N} = \{1, 2, 3, ..., \}$ is countably infinite
- $A = \{\omega : 0 < \omega < 1\}$ is uncountable.

Definitions: Suppose that S is a sample space for a random experiment. An **event** A is a subset of S, that is, $A \subseteq S$. Suppose the experiment produces the outcome ω.

- If $\omega \in A$, we say that "A occurs"
- If $\omega \notin A$, we say that "A does not occur."

The set A is a **subset** of B if

$$\omega \in A \Longrightarrow \omega \in B.$$

This is written $A \subset B$ or $A \subseteq B$. In a random experiment, if the event A occurs, then so does B. The converse is not necessarily true.

Two sets A and B are **equal** if each set is a subset of the other, that is,

$$A = B \Longleftrightarrow A \subseteq B \text{ and } B \subseteq A.$$

In probability, set equality is important. If two events A and B are the same (i.e., $A = B$), then they have the same probability.

The **null set**, denoted by $\emptyset$, is the set that contains no outcomes. Intuitively, it makes sense to assign zero probability to this "event."

Set Operations: Suppose A and B are subsets of S.

- Union: $A \cup B = \{\omega \in S : \omega \in A \text{ or } \omega \in B\}$. This is the set of all outcomes in A or in B (or in both).
- Intersection: $A \cap B = \{\omega \in S : \omega \in A \text{ and } \omega \in B\}$. This is the set of all outcomes in A and B.
- Complementation: $\overline{A} = \{\omega \in S : \omega \notin A\}$. This is the set of all outcomes not in A.

Example 2.5. A medical professional observes adult male patients entering an emergency room. She classifies each patient according to his blood type (AB^+, AB^-, A^+, A^-, B^+, B^-, O^+, and O^-) and whether his systolic blood pressure (SBP) is low (L), normal (N), or high (H). Consider the observation of the next male patient as a random experiment.

The sample space is

$$\begin{aligned} S = \{&(AB^+, L), (AB^-, L), (A^+, L), (A^-, L), (B^+, L), (B^-, L), (O^+, L), (O^-, L), \\ &(AB^+, N), (AB^-, N), (A^+, N), (A^-, N), (B^+, N), (B^-, N), (O^+, N), (O^-, N), \\ &(AB^+, H), (AB^-, H), (A^+, H), (A^-, H), (B^+, H), (B^-, H), (O^+, H), (O^-, H)\}. \end{aligned}$$

Note that this sample space contains $|S| = 24$ outcomes.

This example illustrates many concepts we will discuss in due course:

- There are 8 different blood types. There are 3 different categorizations of SBP. There are $8 \times 3 = 24$ possible outcomes in the sample space, which is formed by combining the two factors. The authors call this "the mn rule."
- Because S is countable, the authors call this a **discrete sample space**.
- Are these 24 outcomes equally likely? Probably not. O^+ is by far the most common blood type among American males (about 38 percent). On the other hand, AB^- is rare (only about 1 percent). Similarly, most American males have either normal or high SBP; much fewer have low SBP.
- Even though we have listed all possible outcomes in S, we have not specified probabilities associated with the outcomes. We cannot assign probability to events like

$$\begin{aligned} A &= \{\text{blood type with a } ^+ \text{ rhesus status}\} \\ B &= \{\text{high SBP}\} \end{aligned}$$

  without having this information.

Exercise: List the outcomes in $A \cup B$, $A \cap B$, and $\overline{A}$.

$$\begin{aligned} A \cup B &= \{\text{outcomes with a } ^+ \text{ rhesus status } \textbf{or} \text{ high SBP}\} \\ &= \{(\text{AB}^+, \text{L}), (\text{A}^+, \text{L}), (\text{B}^+, \text{L}), (\text{O}^+, \text{L}), (\text{AB}^+, \text{N}), (\text{A}^+, \text{N}), (\text{B}^+, \text{N}), (\text{O}^+, \text{N}), \\ &\quad (\text{AB}^+, \text{H}), (\text{AB}^-, \text{H}), (\text{A}^+, \text{H}), (\text{A}^-, \text{H}), (\text{B}^+, \text{H}), (\text{B}^-, \text{H}), (\text{O}^+, \text{H}), (\text{O}^-, \text{H})\} \end{aligned}$$

$$\begin{aligned} A \cap B &= \{\text{outcomes with a } ^+ \text{ rhesus status } \textbf{and} \text{ high SBP}\} \\ &= \{(\text{AB}^+, \text{H}), (\text{A}^+, \text{H}), (\text{B}^+, \text{H}), (\text{O}^+, \text{H})\} \end{aligned}$$

$$\begin{aligned} \overline{A} &= \{\text{outcomes with a } ^- \text{ rhesus status}\} \\ &= \{(\text{AB}^-, \text{L}), (\text{A}^-, \text{L}), (\text{B}^-, \text{L}), (\text{O}^-, \text{L}), (\text{AB}^-, \text{M}), (\text{A}^-, \text{M}), (\text{B}^-, \text{M}), (\text{O}^-, \text{M}), \\ &\quad (\text{AB}^-, \text{H}), (\text{A}^-, \text{H}), (\text{B}^-, \text{H}), (\text{O}^-, \text{H})\} \end{aligned}$$

Exercise: List the outcomes in $\overline{A} \cup B$, $A \cap \overline{B}$, and $\overline{A} \cap \overline{B}$. $\square$

Here are two last set theory results that will prove to be useful.

Distributive Laws:

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned}$$

DeMorgan's Laws:

$$\begin{aligned} \overline{A \cup B} &= \overline{A} \cap \overline{B} \\ \overline{A \cap B} &= \overline{A} \cup \overline{B} \end{aligned}$$

2.3 Axioms of probability

Terminology: We say that two events A and B are **mutually exclusive** or **disjoint** if

$$A \cap B = \emptyset,$$

that is, A and B have no outcomes in common. Mutually exclusive events cannot occur simultaneously. For example, clearly A and $\overline{A}$ are mutually exclusive. If A occurs, then $\overline{A}$ cannot occur and vice versa.

Kolmogorov's Axioms: Suppose S is a sample space and let $\mathcal{B}$ denote the collection of all possible events. Let P be a set function; i.e.,

$$P : \mathcal{B} \to [0, 1],$$

that satisfies the following axioms:

1. $P(A) \geq 0$, for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, ..., \in \mathcal{B}$ are pairwise mutually exclusive; i.e., $A_i \cap A_j = \emptyset \; \forall i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

We call P a **probability set function** (or probability measure).

Remark: Mathematically, we can think of P as a function whose domain is sets (events) and whose range is $[0, 1]$. Therefore, probabilities are numbers between 0 and 1 (inclusive). In a more advanced course, one would describe the collection of events $\mathcal{B}$ much more carefully to avoid certain peculiar mathematical contradictions; we will not.

Consequences: From these 3 axioms, we can develop numerous rules which help us assign probability to events.

1. Complement rule: $P(A) = 1 - P(\overline{A})$, for any event A.
Proof. We can write $S = A \cup \overline{A}$. Because A and $\overline{A}$ are mutually exclusive, Axiom 3 says

$$P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}).$$

However, $P(S) = 1$ by Axiom 2. Therefore,

$$P(A) = 1 - P(\overline{A}). \quad \square$$

Importance: In many problems, it is often much easier to calculate the probability that A does *not* occur. If you can do this, then the complement rule gives $P(A)$ easily.

2. Null set rule: $P(\emptyset) = 0$.
Proof. This follows immediately from the complement rule; take $A = \emptyset$ and $\overline{A} = S$. $\square$

3. Upper bound rule: $P(A) \leq 1$.
Proof. In the proof of the complement rule, we wrote

$$P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}).$$

However, $P(\overline{A}) \geq 0$ by Axiom 1 and $P(S) = 1$ by Axiom 2. $\square$

4. Monotonicity rule: Suppose A and B are events such that $A \subset B$, that is, A is a subset of B. Then

$$P(A) \leq P(B).$$

This result makes sense intuitively. If A occurs, then B must occur. However, the reverse is not true, so $P(B)$ must be larger (or at least not smaller).

Proof. Because $A \subset B$, we can write

$$B = A \cup (B \cap \overline{A});$$

i.e., $B \cap \overline{A}$ captures all outcomes in B and not in A. Clearly, A and $(B \cap \overline{A})$ are mutually exclusive. Thus, from Axiom 3, we have

$$P(B) = P(A) + P(B \cap \overline{A}).$$

However, from Axiom 1, $P(B \cap \overline{A}) \geq 0$, so $P(B) \geq P(A)$. $\square$

5. Additive rule: Suppose A and B are two events.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Remark: We know if A and B are mutually exclusive, then

$$P(A \cup B) = P(A) + P(B).$$

This is what Axiom 3 guarantees. So the additive rule is more general; i.e., A and B need not be mutually exclusive.

Proof. Write $A \cup B = A \cup (\overline{A} \cap B)$. Because A and $(\overline{A} \cap B)$ are mutually exclusive,

$$P(A \cup B) = P(A) + P(\overline{A} \cap B)$$

by Axiom 3. Now, write $B = (A \cap B) \cup (\overline{A} \cap B)$. Clearly, $(A \cap B)$ and $(\overline{A} \cap B)$ are mutually exclusive too. From Axiom 3 again,

$$P(B) = P(A \cap B) + P(\overline{A} \cap B).$$

Combining the expressions for $P(\overline{A} \cap B)$ in both equations above gives the result. $\square$

Example 2.6. A smoke detector system uses two interlinked units. If smoke is present, the probability the first unit will detect it is 0.95 and the probability the second unit will detect it is 0.90. The probability smoke will be detected by both units is 0.88.

(a) If smoke is present, find the probability that the smoke will be detected by either unit or both.
(b) Find the probability the smoke will go undetected.

Solutions. Define the events

$$\begin{aligned} A &= \{\text{first unit detects smoke}\} \\ B &= \{\text{second unit detects smoke}\}. \end{aligned}$$

We are given $P(A) = 0.95$, $P(B) = 0.90$, and $P(A \cap B) = 0.88$.

(a) The probability the system will detect smoke (when present) is

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.95 + 0.90 - 0.88 = 0.97. \end{aligned}$$

(b) Smoke will go undetected when $\overline{A} \cap \overline{B}$ occurs. By DeMorgan's Law,

$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 0.03. \ \square$$

Remark: The additive rule can be generalized for any sequence of events $A_1, A_2, ..., A_n$; i.e.,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i_1<i_2} \sum P(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3} \sum \sum P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} A_i\right).$$

For example, if $n = 3$, then

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).$$

Example 2.7. Prove **Bonferroni's Inequality**; i.e.,

$$P(A \cap B) \geq 1 - P(\overline{A}) - P(\overline{B}).$$

Proof. From the additive rule and complement rule, we know

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= [1 - P(\overline{A})] + [1 - P(\overline{B})] - P(A \cup B) \\ &= 1 - P(\overline{A}) - P(\overline{B}) + [1 - P(A \cup B)]. \end{aligned}$$

However, $1 - P(A \cup B) = P(\overline{A \cup B})$ is itself a probability and hence $1 - P(A \cup B) \geq 0$ by Axiom 1. Thus, we are done. $\square$

Generalization: Bonferroni's Inequality can be generalized for any sequence of events $A_1, A_2, ..., A_n$; i.e.,

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\overline{A}_i).$$

Application: Bonferroni's Inequality is useful in statistics when multiple confidence intervals are being written. In this context, the event

$$A_i = \{i\text{th interval includes its target parameter}\}$$

and the event

$$\bigcap_{i=1}^{n} A_i = \{\text{each confidence interval includes its target parameter}\}.$$

For example, if $n = 5$ and $P(A_i) = 0.95$ (confidence level), then the probability all 5 intervals will include their target parameter is

$$P\left(\bigcap_{i=1}^{5} A_i\right) \geq 1 - 5(0.05) = 0.75.$$

This probability, which corresponds to the family of 5 intervals taken together, can be much lower than each interval's confidence level of 0.95. Furthermore, the fact that this "familywise confidence level" can be so low is concerning.

2.4 Discrete sample spaces

Terminology: Suppose S is a sample space for a random experiment. If S contains a finite or countable number of outcomes, we call S a **discrete sample space**. Recall:

- Finite: $|S| < \infty$; i.e., the number of outcomes in S is finite
- Countable: the number of outcomes may be infinite; i.e., $|S| = +\infty$, but the outcomes can be put into a 1:1 correspondence with $\mathbb{N} = \{1, 2, 3, ...,\}$.

Example 2.8. An American style roulette wheel contains 38 numbered compartments or "pockets." The pockets are either red, black, or green. The numbers 1 through 36 are evenly split between red and black, while 0 and 00 are green pockets. Conceptualize the next spin of the wheel as a random experiment with sample space

$$S = \{1, 2, 3, 4, ..., 34, 35, 36, 0, 00\}.$$

Note that this is a discrete sample space with $|S| = 38$ outcomes (sample points).

Consider the following events (i.e., subsets of S):

$$\begin{aligned} A_1 &= \{13\} \\ A_2 &= \{\text{“red”}\} = \{1,3,5,7,9,12,14,16,18,19,21,23,25,27,30,32,34,36\} \\ A_3 &= \{\text{“green”}\} = \{0,00\}. \end{aligned}$$

Terminology: A **simple event** is an event that consists of exactly one outcome (sample point).

- In Example 2.8, the event $A_1 = \{13\}$ is a simple event.

A **compound event** is an event that contains more than one outcome (sample point). Therefore, any compound event can be written as the (countable) union of simple events. In Example 2.8, the event $A_3 = \{\text{“green”}\} = \{0,00\}$ can be written as

$$A_3 = \{0\} \cup \{00\},$$

the union of 2 simple events. The event $A_2 = \{\text{“red”}\}$ can be written as the union of 18 simple events.

Important: In a discrete sample space, calculating the probability of a compound event A is done by adding up the probabilities associated with each sample point in it. That is,

$$P(A) = \sum_{i:E_i \subset A} P(E_i),$$

where $E_1, E_2, ...,$ denote the simple events whose union makes up A. This strategy to calculate $P(A)$ follows from Axiom 3. If the compound event A can be expressed as $A = E_1 \cup E_2 \cup \cdots \cup E_{n_a}$ (for a finite number of simple events), then

$$\begin{aligned} P(A) &= P(E_1 \cup E_2 \cup \cdots \cup E_{n_a}) \\ &= P(E_1) + P(E_2) + \cdots + P(E_{n_a}). \end{aligned}$$

Example 2.9. Consider the random experiment of observing the number of children born during the next live birth in the United States. A sample space for this experiment is

$$S = \{1,2,3,4,5+\}.$$

Note that this is a discrete sample space with $|S| = 5$ outcomes (sample points).

Let E_1, E_2, E_3, E_4, E_5 denote the five simple events that make up S. The CDC reports the following probabilities during 2015 (among all 3,978,497 live births):

$$\begin{aligned} P(E_1) &= 0.965399 \\ P(E_2) &= 0.033489 \\ P(E_3) &= 0.001047 \\ P(E_4) &= 0.000059 \\ P(E_5) &= 0.000006 \end{aligned}$$

It is easy to see that $P(E_1) + P(E_2) + P(E_3) + P(E_4) + P(E_5) = 1$, as it should (Axiom 2).

Q: Under this model, what is the probability of a multiple birth? Note that the event $A = \{\text{multiple birth}\}$ can be written as the union of the 4 simple events:

$$A = E_2 \cup E_3 \cup E_4 \cup E_5.$$

Therefore,

$$\begin{aligned} P(A) &= P(E_2) + P(E_3) + P(E_4) + P(E_5) \\ &= 0.033489 + 0.001047 + 0.000059 + 0.000006 = 0.034601. \end{aligned}$$

Of course, using the complement rule gives you the same answer:

$$\begin{aligned} P(A) &= 1 - P(E_1) \\ &= 1 - 0.965399 = 0.034601. \ \square \end{aligned}$$

Example 2.10. Two jurors are needed to serve as "alternates" in an attempted murder trial. These two jurors will be selected from 5 potential jurors, three men and two women. Envision the selection of these two jurors as a random experiment with sample space

$$S = \{(\mathrm{M}_1, \mathrm{M}_2), (\mathrm{M}_1, \mathrm{M}_3), (\mathrm{M}_1, \mathrm{W}_1), (\mathrm{M}_1, \mathrm{W}_2), (\mathrm{M}_2, \mathrm{M}_3), (\mathrm{M}_2, \mathrm{W}_1), (\mathrm{M}_2, \mathrm{W}_2), (\mathrm{M}_3, \mathrm{W}_1), (\mathrm{M}_3, \mathrm{W}_2), (\mathrm{W}_1, \mathrm{W}_2)\}.$$

Note that this is a discrete sample space with $|S| = 10$ outcomes (sample points).

Q: What is the probability at least one woman is selected as an alternate juror? We do not have enough information to answer this question because we don't know the probabilities associated with the 10 sample points. However, certainly we can list out the sample points in this event:

$$\begin{aligned} A &= \{\text{at least one woman selected}\} \\ &= \{(\mathrm{M}_1, \mathrm{W}_1), (\mathrm{M}_1, \mathrm{W}_2), (\mathrm{M}_2, \mathrm{W}_1), (\mathrm{M}_2, \mathrm{W}_2), (\mathrm{M}_3, \mathrm{W}_1), (\mathrm{M}_3, \mathrm{W}_2), (\mathrm{W}_1, \mathrm{W}_2)\}. \end{aligned}$$

Note: If we assume the outcomes in S are equally likely; i.e., each with probability

$$\frac{1}{|S|} = \frac{1}{10},$$

then we can compute $P(A)$. It is simply

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{7}{10}.$$

However, it is important to understand that this simple rule for assigning probabilities is only valid when the outcomes in S are equally likely. If the outcomes in S are not equally likely, then this probability assignment is not correct.

Equiprobability model: Suppose the discrete sample space S contains $N = |S| < \infty$ outcomes (sample points), each of which is equally likely. If the event A contains n_a outcomes (sample points), then

$$P(A) = \frac{n_a}{N}.$$

Proof. Write $S = E_1 \cup E_2 \cup \cdots \cup E_N$, where E_i denotes the ith simple event, for $i = 1, 2, ..., N$. Then,

$$1 = P(S) = P(E_1 \cup E_2 \cup \cdots \cup E_N) = \sum_{i=1}^{N} P(E_i),$$

by Axioms 2 and 3. Because $P(E_1) = P(E_2) = \cdots = P(E_N)$,

$$1 = \sum_{i=1}^{N} P(E_i) \quad \Longrightarrow \quad P(E_i) = \frac{1}{N}, \quad i = 1, 2, ..., N.$$

Without loss of generality, take $A = E_1 \cup E_2 \cup \cdots \cup E_{n_a}$. Then,

$$P(A) = P(E_1 \cup E_2 \cup \cdots \cup E_{n_a}) = \sum_{i=1}^{n_a} P(E_i) = \sum_{i=1}^{n_a} \frac{1}{N} = \frac{n_a}{N}. \ \square$$

Implication: Suppose S is a discrete sample space with a finite number of outcomes; i.e., $N < \infty$. If each outcome is equally likely, then finding $P(A)$ reduces to two "counting problems:" one to find N and one to find n_a.

- In simple experiments, like Example 2.10, we can simply list out all outcomes in S and A and count to find N and n_a quickly.
- In more complicated experiments, it may not be possible to do this so quickly. We need combinatoric rules (counting rules) to accomplish this.
- Combinatoric rules are used in probability to count the number of outcomes.

2.5 Tools for counting outcomes (sample points)

2.5.1 Basic counting rule

Basic counting rule ("mn rule"): Suppose we would like to count the number of paired outcomes formed by two factors. The first factor has m outcomes. The second factor has n outcomes. The total number of paired outcomes is mn.

Example 2.11. An experiment consists of rolling a die (with faces 1, 2, ..., 6) and tossing a coin (with sides H and T). The die has $m = 6$ outcomes. The coin has $n = 2$ outcomes. There are $mn = 12$ paired outcomes. The sample space for this experiment is

$$S = \{(1, \text{H}), (2, \text{H}), (3, \text{H}), (4, \text{H}), (5, \text{H}), (6, \text{H}), (1, \text{T}), (2, \text{T}), (3, \text{T}), (4, \text{T}), (5, \text{T}), (6, \text{T})\}.$$

There are $N = 12$ outcomes (sample points) in S. $\square$

Generalization: The basic counting rule can be generalized easily. Suppose there are k factors with

$$\begin{aligned} n_1 &= \text{number of outcomes for factor 1} \\ n_2 &= \text{number of outcomes for factor 2} \\ &\vdots \\ n_k &= \text{number of outcomes for factor } k. \end{aligned}$$

The total number of outcomes is

$$\prod_{i=1}^{k} n_i = n_1 \times n_2 \times \cdots \times n_k.$$

Example 2.12. An experiment consists of selecting a standard South Carolina license plate which consists of 3 letters and 3 numbers. We can think of one outcome (sample point) in the underlying sample space S as having the following structure:

$$(\_\_ \ \_\_ \ \_\_ \quad \_\_ \ \_\_ \ \_\_).$$

Q: How many standard plates are possible; i.e., how many outcomes are in S?
A: There are

$$N = 26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17576000$$

possible outcomes.

Q: Assume each outcome in S is equally likely (e.g., license plate letters/numbers are determined at random). What is the probability a randomly selected plate contains no repeat letters and no repeat numbers?
A: Define the event

$$A = \{\text{no repeat letters/numbers}\}.$$

The number of outcomes in A is

$$n_a = 26 \times 25 \times 24 \times 10 \times 9 \times 8 = 11232000.$$

Therefore,

$$P(A) = \frac{n_a}{N} = \frac{11232000}{17576000} \approx 0.6391. \ \Box$$

Example 2.13. *The birthday problem, revisited.* An experiment consists of observing the birthday of $M = 50$ students. Assume 365 days. There are

$$N = 365 \times 365 \times 365 \times \cdots \times 365 = 365^{50}$$

possible outcomes. We can think of one outcome (sample point) in the underlying sample space S as having the following structure:

$$(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_).$$

Q: Assume each outcome in S is equally likely; e.g., no twins/triplets, etc. What is the probability there will be at least one shared birthday?
A: Define the event

$$\overline{A} = \{\text{no shared birthdays}\}.$$

The number of outcomes in $\overline{A}$ is

$$365 \times 364 \times 363 \cdots \times 317 \times 316 = 50!\binom{365}{50}.$$

Therefore,

$$P(\overline{A}) = \frac{50!\binom{365}{50}}{365^{50}} \approx 0.0296.$$

Using the complement rule, the probability of at least one shared birthday is

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{50!\binom{365}{50}}{365^{50}} \approx 0.9704.$$

Compare this with Example 2.3 where we used simulation to "estimate" this answer. □

2.5.2 Permutations

Remark: We have seen examples where constructing sample spaces requires us to work with distinct "objects;" e.g., license plate digits, students, etc. Counting the number of outcomes (in S or in A) often requires us to count the number of ways distinct objects can be arranged in a sequence.

Terminology: A **permutation** is an arrangement of distinct objects in a particular order. *Order is important.*

Result: Suppose I have n distinct objects. There are

$$n! = n(n-1)(n-2) \times \cdots \times 2 \times 1$$

ways to permute these objects (i.e., to arrange them in a particular order).

Example 2.14. Consider the experiment of arranging 10 distinct books on my bookshelf. There are

$$N = 10! = 3628800$$

possible permutations. We can think of one sample point in the underlying sample space S as having the following structure:

$$(\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ ).$$

Q: Assume each outcome in S is equally likely. If there are 5 math books, 3 physics books, and 2 chemistry books, what is the probability that a randomly selected arrangement will keep like-subject books together?

A: Define the event

$$A = \{\text{like-subject books kept together}\}.$$

The number of outcomes in A can be found using the basic counting rule with

$$\begin{aligned} n_1 &= \text{number of ways to permute M, P, and C ordering} = 3! \\ n_2 &= \text{number of ways to permute M books} = 5! \\ n_3 &= \text{number of ways to permute P books} = 3! \\ n_4 &= \text{number of ways to permute C books} = 2! \end{aligned}$$

The number of outcomes in A is

$$n_a = 3! \times 5! \times 3! \times 2! = 8640.$$

Therefore,

$$P(A) = \frac{n_a}{N} = \frac{8640}{3628800} \approx 0.0024. \ \square$$

Remark: In Example 2.14, our goal was to permute n distinct objects (i.e., books). In other problems, we first select r objects (from the available n) and then permute those only.

Result: From a collection of n distinct objects, we select and permute r of them ($r \leq n$). The number of ways to do this is

$$P_r^n = \frac{n!}{(n-r)!}.$$

The symbol P_r^n is read "the permutation of n things taken r at a time."

Proof. Envision r slots. There are n ways to fill the first slot, $n-1$ ways to fill the second slot, and so on, until we get to the rth slot, where there are $n-r+1$ ways to fill it. From the basic counting rule, there are

$$n(n-1)(n-2) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}$$

different permutations. $\square$

Example 2.15. A personnel director for a corporation has hired 12 new engineers. She must pick 3 engineers to fill distinct positions (team leader, consultant, support staff member). Note that because these positions are inherently different, the selection ordering matters;

- e.g., the outcome (Jim, Mary, Celeste) and the outcome (Celeste, Jim, Mary) are different outcomes.

Conceptualize the selection of 3 engineers from 12 as a random experiment. We can think of one outcome (sample point) in the underlying sample space S as having the following structure:

$$(\ \_\_\ \_\_\ \_\_\ ).$$

Because the ordering within outcomes is important, there are

$$N = P_3^{12} = \frac{12!}{(12-3)!} = 12 \times 11 \times 10 = 1320$$

outcomes in S.

Q: Assume each outcome in S is equally likely. Suppose there are 6 engineers from USC and 6 from Clemson. What is the probability a USC graduate is selected as the team leader and the remaining 2 positions are filled by Clemson graduates?
A: Define the event

$$A = \{\text{USC team leader and Clemson graduates for other 2 positions}\}.$$

The number of outcomes in A can be found using the basic counting rule with

$$\begin{aligned} n_1 &= \text{number of ways to select 1 USC graduate} = 6 \\ n_2 &= \text{number of ways to select 2 Clemson graduates} = P_2^6 \end{aligned}$$

The number of outcomes in A is

$$n_a = 6 \times P_2^6 = 6 \times 30 = 180.$$

Therefore,

$$P(A) = \frac{n_a}{N} = \frac{180}{1320} \approx 0.1363. \ \square$$

2.5.3 Multinomial coefficients

Example 2.16. How many permutations of the letters in the word PEPPER are there?

Solution. Initially treat each of the 6 letters as distinct objects and emphasize this by writing

$$\mathrm{P_1E_1P_2P_3E_2R}.$$

We know there are

$$6! = 720$$

possible permutations of these 6 distinct objects. Now, because the letters in PEPPER really are not distinct, the number of possible permutations is smaller 720. By how much? Note that there are

$$\begin{aligned} &3! \text{ ways to permute the Ps} \\ &2! \text{ ways to permute the Es} \\ &1! \text{ ways to permute the Rs.} \end{aligned}$$

Therefore, 6! is $3! \times 2! \times 1!$ times too large. The number of permutations is

$$\frac{6!}{3!\ 2!\ 1!} = 60. \ \square$$

Terminology: Multinomial coefficients arise in the algebraic expansion of the multinomial expression $(x_1 + x_2 + \cdots + x_k)^n$; i.e.,

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_D \binom{n}{n_1 n_2 \cdots n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where $D = \{(n_1, n_2, ..., n_k) : \sum_{i=1}^k n_i = n\}$. The multinomial coefficient

$$\binom{n}{n_1 n_2 \cdots n_k} = \frac{n!}{n_1!\ n_2!\ \cdots\ n_k!}.$$

Importance: In counting problems, multinomial coefficients are used to count

- the number of ways to permute n objects, of which n_1 are "alike," n_2 are "alike," and so on (see Example 2.16).
- the number of ways to partition n distinct objects into k distinct groups containing $n_1, n_2, ..., n_k$ objects, respectively (where $\sum_{i=1}^k n_i = n$).

Example 2.17. A police department in a small city consists of 10 officers. The department policy is to have 5 officers patrolling the streets, 2 officers working at the station, and 3 officers on reserve. How many divisions (partitions) of the 10 officers are possible?

$$\textbf{A:} \quad \binom{10}{5\ 2\ 3} = \frac{10!}{5!\ 2!\ 3!} = 2520. \ \square$$

Example 2.18. A signal is formed by arranging 9 flags in a line. There are 4 white flags, 3 blue flags, and 2 yellow flags. Envision the process of forming a signal as a random experiment.

We can think of one outcome (sample point) in the underlying sample space S as having the following structure:

$$(\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ ).$$

Q: What is the probability the signal has the 4 white flags grouped together?

Note: We offer two solutions. The solutions differ in the way we conceptualize what a sample point looks like:

1. one solution treats flags of the same color as "indistinguishable" objects
2. one solution treats all 9 flags as distinct objects.

In the first conceptualization, a sample point might look like

$$(\ \text{B W W Y B Y B W W}\ )$$

In the second, a sample point might look like

$$(\text{B}_3 \ \text{W}_2 \ \text{W}_1 \ \text{Y}_1 \ \text{B}_1 \ \text{Y}_2 \ \text{B}_2 \ \text{W}_4 \ \text{W}_3)$$

Important: Different counting rules are needed for each solution. In both solutions, we define

$$A = \{\text{white flags grouped together}\}.$$

and assume that outcomes (sample points) are equally likely.

Solution 1: Treat flags of the same color as "indistinguishable" objects. The number of sample points in S is

$$N = \binom{9}{4\ 3\ 2} = \frac{9!}{4!\ 3!\ 2!} = 1260.$$

This is the number of ways to permute 9 objects, of which 4 are "alike," 3 are "alike," and 2 are "alike." Now, we need to count the number of sample points in A. We can do this using the basic counting rule:

n_1 = number of ways to select 4 adjacent positions for W flags = 6

n_2 = number of ways to permute B/Y flags among the remaining positions = $\binom{5}{3\ 2}$

Therefore,

$$n_a = n_1 \times n_2 = 6 \times \binom{5}{3\ 2} = 60$$

and

$$P(A) = \frac{n_a}{N} = \frac{60}{1260} \approx 0.0476.$$

Solution 2: Treat all 9 flags as distinct objects. The number of sample points in S is

$$N = 9! = 362880.$$

This is the number of ways to permute 9 distinct objects. Now, we need to count the number of sample points in A. We can do this using the basic counting rule:

n_1 = number of ways to select 4 adjacent positions for W flags = 6

n_2 = number of ways to permute W flags = 4!

n_3 = number of ways to permute B/Y flags among the remaining positions = 5!

Therefore,

$$n_a = n_1 \times n_2 \times n_3 = 6 \times 4! \times 5! = 17280$$

and

$$P(A) = \frac{n_a}{N} = \frac{17280}{362880} \approx 0.0476. \ \square$$

2.5.4 Combinations

Result: From a collection of n distinct objects, we choose r of them ($r \leq n$) *without regard to the order in which the objects are chosen*. The number of ways to do this is

$$C_r^n = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

The symbol C_r^n is read "the combination of n things taken r at a time."

Remark: To see why this makes sense, envision n distinct objects. The number of ways to partition these objects into 2 distinct groups, of which r are "alike" (i.e., the chosen objects) and $n-r$ are "alike" (i.e., the objects not chosen) is given by the multinomial coefficient

$$\binom{n}{r \;\; n-r} = \frac{n!}{r!\,(n-r)!}.$$

Remark: We adopt the notation $\binom{n}{r}$, read "n choose r," henceforth as the symbol for C_r^n. The terms $\binom{n}{r}$ are called **binomial coefficients** because they arise in the algebraic expansion of a binomial; i.e.,

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r} b^r.$$

Example 2.19. In Example 2.15, a personnel director was tasked with choosing 3 engineers from 12 to fill distinct positions. If the positions are not distinct, then there are

$$N = \binom{12}{3} = \frac{12!}{3!\,(12-3)!} = 220$$

possible ways to select 3 engineers.

Q: Assume each combination is equally likely. Suppose there are 6 engineers from USC and 6 from Clemson. What is the probability of selecting 1 USC engineer and 2 from Clemson?

A: Define the event

$$A = \{\text{1 USC graduate and 2 Clemson graduates chosen}\}.$$

The number of outcomes in A can be found using the basic counting rule with

$$\begin{aligned} n_1 &= \text{number of ways to select 1 USC graduate} = \binom{6}{1} = 6 \\ n_2 &= \text{number of ways to select 2 Clemson graduates} = \binom{6}{2} = 15 \end{aligned}$$

The number of outcomes in A is

$$n_a = n_1 \times n_2 = 6 \times 15 = 90.$$

Therefore,

$$P(A) = \frac{n_a}{N} = \frac{90}{220} \approx 0.4091. \;\square$$

Remark: From Examples 2.15 and 2.19, one should note that, in general,

$$P_r^n = r! \times \binom{n}{r}.$$

This formula highlights the difference between P_r^n and $\binom{n}{r}$. To count the number of ways to permute n objects chosen r at a time, we first must choose the r objects. The binomial coefficient $\binom{n}{r}$ counts the number of ways to do this. Then, once we have our r chosen objects, there are $r!$ ways to permute them.

Example 2.20. Consider the experiment of drawing 5 cards from a standard deck of 52 cards (without replacement). We can conceptualize the sample space as

$$S = \{[2_S, 2_D, 2_H, 2_C, 3_S], [2_S, 2_D, 2_H, 2_C, 3_D], [2_S, 2_D, 2_H, 2_C, 3_H], ..., [A_S, A_D, A_H, A_C, K_C]\}.$$

The number of outcomes in S is

$$N = \binom{52}{5} = \frac{52!}{5!\,(52-5)!} = 2598960.$$

Q: Assuming that each outcome in S is equally likely, what is the probability of getting "3 of a kind?"

A: Define the event

$$A = \{\text{"3 of a kind"}\}.$$

The number of outcomes in A can be found using the basic counting rule with

$$\begin{aligned}
n_1 &= \text{number of ways to choose denomination} = \binom{13}{1} = 13 \\
n_2 &= \text{number of ways to choose 3 suits} = \binom{4}{3} = 4 \\
n_3 &= \text{number of ways to choose 2 other denominations} = \binom{12}{2} = 66 \\
n_4 &= \text{number of ways to choose 1 card for each "other" denomination} = \binom{4}{1}^2 = 16
\end{aligned}$$

The number of outcomes in A is

$$n_a = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 4 \times 66 \times 16 = 54912.$$

Therefore,

$$P(A) = \frac{n_a}{N} = \frac{54912}{2598960} \approx 0.0211. \ \square$$

Note: In choosing the 2 other denominations above (Step 3), it is important to remember that these 2 denominations must be different. If they are the same, then the hand is a "full house" instead (not a lesser "3 of a kind" hand).

Example 2.21. *The matching problem, revisited.* Suppose M men are at a party, and each man is wearing a hat. Each man throws his hat into the center of the room. Each man then selects a hat at random. What is the probability at least one man selects his own hat; i.e., there is at least one "match"? Define

$$A = \{\text{at least one man selects his own hat}\}$$

and the events

$$A_i = \{\text{the } i\text{th man selects his own hat}\}, \quad i = 1, 2, ..., M,$$

so that

$$A = \bigcup_{i=1}^{M} A_i \implies P(A) = P\left(\bigcup_{i=1}^{M} A_i\right).$$

We now use the additive rule for M events (see pp 11, notes). Note the following:

$$\begin{aligned} P(A_i) &= \frac{(M-1)!}{M!} = \frac{1}{M} \qquad \forall i = 1, 2, ..., M \\ P(A_{i_1} \cap A_{i_2}) &= \frac{(M-2)!}{M!} \qquad 1 \leq i_1 < i_2 \leq M \\ P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) &= \frac{(M-3)!}{M!} \qquad 1 \leq i_1 < i_2 < i_3 \leq M \end{aligned}$$

This pattern continues; the probability of the M-fold intersection is

$$P\left(\bigcap_{i=1}^{M} A_i\right) = \frac{(M-M)!}{M!} = \frac{1}{M!}.$$

Therefore, by the additive rule, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^{M} A_i\right) &= \sum_{i=1}^{M} P(A_i) - \sum\sum_{i_1<i_2} P(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum\sum\sum_{i_1<i_2<i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{M+1} P\left(\bigcap_{i=1}^{M} A_i\right) \\ &= M\left(\frac{1}{M}\right) - \binom{M}{2}\frac{(M-2)!}{M!} + \binom{M}{3}\frac{(M-3)!}{M!} - \cdots + (-1)^{M+1}\frac{1}{M!} \\ &= \sum_{k=1}^{M} (-1)^{k+1} \binom{M}{k} \frac{(M-k)!}{M!} \\ &= \sum_{k=1}^{M} (-1)^{k+1} \frac{M!}{k!(M-k)!}\frac{(M-k)!}{M!} = 1 - \sum_{k=0}^{M} \frac{(-1)^k}{k!}. \end{aligned}$$

Compare this with Example 2.2 where we used simulation to "estimate" this answer for different values of M. $\square$

2.6 Conditional probability and independence

Remark: The probability an event A will often depend on other "related" events. If we know another one of these related events has occurred (or has not occurred), this may change the way we assess the likelihood of A occurring.

Example 2.22. Consider the sample space in Example 2.10,

$$S = \{(\mathrm{M}_1, \mathrm{M}_2), (\mathrm{M}_1, \mathrm{M}_3), (\mathrm{M}_1, \mathrm{W}_1), (\mathrm{M}_1, \mathrm{W}_2), (\mathrm{M}_2, \mathrm{M}_3), (\mathrm{M}_2, \mathrm{W}_1), (\mathrm{M}_2, \mathrm{W}_2), (\mathrm{M}_3, \mathrm{W}_1), (\mathrm{M}_3, \mathrm{W}_2), (\mathrm{W}_1, \mathrm{W}_2)\},$$

where the experiment consisted of choosing two alternate jurors from three men and two women. Define the event

$$A = \{\text{two women are chosen}\}.$$

Assuming each outcome (sample point) in S is equally likely, clearly

$$P(A) = \frac{n_a}{N} = \frac{1}{10}.$$

Now suppose we know that at least one of the jurors chosen is a woman. That is, the event

$$\begin{aligned} B &= \{\text{at least one woman chosen}\} \\ &= \{(\mathrm{M}_1, \mathrm{W}_1), (\mathrm{M}_1, \mathrm{W}_2), (\mathrm{M}_2, \mathrm{W}_1), (\mathrm{M}_2, \mathrm{W}_2), (\mathrm{M}_3, \mathrm{W}_1), (\mathrm{M}_3, \mathrm{W}_2), (\mathrm{W}_1, \mathrm{W}_2)\} \end{aligned}$$

has occurred. How does the knowledge of B occurring influence how we assign probability to A?

In essence, a "new" sample space emerges when we know that B has occurred, namely, the new sample space is B. Continuing to assume outcomes are equally likely, the probability A occurs has now changed to

$$P(A|B) = \frac{1}{7}.$$

We write $P(A|B)$ to emphasize this is a conditional probability. □

Terminology: Let A and B be events in a sample space S. The **conditional probability** of A, given that B has occurred, is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

provided that $P(B) > 0$.

Example 2.23. Brazilian scientists have identified a new strain of the H1N1 virus. The genetic sequence of the new strain consists of alterations in the hemagglutinin protein, making it significantly different than the usual H1N1 strain. Public health officials wish to study the population of residents in Rio de Janeiro.

Suppose that in this population,

- the probability of catching the usual strain is 0.10
- the probability of catching the new strain is 0.05
- the probability of catching both strains is 0.01.

(a) Find the probability of catching the usual strain, given that the new strain is caught.
(b) Find the probability of catching the new strain, given that at least one strain is caught.

Solutions. Define the events

$$\begin{aligned} A &= \{\text{resident catches usual strain}\} \\ B &= \{\text{resident catches new strain}\}. \end{aligned}$$

From the information above, we have $P(A) = 0.10$, $P(B) = 0.05$, and $P(A \cap B) = 0.01$.

(a) Using the definition of conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.01}{0.05} = 0.20.$$

(b) If "at least one strain is caught," this means $A \cup B$ has occurred. Therefore,

$$\begin{aligned} P(B|A \cup B) = \frac{P(B \cap (A \cup B))}{P(A \cup B)} &= \frac{P(B)}{P(A) + P(B) - P(A \cap B)} \\ &= \frac{0.05}{0.10 + 0.05 - 0.01} \approx 0.3571. \end{aligned}$$

Note above that $B \subset (A \cup B)$ so $B \cap (A \cup B) = B$.

Exercise: Find the probability of not catching the usual strain, given that the new strain is not caught. □

Important: Suppose P is a valid probability set function over $(S, \mathcal{B})$; i.e., it satisfies the Kolmogorov axioms. Provided that $P(B) > 0$, the conditional probability assignment

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

also satisfies the Kolmogorov axioms; i.e.,

1. $P(A|B) \geq 0$, for all $A \in \mathcal{B}$
2. $P(B|B) = 1$
3. If $A_1, A_2, ..., \in \mathcal{B}$ are pairwise mutually exclusive; i.e., $A_i \cap A_j = \emptyset \; \forall i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} P(A_i|B).$$

Implication: Because the way we assign conditional probability also satisfies the Kolmogorov axioms, all the probability rules we derived earlier have their respective "conditional versions." For example,

1. **Complement rule:** $P(\overline{A}|B) = 1 - P(A|B)$
2. **Monotonicity:** If $A_1 \subset A_2$, then $P(A_1|B) \leq P(A_2|B)$
3. **Additive rule:** $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$.

Terminology: Suppose A and B are events in S. We say A and B are **independent** if

$$P(A \cap B) = P(A)P(B).$$

If both $P(A) > 0$ and $P(B) > 0$, then the following three conditions for independence are equivalent:

$$\begin{aligned} P(A \cap B) &= P(A)P(B). \\ P(A|B) &= P(A) \\ P(B|A) &= P(B). \end{aligned}$$

Remark: Do not confuse "independence" with what it means for A and B to be mutually exclusive. Independence means that the occurrence of A does not affect whether B occurs (and vice versa). If A and B are mutually exclusive, this means that A and B can not occur simultaneously.

Exercise: Suppose $P(A) > 0$ and $P(B) > 0$. Prove that if A and B are mutually exclusive, then A and B cannot be independent. Now go the other way. Prove that if A and B are independent, then A and B cannot be mutually exclusive.

Example 2.24. An electrical system consists of two components. The probability the second component functions in a satisfactory manner during its design life is 0.90. The probability at least one of the two components does so is 0.96. The probability both components do so is 0.75. Do the two components function independently?

Solution. Define the events

$$\begin{aligned} A &= \{\text{component 1 functions}\} \\ B &= \{\text{component 2 functions}\}. \end{aligned}$$

From the information above, we have $P(B) = 0.90$, $P(A \cup B) = 0.96$, and $P(A \cap B) = 0.75$. The additive rule gives

$$0.96 = P(A) + 0.90 - 0.75 \implies P(A) = 0.81.$$

However,

$$0.75 = P(A \cap B) \neq P(A)P(B) = 0.81(0.90) = 0.729.$$

Therefore, the events A and B are not independent.

Exercise: Check that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$. $\square$

Result: Suppose A and B are events in S. If A and B are independent, then so are

(a) A and $\overline{B}$

(b) $\overline{A}$ and B

(c) $\overline{A}$ and $\overline{B}$.

Proof. We prove part (a) only; the other parts follow similarly. Suppose A and B are independent. Then

$$P(\overline{A} \cap B) = P(\overline{A}|B)P(B) = [1 - P(A|B)]P(B) = [1 - P(A)]P(B) = P(\overline{A})P(B).$$

We used the fact that A and B were independent above when we wrote $P(A|B) = P(A)$. □

Multiplication rule: Suppose A and B are events in S. Then

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A). \end{aligned}$$

This "rule" follows directly from the definition of conditional probability.

Generalization: Suppose $A_1, A_2, ..., A_n$ are events in S. Then

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right).$$

Proof. We use mathematical induction. This is clearly true when $n = 2$ (see above). Assume the result holds for n events. It suffices to show the induction step

$$P\left(\bigcap_{i=1}^{n+1} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right) \times P\left(A_{n+1} \middle| \bigcap_{i=1}^{n} A_i\right).$$

Write

$$P\left(\bigcap_{i=1}^{n+1} A_i\right) = P\left(\bigcap_{i=1}^{n} A_i \cap A_{n+1}\right) = P\left(A_{n+1} \middle| \bigcap_{i=1}^{n} A_i\right) P\left(\bigcap_{i=1}^{n} A_i\right).$$

However, note that

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right)$$

is true by assumption. The result follows immediately. □

Discussion: The multiplication rule allows us to approach calculating the probability of an intersection "sequentially." First, calculate $P(A_1)$ for the first event. Next, calculate $P(A_2|A_1)$ for the second event (given the first). Next, calculate $P(A_3|A_1 \cap A_2)$ for the third event (given the first two), and so on. The next example illustrates this approach.

Example 2.25. I am dealt a hand of 5 cards at random. What is the probability they are all spades?

Solution. Define the events

$$A_i = \{\text{the } i\text{th card is a spade}\}, \quad i = 1, 2, 3, 4, 5.$$

Assuming each card is randomly drawn from the deck,

$$\begin{aligned} P(A_1) &= \frac{13}{52} \\ P(A_2|A_1) &= \frac{12}{51} \\ P(A_3|A_1 \cap A_2) &= \frac{11}{50} \\ P(A_4|A_1 \cap A_2 \cap A_3) &= \frac{10}{49} \\ P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) &= \frac{9}{48}. \end{aligned}$$

Therefore, the probability all five cards are spades is

$$\begin{aligned} P\left(\bigcap_{i=1}^{5} A_i\right) &= P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times P(A_4|A_1 \cap A_2 \cap A_3) \\ &\quad \times P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) \\ &= \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48} \approx 0.0005. \end{aligned}$$

Remark: When I taught this class the last time, a student noted this calculation is easier if you simply regard the cards as belonging to two groups: spades and non-spades. There are $\binom{13}{5}$ ways to draw 5 spades from 13. There are $\binom{52}{5}$ possible hands. Thus, the probability of drawing 5 spades (assuming each hand is equally likely) is $\binom{13}{5}/\binom{52}{5} \approx 0.0005$. $\square$

Terminology: Suppose $A_1, A_2, ..., A_n$ are events in S. We say $A_1, A_2, ..., A_n$ are **mutually independent** if for any sub-collection $A_{i_1}, A_{i_2}, ..., A_{i_k}$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Special case: Take 3 events A_1, A_2, and A_3. For these events to be mutually independent, we need them to be **pairwise independent**:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \end{aligned}$$

and we also need

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3).$$

For $n > 2$, mutual independence is a stronger condition than pairwise independence.

Exercise: Come up with an example of 3 events A_1, A_2, and A_3 that are pairwise independent but not mutually independent. *Hint:* Think of rolling two fair dice with a sample space that regards all $N = 36$ outcomes as being equally likely.

Remark: Many random experiments can be envisioned as consisting of a sequence of n "trials" that are viewed as independent (e.g., flipping a coin 10 times). If A_i denotes the event associated with the ith trial, and the trials are mutually independent, then

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Example 2.26. The State Hygienic Laboratory at the University of Iowa tests thousands of residents for chlamydia every year. Suppose on a given day the lab tests $n = 30$ individual residents. Conceptualizing this as random experiment, the sample space can be written as

$$S = \{(0,0,0,...,0), (1,0,0,...,0), (0,1,0,...,0), ..., (1,1,1,...,1)\},$$

where "0" denotes a negative individual and "1" denotes a positive individual. Note that there are $N = 2^{30} = 1{,}073{,}741{,}824$ outcomes in S. However, these outcomes are probably not equally likely. Define the events

$$A_i = \{i\text{th individual is positive}\}, \quad i = 1, 2, ..., 30.$$

Assume the 30 events $A_1, A_2, ..., A_{30}$ are mutually independent and that $P(A_i) = p$. What is the probability that at least one individual is positive?

Solution. First, note that by using the complement rule, we have $P(\overline{A}_i) = 1 - P(A_i) = 1 - p$, for $i = 1, 2, ..., 30$. Now, the event

$$A = \{\text{at least one individual is positive}\} = \bigcup_{i=1}^{30} A_i.$$

The complement of A is

$$\overline{A} = \{\text{all individuals are negative}\} = \bigcap_{i=1}^{30} \overline{A}_i$$

by DeMorgan's Law. Because $A_1, A_2, ..., A_{30}$ are mutually independent (by assumption), the complements $\overline{A}_1, \overline{A}_2, ..., \overline{A}_{30}$ are also mutually independent. Therefore,

$$P(\overline{A}) = P\left(\bigcap_{i=1}^{30} \overline{A}_i\right) = \prod_{i=1}^{30} P(\overline{A}_i) = (1-p)^{30}.$$

Finally,

$$P(A) = 1 - P(\overline{A}) = 1 - (1-p)^{30}.$$

For example, if $p = 0.01$, then the probability of at least one positive individual among the 30 tested is $P(A) = 0.2603$. $\square$

2.7 Law of Total Probability and Bayes' Rule

Law of Total Probability: Suppose A and B are events in S. We can express A as the union of two mutually exclusive events

$$A = (A \cap B) \cup (A \cap \overline{B}).$$

Therefore, by Axiom 3,

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) \\ &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}). \end{aligned}$$

This is called the **Law of Total Probability**.

Remark: The Law of Total Probability (LOTP) gives us a way to calculate $P(A)$ by relying instead on the conditional probabilities $P(A|B)$ and $P(A|\overline{B})$ and the (unconditional) probability of a related event B. More specifically, $P(A)$ is a linear combination of the conditional probabilities $P(A|B)$ and $P(A|\overline{B})$. The "weights" in the linear combination, $P(B)$ and $P(\overline{B})$, add to 1.

Example 2.27. An insurance company classifies drivers as "accident-prone" and "non-accident-prone." The probability an accident-prone driver has an accident is 0.4. The probability a non-accident-prone driver has an accident is 0.2. The population is estimated to be 30 percent accident-prone.

(a) What is the probability that a policy-holder will have an accident?
(b) If a policy-holder has an accident, what is the probability that s/he was "accident-prone?"

Solutions. Define the events

$$\begin{aligned} A &= \{\text{policy holder has an accident}\} \\ B &= \{\text{policy holder is accident-prone}\}. \end{aligned}$$

We are given $P(A|B) = 0.4$, $P(A|\overline{B}) = 0.2$, and $P(B) = 0.3$.

(a) We want to calculate $P(A)$. By the LOTP,

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) \\ &= 0.4(0.3) + 0.2(0.7) = 0.26. \end{aligned}$$

(b) We want to calculate $P(B|A)$. Note that

$$\begin{aligned} P(B|A) = \frac{P(A \cap B)}{P(A)} &= \frac{P(A|B)P(B)}{P(A)} \\ &= \frac{0.4(0.3)}{0.26} \approx 0.462. \ \square \end{aligned}$$

Note: From Example 2.27(b), we see that, in general,

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

This is a special case of **Bayes' Rule**.

Example 2.28. *Diagnostic testing.* A lab test is 95% effective at detecting a disease when it is present. It is 99% effective at declaring a subject negative when the subject is truly negative for the disease. Suppose 8% of the population has the disease.

(a) What is the probability a randomly selected subject will test positively?
(b) What is the probability a subject has the disease if his test is positive?

Solutions. Define the events

$$\begin{aligned} D &= \{\text{disease is present}\} \\ A &= \{\text{test is positive}\}. \end{aligned}$$

We are given

$$\begin{aligned} P(A|D) &= 0.95 \quad (\text{"sensitivity"}) \\ P(\overline{A}|\overline{D}) &= 0.99 \quad (\text{"specificity"}) \\ P(D) &= 0.08 \quad (\text{"prevalence"}). \end{aligned}$$

(a) We want $P(A)$. By the LOTP,

$$\begin{aligned} P(A) &= P(A|D)P(D) + P(A|\overline{D})P(\overline{D}) \\ &= 0.95(0.08) + 0.01(0.92) \approx 0.0852. \end{aligned}$$

(b) We want $P(D|A)$. By Bayes' Rule,

$$\begin{aligned} P(D|A) &= \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|\overline{D})P(\overline{D})} \\ &= \frac{0.95(0.08)}{0.95(0.08) + 0.01(0.92)} \approx 0.892. \end{aligned}$$

Remark: Bayes' Rule allows us to "update" probabilities on the basis of observed information (in Example 2.28, this "observed information" is the test result):

Prior probability		Test result		**Posterior** probability
$P(D) = 0.08$	$\longrightarrow$	A	$\longrightarrow$	$P(D\|A) \approx 0.892$
$P(D) = 0.08$	$\longrightarrow$	$\overline{A}$	$\longrightarrow$	$P(D\|\overline{A}) \approx 0.004$

Note: $P(D|A)$ in this example is called the "positive predictive value" (PPV). Calculate $P(\overline{D}|\overline{A})$, the "negative predictive value" (NPV). □

Remark: For two events A and B, the formulas for LOTP and Bayes' Rule are given below:

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) \\ P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})}. \end{aligned}$$

Both of these formulas arise because the sample space S can be written as $S = B \cup \overline{B}$, the union of two mutually exclusive events. In other words, the events B and $\overline{B}$ "partition" the sample space. We now generalize LOTP and Bayes' Rule for an arbitrary partition of S.

Terminology: A collection of events $B_1, B_2, ..., B_k$ forms a **partition** of the sample space S if

$$\bigcup_{i=1}^{k} B_i = B_1 \cup B_2 \cup \cdots \cup B_k = S$$

and $B_i \cap B_j = \emptyset$, for $i \neq j$.

LOTP: Suppose A is an event in S and suppose $B_1, B_2, ..., B_k$ forms a partition of S. Then

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$

Proof. The event A can be written as

$$A = A \cap S = A \cap \bigcup_{i=1}^{k} B_i = \bigcup_{i=1}^{k} (A \cap B_i).$$

Because $B_1, B_2, ..., B_k$ partition S, the events $A \cap B_1, A \cap B_2, ..., A \cap B_k$ are pairwise mutually exclusive. Therefore,

$$P(A) = P\left(\bigcup_{i=1}^{k} (A \cap B_i)\right) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i),$$

the last step following from the multiplication rule. □

Bayes' Rule: Suppose A is an event in S and suppose $B_1, B_2, ..., B_k$ forms a partition of S. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

Proof. From the definition of conditional probability and the multiplication rule, note that

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{P(A)}.$$

Now just write $P(A)$ out in its LOTP expansion. □

Example 2.29. For policy holders of a certain age, a life insurance company issues standard, preferred, and ultra-preferred policies. Among these policy holders,

- 60 percent are standard with a probability of 0.05 of dying next year.
- 30 percent are preferred with a probability of 0.03 of dying next year.
- 10 percent are ultra-preferred with a probability of 0.01 of dying next year.

(a) What is the probability a policy holder of this certain age dies next year?
(b) A policy holder of this certain age dies next year. What is the probability the deceased was a preferred policy holder?
(c) A policy holder of this certain age does not die next year. What is the probability this policy holder is an ultra-preferred policy holder?

Solutions. Define the events $A = \{\text{policy holder dies next year}\}$ and

$$\begin{aligned} B_1 &= \{\text{policy holder has standard policy}\} \\ B_2 &= \{\text{policy holder has preferred policy}\} \\ B_3 &= \{\text{policy holder has ultra-preferred policy}\}. \end{aligned}$$

Note that $\{B_1, B_2, B_3\}$ partition the sample space with $P(B_1) = 0.60$, $P(B_2) = 0.30$, and $P(B_3) = 0.10$. We are also given $P(A|B_1) = 0.05$, $P(A|B_2) = 0.03$, and $P(A|B_3) = 0.01$.

(a) We want $P(A)$. By the LOTP,

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= 0.05(0.60) + 0.03(0.30) + 0.01(0.10) = 0.04. \end{aligned}$$

(b) We want $P(B_2|A)$. By Bayes' Rule,

$$\begin{aligned} P(B_2|A) &= \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{0.03(0.30)}{0.05(0.60) + 0.03(0.30) + 0.01(0.10)} = 0.225. \end{aligned}$$

Note how the "prior probability" $P(B_2) = 0.30$ has changed to $P(B_2|A) = 0.225$ when we learn that A has occurred.

(c) We want $P(B_3|\overline{A})$. By Bayes' Rule,

$$P(B_3|\overline{A}) = \frac{P(\overline{A}|B_3)P(B_3)}{P(\overline{A})} = \frac{[1 - P(A|B_3)]P(B_3)}{1 - P(A)} = \frac{(1 - 0.01)(0.10)}{1 - 0.04} \approx 0.103.$$

Note how the "prior probability" $P(B_3) = 0.10$ has changed to $P(B_3|\overline{A}) \approx 0.103$ when we learn that $\overline{A}$ has occurred.

3 Discrete Random Variables and their Probability Distributions

3.1 Introduction

Recall: In Example 2.26 (pp 30, notes), we considered the problem of testing 30 Iowa residents for chlamydia. Conceptualizing this as random experiment, we wrote the sample space as

$$S = \{(0,0,0,...,0),(1,0,0,...,0),(0,1,0,...,0),...,(1,1,1,...,1)\},$$

where "0" denotes a negative individual and "1" denotes a positive individual. Note that there are $N = 2^{30} = 1{,}073{,}741{,}824$ outcomes in S.

Remark: Keeping track of outcomes in large unwieldy sample spaces like this is not practical. In this and other random experiments, it is much easier to reduce each outcome (sample point) to a numerical value.

Terminology: A **random variable** Y is a function whose domain is the sample space S and whose range is the set of real numbers $\mathbb{R} = (-\infty, \infty)$. That is,

$$Y : S \to \mathbb{R}$$

takes outcomes (sample points) in S and assigns them a real number.

Note: In the example above, define

$$Y = \text{ number of positives (out of 30).}$$

Thinking of Y as a function, we see that, for example,

$$\begin{aligned} Y((0,0,0,...,0)) &= 0 \\ Y((1,0,0,...,0)) &= 1 \\ Y((1,1,0,...,0)) &= 2 \\ Y((1,1,1,...,1)) &= 30. \end{aligned}$$

The domain of Y is all 1,073,741,824 outcomes in S. The range of Y is

$$R = \{0,1,2,3,...,30\}.$$

Terminology: The **support** of a random variable Y is the set of all possible values that Y can assume; i.e., it is the range of Y under the mapping $Y : S \to \mathbb{R}$. We will denote the support by R. It is understood that $R \subseteq \mathbb{R}$.

Terminology: We call a random variable Y **discrete** if its support R is a finite or countable set. In other words, Y can assume a finite or (at most) a countable number of values.

Example 3.1. Consider the random experiment of rolling two dice and observing the face on each. The sample space for this experiment is

$$\begin{aligned} S = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}. \end{aligned}$$

Assume the dice are "fair" so that each outcome (sample point) is equally likely; i.e., each outcome has probability 1/36.

Define the random variable

$$Y = \text{ sum of the two faces.}$$

The support of Y is

$$R = \{2,3,4,5,6,7,8,9,10,11,12\}.$$

Because R is a finite set, Y is a discrete random variable.

Q: How do we calculate probabilities like $P(Y = 2)$? like $P(Y = 7)$? like $P(Y = 21)$?
A: The "first principles" approach to doing this would be to find the **inverse image** of events like $\{Y = 2\}$, $\{Y = 7\}$, and $\{Y = 21\}$ back on the original sample space S and then carry out the calculations there. For example,

$$P(Y = 2) = P(\{(1,1)\}) = \frac{1}{36}.$$

Similarly,

$$P(Y = 7) = P(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = \frac{6}{36}$$

and

$$P(Y = 21) = P(\emptyset) = 0.$$

Important: In general, the probability Y takes on the value y, written $P(Y = y)$, is the sum of the probabilities of the outcomes (sample points) in S that are assigned the value y under the mapping $Y : S \to \mathbb{R}$. In notation,

$$P(Y = y) = P(\{\text{all } \omega \in S : Y(\omega) = y\}) = \sum_{\substack{\omega \in S \\ Y(\omega) = y}} P(\{\omega\}),$$

where recall ω denotes an outcome (sample point) in S.

Terminology: The **probability mass function (pmf)** of a discrete random variable Y is the function defined by

$$p_Y(y) = P(Y = y), \quad \text{for all } y.$$

If the value y is not in the support R, then it is understood that $p_Y(y) = 0$. A discrete random variable's pmf describes its **probability distribution**.

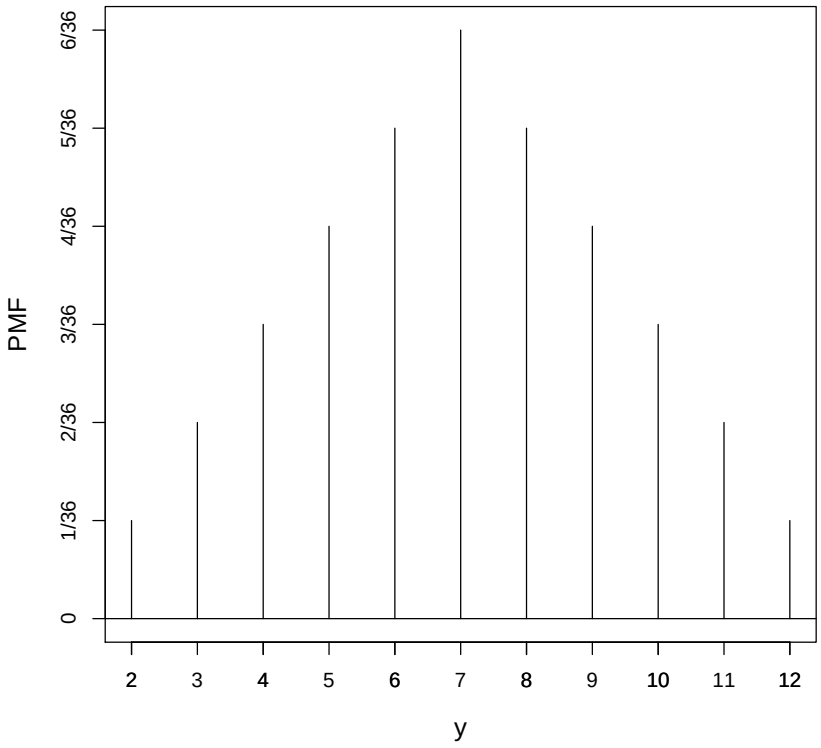

Figure 3.1: Probability mass function (pmf) of Y in Example 3.1.

Important: The pmf of Y in Example 3.1 (and in other examples) can be described by using a table, a graph, or a formula. In tabular form, we can write

y	2	3	4	5	6	7	8	9	10	11	12
$p_Y(y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

A graph of the pmf of Y is shown in Figure 3.1 above. Finally, it is also possible to represent the pmf of Y as a formula; i.e.,

$$p_Y(y) = \begin{cases} \frac{1}{36}\left(6 - |7-y|\right), & y = 2, 3, ..., 12 \\ 0, & \text{otherwise.} \end{cases}$$

Exercise: In Example 3.1, find the pmf of

$$Y = \text{ absolute difference of the two faces.}$$

For example, $Y((1,1)) = |1-1| = 0$, $Y((1,2)) = |1-2| = 1$, and so on. Depict your pmf in a table and a graph like above. $\square$

Properties: The pmf of a discrete random variable Y has the following properties:

1. $0 \leq p_Y(y) \leq 1$, for all y
2. the sum of the probabilities over all y equals 1; i.e.,

$$\sum_{y \in R} p_Y(y) = 1.$$

These properties arise as a consequence of Axioms 1 and 2 (see pp 9, notes).

Example 3.2. I recently had a flight from Washington DC to Columbia. The plane had 66 seats on it and each seat was occupied; there were 36 females and 30 males on the flight. Suppose I selected 5 passengers at random and recorded

$$Y = \text{ number of males (out of 5).}$$

Find the pmf of Y.

Solution. We can think of one outcome (sample point) in the underlying sample space S as having the following structure:

$$(\ \_\_\ \_\_\ \_\_\ \_\_\ \_\_\ ).$$

For example, the outcomes

$$(\ \mathrm{F}_1\ \mathrm{F}_2\ \mathrm{F}_3\ \mathrm{F}_4\ \mathrm{F}_5\ ) \quad \text{and} \quad (\ \mathrm{M}_1\ \mathrm{F}_1\ \mathrm{F}_2\ \mathrm{M}_2\ \mathrm{M}_3\ ),$$

would produce the values $y = 0$ and $y = 3$, respectively. Note that there are

$$N = \binom{66}{5} = 8936928$$

outcomes in the sample space S (the ordering of passenger selection doesn't matter).

The pmf of Y is the function $p_Y(y) = P(Y = y)$, which is nonzero when $y = 0, 1, 2, 3, 4, 5$. The number of outcomes in S with y males can be found by using the basic rule of counting:

$$\begin{aligned} n_1 &= \text{number of ways to select } y \text{ males from } 30 = \binom{30}{y} \\ n_2 &= \text{number of ways to select } 5-y \text{ females from } 36 = \binom{36}{5-y} \end{aligned}$$

Therefore, there are

$$\binom{30}{y}\binom{36}{5-y}$$

outcomes in S with y males. Assuming each outcome is equally likely,

$$p_Y(y) = \begin{cases} \dfrac{\binom{30}{y}\binom{36}{5-y}}{\binom{66}{5}}, & y = 0, 1, 2, 3, 4, 5 \\ 0, & \text{otherwise.} \end{cases}$$

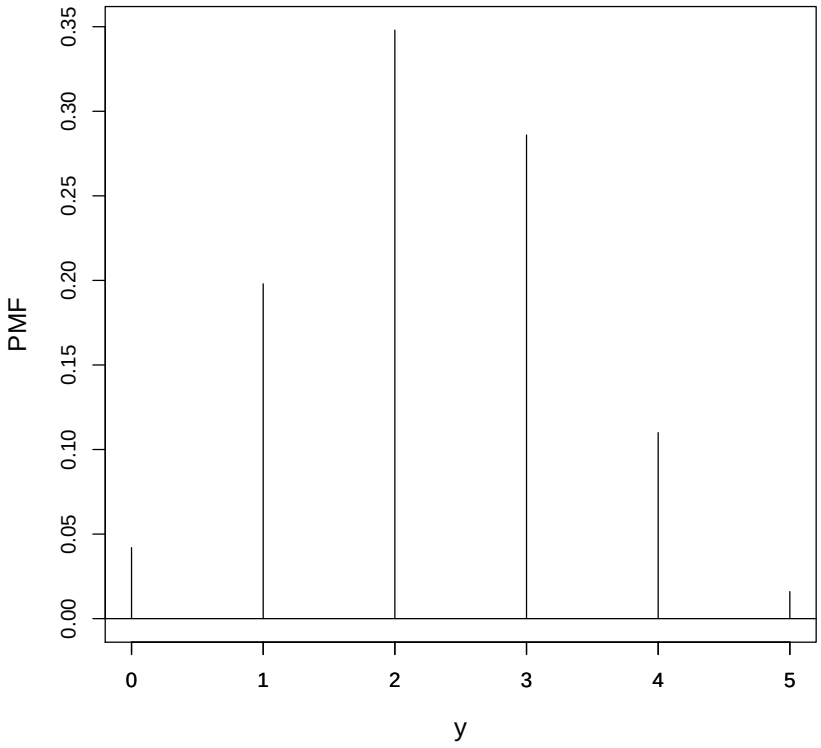

Figure 3.2: Probability mass function (pmf) of Y in Example 3.2.

Here are the probabilities $p_Y(y) = P(Y = y)$ listed out in a table (to 3 dp):

y	0	1	2	3	4	5
$p_Y(y)$	0.042	0.198	0.348	0.286	0.110	0.016

Note that these probabilities sum to 1, as required. A graph of the pmf of Y is shown in Figure 3.2 above.

Q: What is the probability I select at least 4 males?
A: We can work directly from the pmf:

$$\begin{aligned} P(Y \geq 4) = P(Y = 4) + P(Y = 5) &= p_Y(4) + p_Y(5) \\ &\approx 0.110 + 0.016 = 0.126. \ \square \end{aligned}$$

This example illustrates the following general result.

Result: Suppose Y is a **discrete** random variable with pmf $p_Y(y)$. The probability of an event $\{Y \in B\}$ is found by adding the probabilities $p_Y(y)$ for all $y \in B$; i.e.,

$$P(Y \in B) = \sum_{y \in B} p_Y(y).$$

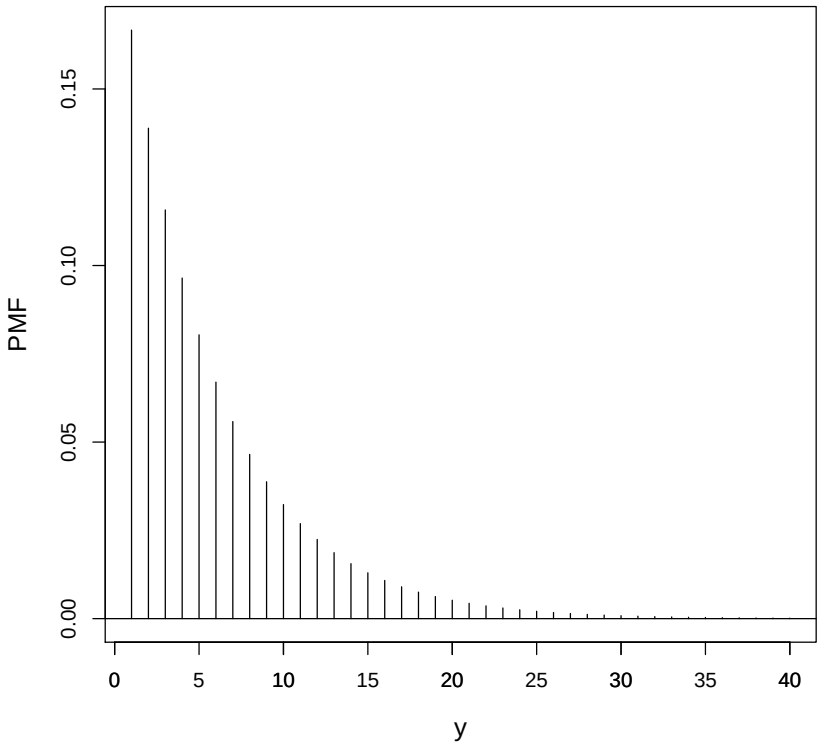

Figure 3.3: Probability mass function (pmf) of Y in Example 3.3.

Example 3.3. An experiment consists of rolling an unbiased die until the first "6" is observed. Let Y denote the number of rolls needed. The pmf of Y is given by

$$p_Y(y) = \begin{cases} \frac{1}{6}\left(\frac{5}{6}\right)^{y-1}, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

A graph of the pmf of Y is shown in Figure 3.3 above.

Q: Is this a valid pmf?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 1, 2, 3, ...$ Do the probabilities $p_Y(y)$ sum to 1?

Recall: If $a \in \mathbb{R}$ and $|r| < 1$, then

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}.$$

This is the formula for an **infinite geometric sum**. The condition $|r| < 1$ is needed or else the sum does not converge.

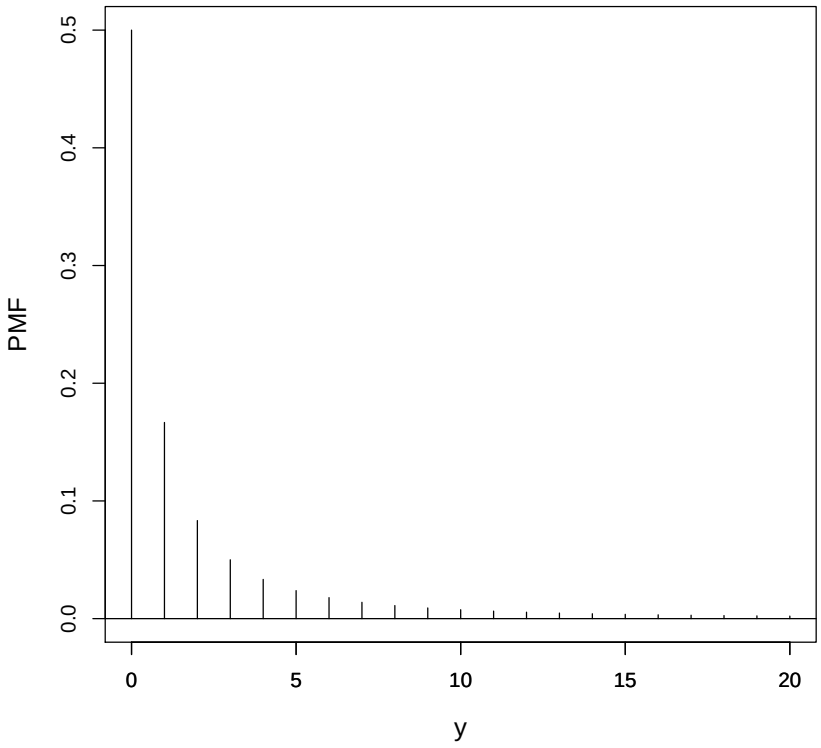

Figure 3.4: Probability mass function (pmf) of Y in Example 3.4.

Note that

$$\sum_{y=1}^{\infty} p_Y(y) = \sum_{y=1}^{\infty} \frac{1}{6}\left(\frac{5}{6}\right)^{y-1} = \sum_{k=0}^{\infty} \frac{1}{6}\left(\frac{5}{6}\right)^{k} = \frac{\frac{1}{6}}{1-\frac{5}{6}} = 1.$$

Therefore, the pmf $p_Y(y)$ is valid. $\square$

Example 3.4. An insurance company models the number of claims per day, Y, as a discrete random variable with pmf

$$p_Y(y) = \begin{cases} \dfrac{1}{(y+1)(y+2)}, & y = 0, 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

A graph of the pmf of Y is shown in Figure 3.4 above.

Q: Is this a valid pmf?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 0, 1, 2, 3, ...$ Do the probabilities $p_Y(y)$ sum to 1? Note that we can rewrite

$$\frac{1}{(y+1)(y+2)} = \frac{1}{y+1} - \frac{1}{y+2}.$$

It follows that $\sum_{y=0}^{\infty} p_Y(y)$ is a **telescoping sum**; i.e.,

$$\begin{aligned}\sum_{y=0}^{\infty} p_Y(y) = \sum_{y=0}^{\infty} \frac{1}{(y+1)(y+2)} &= \sum_{y=0}^{\infty}\left(\frac{1}{y+1} - \frac{1}{y+2}\right) \\ &= \left(1-\frac{1}{2}\right) + \left(\frac{1}{2}-\frac{1}{3}\right) + \left(\frac{1}{3}-\frac{1}{4}\right) + \left(\frac{1}{4}-\frac{1}{5}\right) + \cdots = 1.\end{aligned}$$

Therefore, the pmf $p_Y(y)$ is valid. □

3.2 Mathematical expectation

3.2.1 Expected value

Terminology: Suppose Y is a discrete random variable with pmf $p_Y(y)$ and support R. The **expected value** of Y is

$$E(Y) = \sum_{y \in R} y p_Y(y).$$

In other words, $E(Y)$ is a weighted average of the possible values of Y. Each $y \in R$ is weighted by its corresponding probability $p_Y(y)$.

Technical note: If the support R is countable but not finite, then $E(Y)$ may not exist. This occurs when the sum above does not converge absolutely. In other words, for $E(Y)$ to exist, we need

$$\sum_{y \in R} |y| p_Y(y) < \infty.$$

Of course, if R is a finite set, then the sum $\sum_{y \in R} |y| p_Y(y)$ is finite and hence $E(Y)$ exists.

Exercise: Show that $E(Y)$ in Example 3.4 does not exist.

Example 3.5. Patient responses to a generic drug to control pain are scored on 5-point scale (1 = lowest pain level; 5 = highest pain level). In a certain population of patients, the pmf of the response Y is given by

y	1	2	3	4	5
$p_Y(y)$	0.38	0.27	0.18	0.11	0.06

A graph of the pmf of Y is shown in Figure 3.5 (next page). The expected value of Y is

$$\begin{aligned}E(Y) &= \sum_{y=1}^{5} y p_Y(y) \\ &= 1(0.38) + 2(0.27) + 3(0.18) + 4(0.11) + 5(0.06) = 2.2.\end{aligned}$$

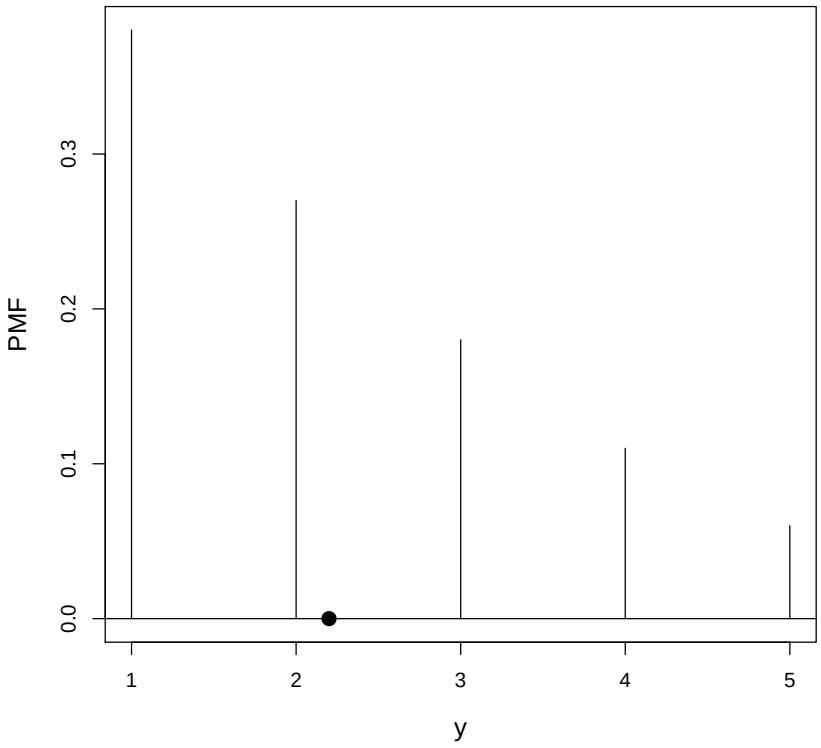

Figure 3.5: Pmf of Y in Example 3.5. A solid circle identifies where $E(Y)$ falls.

Interpretations: The expected value, or **mean**, of Y can be interpreted in different ways:

- $E(Y)$ is the "center of gravity" on the pmf of Y (see above). It's located where the pmf would balance.
- $E(Y)$ is a "long run average." In other words, if we observed the value of Y over and over again (e.g., for a large number of patients in Example 3.5), then the average value would be close to $E(Y)$.

To illustrate this last interpretation, I used R's `sample` function to sample 1000 values of Y according to the pmf in Example 3.5:

```
y = c(1,2,3,4,5)
prob = c(0.38,0.27,0.18,0.11,0.06)
sample.values = sample(y,1000,replace=TRUE,prob=prob)
> mean(sample.values)
[1] 2.203
```

The mean of these 1000 values was 2.203, which is very close to $E(Y) = 2.2$. □

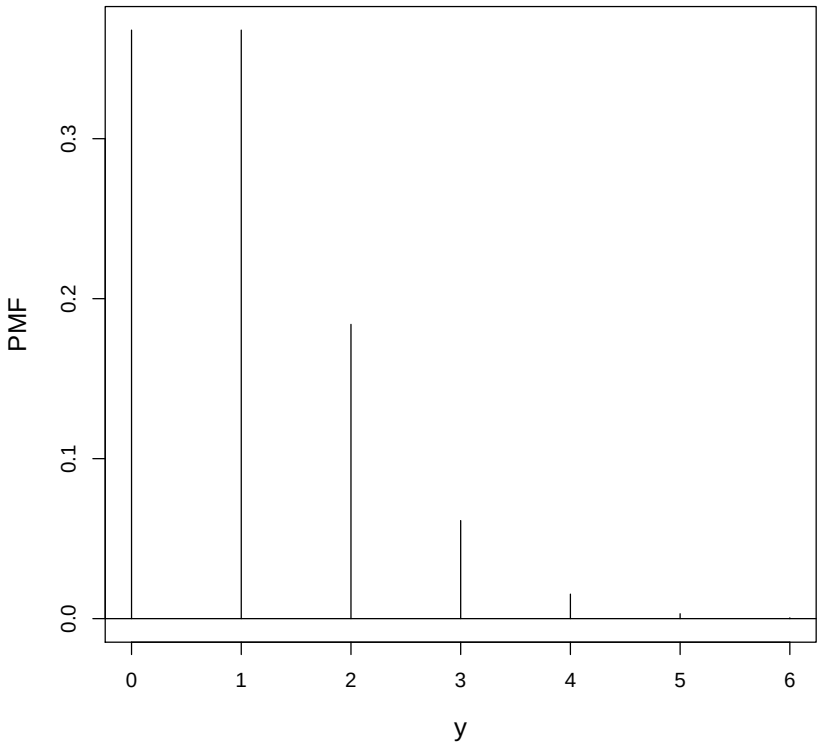

Figure 3.6: Probability mass function (pmf) of Y in Example 3.6.

Remark: In statistical applications, the expected value $E(Y)$ is called the **population mean**. This is the average value of Y that would result from measuring every individual in the population (provided, of course, that we could and that the pmf of Y was an accurate model for the population).

Example 3.6. An entomologist records Y, the number of insects that occupy a test plant. The pmf of Y is given by

$$p_Y(y) = \begin{cases} \dfrac{e^{-1}}{y!}, & y = 0, 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

A graph of the pmf of Y is shown in Figure 3.6 above. Find $E(Y)$.

Solution. The expected value of Y is

$$E(Y) = \sum_{y=0}^{\infty} y p_Y(y) = \sum_{y=0}^{\infty} y \, \frac{e^{-1}}{y!} = \sum_{y=1}^{\infty} y \, \frac{e^{-1}}{y!} = \sum_{y=1}^{\infty} \frac{e^{-1}}{(y-1)!} = e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!} = e^{-1}e^{1} = 1. \; \square$$

Recall: The McLaurin series expansion of $f(x) = e^x$ is

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots .$$

This expansion is valid for all $x \in \mathbb{R}$. For example, when $x = 1$, we have

$$e = e^1 = \sum_{k=0}^{\infty} \frac{1^k}{k!} = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

In general, recall that the Taylor series expansion of the function $f(x)$ about the point $x = a$ is given by

$$\begin{aligned} f(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(a)(x-a)^k}{k!} \\ &= f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f^{(3)}(a)(x-a)^3 + \frac{1}{24}f^{(4)}(a)(x-a)^4 + \cdots . \end{aligned}$$

A McLaurin series expansion is a Taylor series expansion when $a = 0$.

Exercise: Write out $f(x) = \ln(1+x)$, $f(x) = \cos x$, and $f(x) = 1/(1-x)$ in their McLaurin series expansions. Note for which values of $x \in \mathbb{R}$ the expansion is valid.

Example 3.7. *Discrete uniform distribution.* Suppose the random variable Y has pmf

$$p_Y(y) = \begin{cases} \frac{1}{N}, & y = 1, 2, ..., N \\ 0, & \text{otherwise,} \end{cases}$$

where N is a positive integer larger than 1. Find $E(Y)$.

Solution. The expected value of Y is

$$E(Y) = \sum_{y=1}^{N} y\left(\frac{1}{N}\right) = \frac{1}{N}\sum_{y=1}^{N} y = \frac{1}{N}\left[\frac{N(N+1)}{2}\right] = \frac{N+1}{2}.$$

Here, we have used the well known fact that the sum of the first N positive integers; i.e.,

$$\sum_{y=1}^{N} y = 1 + 2 + 3 + \cdots + N = \frac{N(N+1)}{2}.$$

This can be proven using induction. If $N = 6$, then the discrete uniform distribution applies for the outcome of a fair die:

y	1	2	3	4	5	6
$p_Y(y)$	1/6	1/6	1/6	1/6	1/6	1/6

The expected value of Y is $E(Y) = (6+1)/2 = 3.5$. □

3.2.2 Functions of random variables

Terminology: Suppose Y is a discrete random variable with pmf $p_Y(y)$ and support R. The **expected value** of $g(Y)$ is

$$E[g(Y)] = \sum_{y \in R} g(y) p_Y(y).$$

In other words, $E[g(Y)]$ is a weighted average of the possible values of $g(Y)$, where the probabilities $p_Y(y)$ play the role of the weights.

Technical note: If the support R is countable but not finite, then $E[g(Y)]$ may not exist. This occurs when the sum above does not converge absolutely. In other words, for $E[g(Y)]$ to exist, we need

$$\sum_{y \in R} |g(y)| p_Y(y) < \infty.$$

Example 3.8. In Example 3.5, we used the pmf below to describe the population of patients' responses to a generic drug to control pain:

y	1	2	3	4	5
$p_Y(y)$	0.38	0.27	0.18	0.11	0.06

Find $E(Y^2)$, $E(\sqrt{Y})$, and $E(e^{tY})$, where t is a constant.

Solutions. Using the result above, we have

$$E(Y^2) = 1^2(0.38) + 2^2(0.27) + 3^2(0.18) + 4^2(0.11) + 5^2(0.06) = 6.34$$

$$E(\sqrt{Y}) = \sqrt{1}(0.38) + \sqrt{2}(0.27) + \sqrt{3}(0.18) + \sqrt{4}(0.11) + \sqrt{5}(0.06) \approx 1.43$$

and

$$\begin{aligned} E(e^{tY}) &= e^{t(1)}(0.38) + e^{t(2)}(0.27) + e^{t(3)}(0.18) + e^{t(4)}(0.11) + e^{t(5)}(0.06) \\ &= 0.38e^t + 0.27e^{2t} + 0.18e^{3t} + 0.11e^{4t} + 0.06e^{5t}. \ \square \end{aligned}$$

Properties: In general, the expectation operator $E(\cdot)$ has certain properties. First, the expected value of a **constant** c is c; i.e.,

$$E(c) = c.$$

This is easy to show when Y is discrete with pmf $p_Y(y)$; note that

$$E(c) = \sum_{y \in R} c p_Y(y) = c \sum_{y \in R} p_Y(y) = c(1) = c.$$

Second, multiplicative constants can be moved outside the expectation; i.e.,

$$E[cg(Y)] = cE[g(Y)].$$

This is also easy to prove provided that $E[g(Y)]$ exists; note that

$$E[cg(Y)] = \sum_{y \in R} cg(y)p_Y(y) = c\sum_{y \in R} g(y)p_Y(y) = cE[g(Y)].$$

Finally, taking expectations is **additive**; i.e.,

$$E\left[\sum_{j=1}^{k} g_j(Y)\right] = \sum_{j=1}^{k} E[g_j(Y)],$$

provided that $E[g_j(Y)]$ exists for each $j = 1, 2, ..., k$. Note that

$$\begin{aligned} E\left[\sum_{j=1}^{k} g_j(Y)\right] &= \sum_{y \in R}\sum_{j=1}^{k} g_j(y)p_Y(y) \\ &= \sum_{y \in R} g_1(y)p_Y(y) + \sum_{y \in R} g_2(y)p_Y(y) + \cdots + \sum_{y \in R} g_k(y)p_Y(y) \\ &= E[g_1(Y)] + E[g_2(Y)] + \cdots + E[g_k(Y)] = \sum_{j=1}^{k} E[g_j(Y)]. \end{aligned}$$

These are called the **linearity properties** of the expectation.

Another useful result: If $g(y) \geq 0$ for all y, then $E[g(Y)] \geq 0$. In other words, random variables that are nonnegative have nonnegative expectations.

3.2.3 Variance

Terminology: Suppose Y is a discrete random variable with mean $E(Y) = \mu$. The **variance** of Y is

$$\sigma^2 = V(Y) = E[(Y - \mu)^2] = \sum_{y \in R} (y - \mu)^2 p_Y(y).$$

In other words, $V(Y)$ is a weighted average of the possible values of $g(Y) = (Y - \mu)^2$, where the probabilities $p_Y(y)$ play the role of the weights.

Note: The variance $V(Y)$ is the expected value of a "special" function of Y, namely $g(Y) = (Y - \mu)^2$. Similar technical requirements arise regarding existence; i.e., we need

$$\sum_{y \in R} (y - \mu)^2 p_Y(y) < \infty$$

for $V(Y)$ to exist.

Terminology: The **standard deviation** of Y is the (positive) square root of the variance; i.e.,

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(Y)}.$$

Example 3.9. In Example 3.5, we used the pmf below to describe the population of patients' responses to a generic drug to control pain:

y	1	2	3	4	5
$p_Y(y)$	0.38	0.27	0.18	0.11	0.06

Calculate $\sigma^2 = V(Y)$ and the standard deviation σ.

Solution. In Example 3.5, we calculated

$$E(Y) = \mu = 2.2.$$

Therefore, the variance of Y is

$$\begin{aligned}\sigma^2 &= \sum_{y=1}^{5}(y-\mu)^2 p_Y(y) \\ &= (1-2.2)^2(0.38) + (2-2.2)^2(0.27) + (3-2.2)^2(0.18) \\ &\qquad + (4-2.2)^2(0.11) + (5-2.2)^2(0.06) = 1.5.\end{aligned}$$

The standard deviation of Y is

$$\sigma = \sqrt{1.5} \approx 1.22.$$

Properties: The variance of a discrete random variable Y has the following properties and interpretations:

1. The variance is nonnegative; i.e., $V(Y) \geq 0$. This is easy to see because

$$V(Y) = E[(Y-\mu)^2]$$

and $g(y) = (y-\mu)^2 \geq 0$ for all y; see the last result at the end of Section 3.2.2 (notes). Clearly $\sigma \geq 0$ as well.

2. Can $V(Y)$ ever be zero? It can, but only when all of the probability mass for Y resides at one point, namely $y = \mu$. A random variable Y with this property is called a **degenerate** random variable. Any constant c can be thought of as a degenerate random variable.

3. Whereas the expected value $E(Y) = \mu$ measures the "center" or the "balance point" of a distribution, the variance $V(Y) = \sigma^2$ (and the standard deviation) measures the "spread" in the distribution. The larger $V(Y)$ is, the larger the spread.

4. The variance $V(Y) = \sigma^2$ is measured in the squared units of Y. The standard deviation σ is measured in the same units as Y. Because of this, the standard deviation is easier for interpretation purposes.

Variance computing formula: Suppose Y is a (discrete) random variable with mean $E(Y) = \mu$. The variance of Y can be calculated as

$$V(Y) = E(Y^2) - [E(Y)]^2.$$

Proof. Using the definition of $V(Y)$, we have

$$\begin{aligned} V(Y) = E[(Y-\mu)^2] &= E(Y^2 - 2\mu Y + \mu^2) \\ &= E(Y^2) - 2\mu E(Y) + \mu^2 \\ &= E(Y^2) - \mu^2 \\ &= E(Y^2) - [E(Y)]^2. \ \square \end{aligned}$$

Remark: The variance computing formula is helpful because you only need to have $E(Y)$ and $E(Y^2)$ to find $V(Y)$. Note that, in general,

$$E(Y^2) \neq [E(Y)]^2.$$

The only time this is true is when $V(Y) = 0$; i.e., Y is a degenerate random variable.

Example 3.10. *Discrete uniform distribution.* Suppose the random variable Y has pmf

$$p_Y(y) = \begin{cases} \dfrac{1}{N}, & y = 1, 2, ..., N \\ 0, & \text{otherwise,} \end{cases}$$

where N is a positive integer larger than 1. Find $V(Y)$.

Solution. We showed in Example 3.7 that

$$E(Y) = \frac{N+1}{2}.$$

Therefore, we only need to find $E(Y^2)$. From the definition of expectation, we have

$$E(Y^2) = \sum_{y=1}^{N} y^2 \left(\frac{1}{N}\right) = \frac{1}{N}\sum_{y=1}^{N} y^2 = \frac{1}{N}\left[\frac{N(N+1)(2N+1)}{6}\right] = \frac{(N+1)(2N+1)}{6}.$$

Here, we have used the well known fact that

$$\sum_{y=1}^{N} y^2 = 1^2 + 2^2 + 3^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}.$$

Therefore, from the variance computing formula, we have

$$\begin{aligned} V(Y) = E(Y^2) - [E(Y)]^2 &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{N^2-1}{12}. \ \square \end{aligned}$$

Result: Suppose Y is a (discrete) random variable and a and b are constants. Then

$$V(a+bY) = b^2V(Y).$$

Taking $b = 0$, we see that $V(a) = 0$ for any constant a. This makes sense intuitively. The variance is a measure of variability for a random variable; a constant (such as a) does not vary. Also, by taking $a = 0$, we see that $V(bY) = b^2V(Y)$.

3.3 Moment-generating functions

Terminology: The kth **moment** of a (discrete) random variable Y is

$$\mu_k' = E(Y^k).$$

For example, the first four moments are

$$\begin{aligned} E(Y) &= \text{1st moment} \\ E(Y^2) &= \text{2nd moment} \\ E(Y^3) &= \text{3rd moment} \\ E(Y^4) &= \text{4th moment} \end{aligned}$$

Remark: Note that the first moment $E(Y)$ is simply the expected value (or mean) or Y, which describes the "center" of the distribution of Y. Recall that

$$V(Y) = E(Y^2) - [E(Y)]^2$$

so the first two moments can be used to find $V(Y)$, which describes the "spread" in the distribution of Y.

Terminology: Suppose Y is a discrete random variable with pmf $p_Y(y)$ and support R. The **moment-generating function (mgf)** of Y is

$$m_Y(t) = E(e^{tY}) = \sum_{y \in R} e^{ty} p_Y(y),$$

provided this expectation is finite for all t in an open neighborhood about $t = 0$; i.e., $\exists b > 0$ such that $E(e^{tY}) < \infty \; \forall t \in (-b, b)$. If no such $b > 0$ exists, then the moment generating function of Y does not exist.

Example 3.11. In Example 3.6, we considered the discrete random variable Y with pmf

$$p_Y(y) = \begin{cases} \dfrac{e^{-1}}{y!}, & y = 0, 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

Find the mgf of Y.

Solution. The mgf of Y is

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty} \frac{e^{-1}}{y!} = e^{-1} \sum_{y=0}^{\infty} \frac{(e^t)^y}{y!} = e^{-1} \exp(e^t) = \exp(e^t - 1).$$

Above we used the fact that $\sum_{y=0}^{\infty} (e^t)^y/y!$ is the McLaurin series expansion of $\exp(e^t)$, which is a valid expansion for all $t \in \mathbb{R}$.

Example 3.12. Suppose Y is a discrete random variable with pmf

$$p_Y(y) = \begin{cases} \left(\frac{1}{2}\right)^y, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

Find the mgf of Y.

Solution. The mgf of Y is

$$m_Y(t) = E(e^{tY}) = \sum_{y=1}^{\infty} e^{ty} \left(\frac{1}{2}\right)^y = \sum_{y=1}^{\infty} \left(\frac{e^t}{2}\right)^y = \left[\sum_{y=0}^{\infty} \left(\frac{e^t}{2}\right)^y\right] - 1.$$

Note that the sum

$$\sum_{y=0}^{\infty} \left(\frac{e^t}{2}\right)^y$$

is an infinite geometric sum with common ratio

$$r = \frac{e^t}{2} < 1 \iff t < \ln 2.$$

Therefore, for all $t < \ln 2$, this sum converges and hence

$$\begin{aligned} m_Y(t) = \left[\sum_{y=0}^{\infty} \left(\frac{e^t}{2}\right)^y\right] - 1 &= \frac{1}{1 - \frac{e^t}{2}} - 1 \\ &= \frac{2}{2 - e^t} - 1 = \frac{e^t}{2 - e^t}. \end{aligned}$$

Q: Why are mgfs useful?
A: Moment generating functions are functions that generate moments.

Important: If Y is a random variable with mgf $m_Y(t)$, then

$$E(Y^k) = m_Y^{(k)}(0),$$

where

$$m_Y^{(k)}(0) = \frac{d^k}{dt^k} m_Y(t) \bigg|_{t=0}.$$

This shows how the moments of Y can be found by differentiation. Note that derivatives are taken with respect to t.

Proof. Assume Y is a discrete random variable with pmf $p_Y(y)$ and support R. For $k = 1$,

$$\begin{aligned}\frac{d}{dt}m_Y(t) = \frac{d}{dt}\sum_{y\in R} e^{ty}p_Y(y) &\stackrel{?}{=} \sum_{y\in R}\frac{d}{dt}e^{ty}p_Y(y) \\ &= \sum_{y\in R} ye^{ty}p_Y(y) = E(Ye^{tY}).\end{aligned}$$

Thus,

$$\frac{d}{dt}m_Y(t)\bigg|_{t=0} = E(Ye^{tY})\Big|_{t=0} = E(Y).$$

Taking higher-order derivatives, it follows that

$$\frac{d^k}{dt^k}m_Y(t)\bigg|_{t=0} = E(Y^k),$$

for any integer $k \geq 1$. □

Remark: In the argument above, we needed to assume that the interchange of the derivative and sum is justified. When the mgf exists, this interchange is justified.

Interesting: Writing $m_Y(t)$ in its McLaurin series expansion, we see that

$$\begin{aligned}m_Y(t) &= m_Y(0) + \frac{m_Y^{(1)}(0)}{1!}(t-0) + \frac{m_Y^{(2)}(0)}{2!}(t-0)^2 + \frac{m_Y^{(3)}(0)}{3!}(t-0)^3 + \cdots \\ &= 1 + E(Y)t + \frac{E(Y^2)}{2}t^2 + \frac{E(Y^3)}{6}t^3 + \frac{E(Y^4)}{24}t^4 + \cdots \\ &= \sum_{k=0}^{\infty}\frac{E(Y^k)}{k!}t^k.\end{aligned}$$

You can also see that

$$E(Y^k) = \frac{d^k}{dt^k}m_Y(t)\bigg|_{t=0}$$

by differentiating the RHS of $m_Y(t)$ written in its expansion (and evaluating at $t = 0$).

Example 3.13. Suppose Y is a discrete random variable with pmf

$$p_Y(y) = \begin{cases} \left(\frac{1}{2}\right)^y, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

Find $E(Y)$ and $V(Y)$.

Solution. Using the definition of mathematical expectation, the first two moments of Y are

$$\begin{aligned}E(Y) &= \sum_{y=1}^{\infty} y\left(\frac{1}{2}\right)^y \\ E(Y^2) &= \sum_{y=1}^{\infty} y^2\left(\frac{1}{2}\right)^y.\end{aligned}$$

Neither of these sums are straightforward to calculate. Let's use the mgf of Y instead. Recall in Example 3.12 we found the mgf of Y to be

$$m_Y(t) = \frac{e^t}{2 - e^t}.$$

The first two derivatives of $m_Y(t)$ are

$$\begin{aligned} \frac{d}{dt} m_Y(t) &= \frac{2e^t}{(2-e^t)^2} \\ \frac{d^2}{dt^2} m_Y(t) &= \frac{2e^t(e^t+2)}{(2-e^t)^3}. \end{aligned}$$

Therefore,

$$E(Y) = \frac{d}{dt} m_Y(t)\bigg|_{t=0} = \frac{2e^0}{(2-e^0)^2} = 2.$$

The second moment is

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t)\bigg|_{t=0} = \frac{2e^0(e^0+2)}{(2-e^0)^3} = 6.$$

Applying the variance computing formula, we have

$$V(Y) = E(Y^2) - [E(Y)]^2 = 6 - 4 = 2.$$

Lesson: In this example and elsewhere, finding $E(Y)$ and $E(Y^2)$ using the definition of mathematical expectation can be difficult. Using mgfs can be much easier. In other examples (e.g., Example 3.5, etc.), finding $E(Y)$ and $E(Y^2)$ using the definition of mathematical expectation is easy. There is no need to use mgfs in these examples.

3.4 Binomial distribution

Important: Many experiments consist of a sequence of "trials," where

(i) each trial results in either a "success" or a "failure"

(ii) the probability of "success," denoted by p, $0 < p < 1$, is the same on every trial

(iii) the trials are mutually independent.

Trials that obey these three properties are called **Bernoulli trials**.

Terminology: Let Y denote the number of successes out of n Bernoulli trials. Then Y has a **binomial distribution** with parameters n (the number of trials) and probability of success p. We write $Y \sim b(n, p)$.

Example 3.14. Consider each of the following situations involving a binomial random variable. Are you satisfied with the three Bernoulli trial assumptions in each case?

- I flip a coin $n = 25$ times and record Y, the number of tails. If the coin is fair, then $Y \sim b(n = 25, p = 0.5)$.
- In an agricultural study, it is determined that 40 percent of all plots respond to a certain treatment. Four plots are observed. If Y denotes the number of plots that respond to the treatment, then $Y \sim b(n = 4, p = 0.4)$.
- In a biology experiment, 30 albino rats are injected with a drug that inhibits the synthesis of protein. The probability an individual rat will die from the drug before the study is complete is 0.15. If Y denotes the number of rats that die before the study is complete, then $Y \sim b(n = 30, p = 0.15)$.
- Auditors estimate that 22 percent of insurance claims of a certain type are fraudulent. There are 189 claims this year. If Y denotes the number of fraudulent claims this year, then $Y \sim b(n = 189, p = 0.22)$. $\square$

Note: Our goal is to derive the pmf of $Y \sim b(n, p)$; i.e., to derive a formula for

$$p_Y(y) = P(Y = y).$$

Among n Bernoulli trials, how can we get exactly y successes? We can think of one outcome (sample point) in the underlying sample space S as having the following structure

$$(\ \_\ \_\ \_\ \_\ \_\ \cdots\ \_\ \_\ )$$

where each position (trial) is occupied by an S (for a "success") or an F (for a "failure"). For example, if $n = 10$, one possible outcome looks like

$$(\ \text{S F F S S F S S F S}\ )$$

and corresponds to $y = 6$ successes.

In general, any ordering of y successes (S's) and $n - y$ failures (F's) occurs with probability

$$\underbrace{p \times p \times \cdots \times p}_{y \text{ successes}} \times \underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{n-y \text{ failures}} = p^y(1-p)^{n-y}.$$

This is true because the trials are mutually independent and the probability of success (and the probability of failure) is the same on every trial. Thus, all we have to do is count the number of outcomes in the sample space with y successes; each one of these outcomes has the same probability $p^y(1-p)^{n-y}$. Counting this is the same as counting the number of ways to choose y positions (among the n) to contain a success S; there are $\binom{n}{y}$ ways to do this.

PMF: The pmf of $Y \sim b(n, p)$ is

$$p_Y(y) = \begin{cases} \binom{n}{y} p^y(1-p)^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

Recall: The binomial expansion of $(a+b)^n$ is given by

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r} b^r.$$

Q: Is the binomial pmf $p_Y(y)$ valid?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 0, 1, 2, ..., n$. Do the probabilities $p_Y(y)$ sum to 1? Letting $a = 1-p$, $b = p$, and $r = y$ in the binomial expansion formula above, we have

$$[(1-p)+p]^n = \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y}.$$

The LHS clearly equals 1. The RHS is the $b(n,p)$ pmf. Thus, $p_Y(y)$ is valid. $\square$

MGF: The mgf of $Y \sim b(n,p)$ is

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1-p)^{n-y} = (q+pe^t)^n,$$

where $q = 1-p$. The last step follows from noting $\sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1-p)^{n-y}$ is the binomial expansion of $(q+pe^t)^n$. $\square$

Mean/Variance: The mean and variance of $Y \sim b(n,p)$ are

$$\begin{aligned} E(Y) &= np \\ V(Y) &= np(1-p). \end{aligned}$$

Proof. The first derivative of $m_Y(t)$ with respect to t is

$$m_Y'(t) = \frac{d}{dt} m_Y(t) = \frac{d}{dt}(q+pe^t)^n = n(q+pe^t)^{n-1} pe^t.$$

Thus,

$$E(Y) = \frac{d}{dt} m_Y(t)\bigg|_{t=0} = n(q+pe^0)^{n-1} pe^0 = n(q+p)^{n-1} p = np,$$

because $q+p = 1$. To find $V(Y)$, we can find the second moment $E(Y^2)$ and then use the variance computing formula. The second derivative of $m_Y(t)$ with respect to t is

$$\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \underbrace{n(q+pe^t)^{n-1} pe^t}_{m_Y'(t)} = n(n-1)(q+pe^t)^{n-2}(pe^t)^2 + n(q+pe^t)^{n-1} pe^t.$$

Thus,

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t)\bigg|_{t=0} = n(n-1)(q+pe^0)^{n-2}(pe^0)^2 + n(q+pe^0)^{n-1} pe^0 = n(n-1)p^2 + np.$$

Finally,

$$\begin{aligned} V(Y) = E(Y^2) - [E(Y)]^2 &= n(n-1)p^2 + np - (np)^2 \\ &= np(1-p). \ \square \end{aligned}$$

Note: WMS show how to derive $E(Y)$ by writing

$$E(Y) = \sum_{y=0}^{n} y\binom{n}{y}p^y(1-p)^{n-y}$$

and then manipulating this sum (see pp 107-108). Calculating $E(Y^2)$ directly is difficult, so the authors instead find the **second factorial moment**

$$E[Y(Y-1)] = \sum_{y=0}^{n} y(y-1)\binom{n}{y}p^y(1-p)^{n-y}.$$

Note that

$$E[Y(Y-1)] = E(Y^2 - Y) = E(Y^2) - E(Y) \implies E(Y^2) = E[Y(Y-1)] + E(Y).$$

Factorial moments are discussed in Section 3.10 (pp 143-146, WMS).

Example 3.15. Physicians conjecture that 35 percent of renal cell carcinoma patients will respond positively to a new drug treatment. A small clinical trial tests the new drug in 30 patients. Let Y denote the number of patients who will respond positively to the drug. If the Bernoulli trial assumptions hold for the patients (and the physicians' conjecture is correct), then $Y \sim b(n = 30, p = 0.35)$. The pmf of Y is shown in Figure 3.7 (next page).

Q: What is the probability exactly 10 patients respond positively? at most 10? at least 10?
A: We use the $b(n = 30, p = 0.35)$ pmf. The probability exactly 10 patients respond positively is

$$P(Y = 10) = p_Y(10) = \binom{30}{10}(0.35)^{10}(1-0.35)^{30-10} \approx 0.150.$$

The probability at most 10 patients respond positively is

$$P(Y \leq 10) = \sum_{y=0}^{10}\binom{30}{y}(0.35)^y(1-0.35)^{30-y} \approx 0.508.$$

The probability at least 10 patients respond positively is

$$P(Y \geq 10) = \sum_{y=10}^{30}\binom{30}{y}(0.35)^y(1-0.35)^{30-y} = 1 - \underbrace{\sum_{y=0}^{9}\binom{30}{y}(0.35)^y(1-0.35)^{30-y}}_{= P(Y \leq 9)} \approx 0.642.$$

Here is the R code that will perform these calculations:

```
> dbinom(10,30,0.35)
[1] 0.1502173
> pbinom(10,30,0.35)
[1] 0.5077582
> 1-pbinom(9,30,0.35)
[1] 0.6424591
```

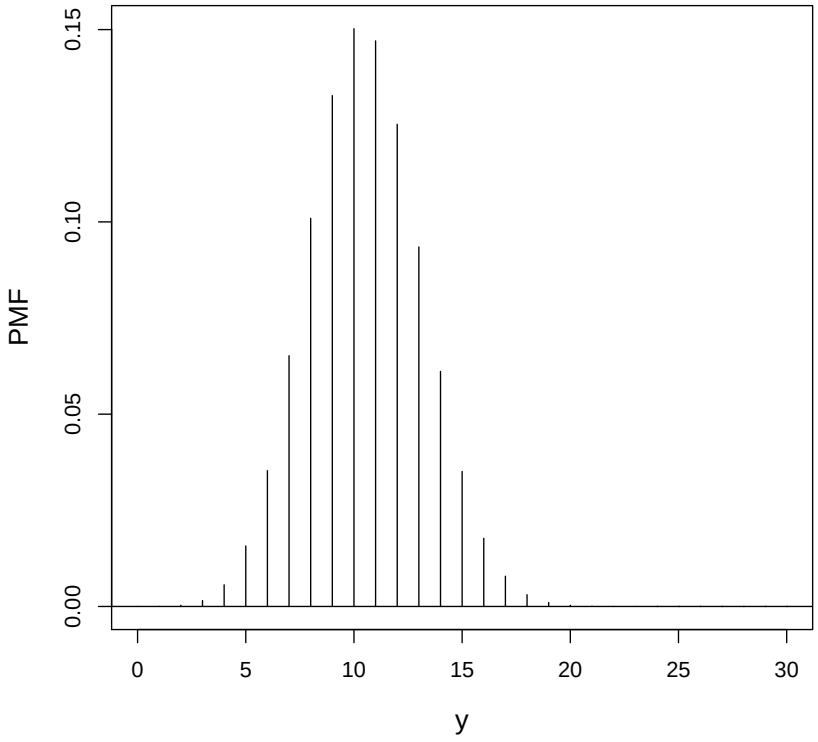

Figure 3.7: Pmf of $Y \sim b(n = 30, p = 0.35)$ in Example 3.15.

Q: What are $E(Y)$ and $V(Y)$?
A: The mean of Y is

$$E(Y) = np = 30(0.35) = 10.5 \text{ patients.}$$

Therefore, we would expect 10.5 patients to respond positively. The variance of Y is

$$V(Y) = np(1-p) = 30(0.35)(1-0.35) = 6.825 \text{ (patients)}^2.$$

The standard deviation is $\sigma = \sqrt{6.825} \approx 2.61$ patients. □

Important: In the $b(n, p)$ family, when $n = 1$, the binomial pmf reduces to

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

This is called the **Bernoulli distribution**. Shorthand notation is $Y \sim b(1, p)$ or $Y \sim$ Bernoulli(p). The Bernoulli distribution is used to model binary (0-1) outcomes; e.g., success/failure, agree/disagree, disease/healthy, etc.

3.5 Geometric distribution

Note: Recall the Bernoulli trial assumptions:

(i) each trial results in either a "success" or a "failure"

(ii) the probability of "success," denoted by p, $0 < p < 1$, is the same on every trial

(iii) the trials are mutually independent.

Terminology: Suppose Bernoulli trials are continually observed. Let Y denote the number of trials to observe the first success. Then Y has a **geometric distribution** with probability of success p. We write $Y \sim \text{geom}(p)$.

PMF: The pmf of $Y \sim \text{geom}(p)$ is

$$p_Y(y) = \begin{cases} (1-p)^{y-1}p, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

The form of this pmf makes sense; i.e., if the first success occurs on the yth trial, then the first $y-1$ trials were failures. Each failure occurs with probability $1-p$. The yth trial is a success (with probability p). Everything gets multiplied together because the Bernoulli trial outcomes are mutually independent.

Q: Is the geometric pmf $p_Y(y)$ valid?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 1, 2, 3, ...$ Do the probabilities $p_Y(y)$ sum to 1? We have

$$\sum_{y=1}^{\infty}(1-p)^{y-1}p = p\sum_{x=0}^{\infty}(1-p)^x = \frac{p}{1-(1-p)} = 1.$$

In the last step, we realize that $\sum_{x=0}^{\infty}(1-p)^x$ is an infinite geometric sum with common ratio $1-p$. $\square$

MGF: The mgf of $Y \sim \text{geom}(p)$ is

$$\begin{aligned} m_Y(t) = E(e^{tY}) = \sum_{y=1}^{\infty} e^{ty}(1-p)^{y-1}p &= \frac{p}{q}\sum_{y=1}^{\infty}(qe^t)^y \\ &= \frac{p}{q}\left[\sum_{y=0}^{\infty}(qe^t)^y - 1\right] \\ &= \frac{p}{q}\left(\frac{1}{1-qe^t} - 1\right) = \frac{pe^t}{1-qe^t}, \end{aligned}$$

where $q = 1-p$. Note that the infinite geometric sum $\sum_{y=0}^{\infty}(qe^t)^y$ above converges and is equal to $1/(1-qe^t)$ if and only if

$$qe^t < 1 \Longleftrightarrow t < -\ln q.$$

Therefore, the mgf exists and is given by the formula above. $\square$

Mean/Variance: The mean and variance of $Y \sim \text{geom}(p)$ are

$$\begin{aligned} E(Y) &= \frac{1}{p} \\ V(Y) &= \frac{q}{p^2}, \end{aligned}$$

where $q = 1 - p$.

Proof. The first derivative of $m_Y(t)$ with respect to t is

$$m_Y'(t) = \frac{d}{dt} m_Y(t) = \frac{d}{dt}\left(\frac{pe^t}{1-qe^t}\right) = \frac{pe^t(1-qe^t) - pe^t(-qe^t)}{(1-qe^t)^2} = \frac{pe^t}{(1-qe^t)^2}.$$

Therefore,

$$E(Y) = \frac{d}{dt} m_Y(t)\bigg|_{t=0} = \frac{pe^0}{(1-qe^0)^2} = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

To find $V(Y)$, we can find the second moment $E(Y^2)$ and then use the variance computing formula. The second derivative of $m_Y(t)$ with respect to t is

$$\frac{d^2}{dt^2} m_Y(t) = \frac{pe^t(1-qe^t)^2 - 2pe^t(1-qe^t)(-qe^t)}{(1-qe^t)^4}.$$

Therefore,

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t)\bigg|_{t=0} = \frac{pe^0(1-qe^0)^2 - 2pe^0(1-qe^0)(-qe^0)}{(1-qe^0)^4} = \frac{p^3 + 2p^2q}{p^4} = \frac{p+2q}{p^2}.$$

Finally,

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{p+2q}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}. \ \square$$

Note: WMS show how to derive $E(Y)$ and $V(Y)$ directly by writing

$$\begin{aligned} E(Y) &= \sum_{y=1}^{\infty} y(1-p)^{y-1}p \\ E[Y(Y-1)] &= \sum_{y=1}^{\infty} y(y-1)(1-p)^{y-1}p \end{aligned}$$

and then manipulating these sums; see pp 116-117.

Example 3.16. An EPA engineer is tasked with observing water specimens from lakes in northeast Georgia. In this region, each specimen has a 20 percent chance of containing a particular organic pollutant. Let Y denote the number of specimens observed to find the first one containing the pollutant. If the Bernoulli trial assumptions hold for the specimens, then $Y \sim \text{geom}(p = 0.20)$. The pmf of Y is shown in Figure 3.8 (next page).

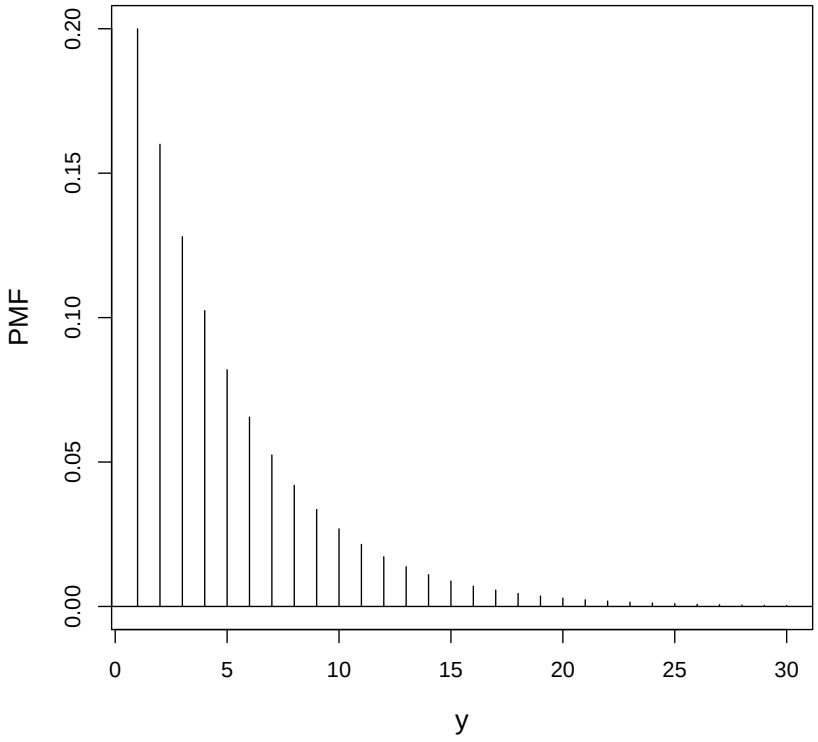

Figure 3.8: Pmf of $Y \sim \text{geom}(p = 0.20)$ in Example 3.16.

Here are the first few probabilities:

$$\begin{aligned}
P(Y=1) = p_Y(1) &= (1-0.20)^{1-1}(0.20) &= 0.20 \\
P(Y=2) = p_Y(2) &= (1-0.20)^{2-1}(0.20) &= 0.16 \\
P(Y=3) = p_Y(3) &= (1-0.20)^{3-1}(0.20) &= 0.128 \\
P(Y=4) = p_Y(4) &= (1-0.20)^{4-1}(0.20) &= 0.1024 \\
P(Y=5) = p_Y(5) &= (1-0.20)^{5-1}(0.20) &= 0.08192.
\end{aligned}$$

The probability the first water specimen containing the pollutant is observed among the first five specimens is

$$\begin{aligned}
P(Y \leq 5) &= P(Y=1) + P(Y=2) + P(Y=3) + P(Y=4) + P(Y=5) \\
&= \sum_{y=1}^{5} (1-0.20)^{y-1}(0.20) = 0.67232. \ \square
\end{aligned}$$

```
> pgeom(5-1,0.20)
[1] 0.67232
```

3.6 Negative binomial distribution

Note: Recall the Bernoulli trial assumptions:

(i) each trial results in either a "success" or a "failure"

(ii) the probability of "success," denoted by p, $0 < p < 1$, is the same on every trial

(iii) the trials are mutually independent.

Terminology: Suppose Bernoulli trials are continually observed. Let Y denote the number of trials to observe the rth success, where $r \geq 1$. Then Y has a **negative binomial distribution** with waiting parameter r and probability of success p. We write $Y \sim \text{nib}(r, p)$.

Note: When $r = 1$, the nib(r, p) distribution reduces to the geom(p) distribution. We can think of the negative binomial distribution as a generalization of the geometric; i.e., where one is "waiting" for more successes.

PMF: The pmf of $Y \sim \text{nib}(r, p)$ is

$$p_Y(y) = \begin{cases} \binom{y-1}{r-1} p^r (1-p)^{y-r}, & y = r, r+1, r+2, ... \\ 0, & \text{otherwise.} \end{cases}$$

The form of this pmf can be explained intuitively. If the rth success occurs on the yth trial, then $r-1$ successes must have occurred during the first $y-1$ trials. The number of sample points (in the underlying sample space) where this occurs is $\binom{y-1}{r-1}$, which counts the number of ways one can choose the locations of $r-1$ successes among the 1st $y-1$ trials. Because the trials are mutually independent, the probability of each of these sample points is $p^{r-1}(1-p)^{y-r}$. Therefore, the probability of exactly $r-1$ successes among the first $y-1$ trials is $\binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}$. On the yth trial, we observe the rth success (this occurs with probability p). Because the yth trial is independent of the previous $y-1$ trials, we have

$$P(Y = y) = \underbrace{\binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}}_{\text{pertains to 1st } y-1 \text{ trials}} \times p = \binom{y-1}{r-1} p^r (1-p)^{y-r}.$$

MGF: The mgf of $Y \sim \text{nib}(r, p)$ is

$$\left(\frac{pe^t}{1 - qe^t} \right)^r,$$

for $t < -\ln q$, where $q = 1 - p$.

Note: When $r = 1$, the nib(r, p) mgf reduces to the geom(p) mgf. Interesting!

Proof. The mgf of $Y \sim \text{nib}(r, p)$ is

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=r}^{\infty} e^{ty}\binom{y-1}{r-1}p^r(1-p)^{y-r} \\ &= (pe^t)^r \underbrace{\sum_{y=r}^{\infty}\binom{y-1}{r-1}(qe^t)^{y-r}}_{=\,(1-qe^t)^{-r}} = \left(\frac{pe^t}{1-qe^t}\right)^r, \end{aligned}$$

for $1 - qe^t > 0 \Longleftrightarrow t < -\ln q$. That $\sum_{y=r}^{\infty}\binom{y-1}{r-1}(qe^t)^{y-r} = (1-qe^t)^{-r}$ follows from the lemma below. □

LEMMA. Suppose r is a nonnegative integer. Then

$$\sum_{y=r}^{\infty}\binom{y-1}{r-1}(qe^t)^{y-r} = (1-qe^t)^{-r}.$$

Proof. Consider the function $f(w) = (1-w)^{-r}$, where r is a nonnegative integer. It is easy to show that

$$\begin{aligned} f'(w) &= r(1-w)^{-(r+1)} \\ f''(w) &= r(r+1)(1-w)^{-(r+2)} \\ f'''(w) &= r(r+1)(r+2)(1-w)^{-(r+3)}, \end{aligned}$$

and so on. In general, $f^{(z)}(w) = r(r+1)\cdots(r+z-1)(1-w)^{-(r+z)}$, where $f^{(z)}(w)$ denotes the zth derivative of f with respect to w. Note that

$$f^{(z)}(w)\Big|_{w=0} = r(r+1)\cdots(r+z-1).$$

Now writing $f(w)$ in its McLaurin Series expansion, we have

$$f(w) = \sum_{z=0}^{\infty}\frac{f^{(z)}(0)}{z!}w^z = \sum_{z=0}^{\infty}\frac{r(r+1)\cdots(r+z-1)}{z!}w^z = \sum_{z=0}^{\infty}\binom{r+z-1}{r-1}w^z.$$

Letting $w = qe^t$ and $z = y - r$ proves the lemma. □

Mean/Variance: The mean and variance of $Y \sim \text{nib}(r, p)$ are

$$\begin{aligned} E(Y) &= \frac{r}{p} \\ V(Y) &= \frac{rq}{p^2}, \end{aligned}$$

where $q = 1 - p$. Note again these formulae for $E(Y)$ and $V(Y)$ reduce to those for the geometric distribution when $r = 1$.

Proof. Exercise. □

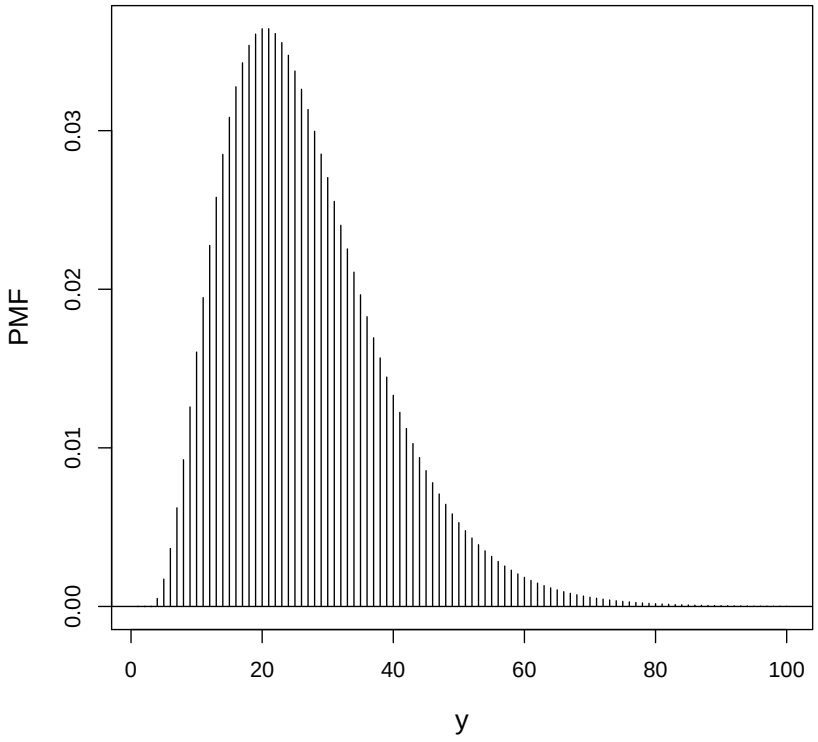

Figure 3.9: Pmf of $Y \sim \text{nib}(r = 4, p = 0.15)$ in Example 3.17.

Example 3.17. At an automotive plant, 15 percent of all paint batches sent to the lab for chemical analysis do not conform to specifications. Let Y denote the number of batches to find the 4th one that does not conform. If the Bernoulli trial assumptions hold for the batches, then $Y \sim \text{nib}(r = 4, p = 0.15)$. The pmf of Y is shown in Figure 3.9 (above).

Q: What is the probability no more than three nonconforming batches will be observed among the first 30 batches sent to the lab?
A: This will occur when the fourth nonconforming batch is observed on the 31st batch sent to the lab, the 32nd, the 33rd, etc. Therefore,

$$\begin{aligned} P(Y \geq 31) &= 1 - P(Y \leq 30) \\ &= 1 - \sum_{y=4}^{30} \binom{y-1}{4-1} (0.15)^4 (0.85)^{y-4} \approx 0.322. \ \square \end{aligned}$$

```
> 1-pnbinom(30-4,4,0.15)
[1] 0.3216599
```

3.7 Hypergeometric distribution

Setting: Consider a population of N objects and suppose each object belongs to one of two dichotomous classes: Class 1 or Class 2. For example, the objects and classes might be

Poker chips:	red/blue
People:	diseased/healthy
Plots of land:	respond to treatment/not.

In the population of interest, we have

$$\begin{aligned} N &= \text{total number of objects} \\ r &= \text{number of objects in Class 1} \\ N-r &= \text{number of objects in Class 2.} \end{aligned}$$

We sample n objects from the population at random and without replacement. Define

$$Y = \text{ number of objects in Class 1 (among the } n \text{ sampled).}$$

Then Y has a **hypergeometric distribution** with population size N, sample size n, and number of Class 1 objects r. We write $Y \sim \text{hyper}(N, n, r)$.

Remark: We have already seen an example of this distribution in Example 3.2 (notes). In this example, the "objects" were passengers and the classes were male/female; i.e.,

$$\begin{aligned} N &= \text{total number of passengers } = 66 \\ r &= \text{number of males } = 30 \\ N-r &= \text{number of females } = 36. \end{aligned}$$

We sampled $n = 5$ passengers at random and without replacement from the population of 66 passengers and recorded Y, the number of males among those sampled. In this example, $Y \sim \text{hyper}(N = 66, n = 5, r = 30)$. By conceptualizing the selection of $n = 5$ passengers as a random experiment, we derived the pmf of Y to be

$$p_Y(y) = \begin{cases} \dfrac{\binom{30}{y}\binom{36}{5-y}}{\binom{66}{5}}, & y = 0, 1, 2, 3, 4, 5 \\ 0, & \text{otherwise.} \end{cases}$$

The hypergeometric pmf derivation generalizes immediately.

PMF: The pmf of $Y \sim \text{hyper}(N, n, r)$ is

$$p_Y(y) = \begin{cases} \dfrac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & \begin{array}{c} y = 0, 1, 2, ..., n \\ y \leq r,\ n-y \leq N-r \end{array} \\ 0, & \text{otherwise.} \end{cases}$$

Comparison: The motivation for the hypergeometric distribution should remind us of the underlying framework for the binomial; i.e., we record the number of Class 1 objects ("successes") out of n ("trials"). The difference here is that

- the population size N is finite
- sampling is done without replacement.

To understand further, suppose

$$p = \frac{r}{N} = \text{proportion of Class 1 objects in the population.}$$

Because sampling from the population is done **without replacement**, the value of p changes from trial to trial. This violates the Bernoulli trial assumptions, so technically the binomial model does not apply. However, one can show mathematically that

$$\lim_{\substack{N\to\infty \\ r/N\to p}} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \underbrace{\binom{n}{y}p^y(1-p)^{n-y}}_{b(n,p)\text{ pmf}}.$$

This result implies that if the population size N is "large," the hyper(N, n, r) distribution and the $b(n, p = r/N)$ distribution should be very close to each other even when one samples without replacement. Of course, if one samples from a population **with replacement**, then $p = r/N$ remains fixed and hence the binomial model applies regardless of how large N is.

Example 3.18. A supplier ships parts to a company in lots of 1000 parts. Suppose a lot contains 100 defective parts and 900 non-defective parts. An operator selects 10 parts at random and without replacement. What is the probability he selects no more than 2 defective parts?

Hypergeometric: Because sampling is done without replacement, a hypergeometric model applies. We recognize

$$\begin{aligned} N &= \text{total number of parts } = 1000 \\ r &= \text{number of defectives } = 100 \\ N-r &= \text{number of non-defectives } = 900. \end{aligned}$$

Let Y denote the number of defective parts (i.e., "Class 1 objects") out of $n = 10$. Then $Y \sim$ hyper$(N = 1000, n = 10, r = 100)$ and

$$\begin{aligned} P(Y \le 2) &= P(Y=0) + P(Y=1) + P(Y=2) \\ &= \frac{\binom{100}{0}\binom{900}{10}}{\binom{1000}{10}} + \frac{\binom{100}{1}\binom{900}{9}}{\binom{1000}{10}} + \frac{\binom{100}{2}\binom{900}{8}}{\binom{1000}{10}} \\ &\approx 0.3469 + 0.3894 + 0.1945 = 0.9308. \end{aligned}$$

```
> phyper(2,100,900,10)
[1] 0.9307629
```

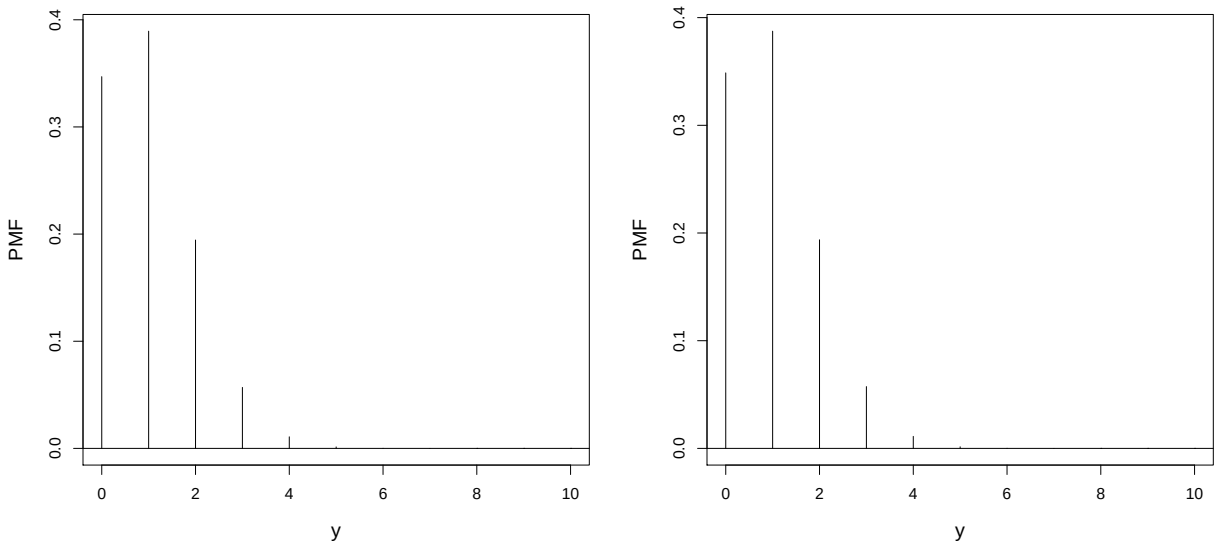

Figure 3.10: Example 3.18. Left: Pmf of $Y \sim \text{hyper}(N = 1000, n = 10, r = 100)$. Right: Pmf of $Y \sim b(n = 10, p = 0.10)$.

Binomial: The population proportion of defective parts is

$$p = \frac{100}{1000} = 0.10.$$

Therefore, the $b(n = 10, p = 0.10)$ model should offer a good approximation to the (exact) answer obtained from the hypergeometric calculation. We have

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= \binom{10}{0}(0.10)^0(0.90)^{10} + \binom{10}{1}(0.10)^1(0.90)^9 + \binom{10}{2}(0.10)^2(0.90)^8 \\ &\approx 0.3487 + 0.3874 + 0.1937 = 0.9298. \end{aligned}$$

```
> pbinom(2,10,0.10)
[1] 0.9298092
```

Figure 3.10 (above) shows the hypergeometric and binomial pmfs used in this problem. Note that they are nearly identical in appearance. □

Q: Is the hypergeometric pmf $p_Y(y)$ valid?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 1, 2, 3, ...$ Do the probabilities $p_Y(y)$ sum to 1? The answer is yes, of course, but showing this is not trivial. It suffices to show

$$\sum_{y=0}^{n} \binom{r}{y}\binom{N-r}{n-y} = \binom{N}{n}.$$

See Exercise 3.216 (pp 156, WMS).

Mean/Variance: The mean and variance of $Y \sim \text{hyper}(N, n, r)$ are

$$\begin{aligned} E(Y) &= n\left(\frac{r}{N}\right) \\ V(Y) &= n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right). \end{aligned}$$

Deriving these formulas is not trivial either. The mgf of $Y \sim \text{hyper}(N, n, r)$ exists, but its form is not very friendly. Therefore, to derive $E(Y)$, we will have to appeal directly to the definition of expected value; note that

$$E(Y) = \sum_{y=0}^{n} y p_Y(y) = \sum_{y=0}^{n} y \, \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \sum_{y=1}^{n} y \, \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}.$$

The denominator in the pmf of Y can be written as

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N}{n}\left[\frac{(N-1)!}{(n-1)!(N-n)!}\right] = \frac{N}{n}\binom{N-1}{n-1}.$$

Therefore,

$$\begin{aligned} E(Y) = \frac{n}{N}\sum_{y=1}^{n} y \, \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N-1}{n-1}} &= \frac{n}{N}\sum_{y=1}^{n} y \, \frac{\frac{r!}{y!(r-y)!}\frac{(N-r)!}{(n-y)!(N-r-n+y)!}}{\binom{N-1}{n-1}} \\ &= \frac{nr}{N}\sum_{y=1}^{n} \frac{\frac{(r-1)!}{(y-1)!(r-y)!}\frac{(N-r)!}{(n-y)!(N-r-n+y)!}}{\binom{N-1}{n-1}} \\ &\overset{x=y-1}{=} \frac{nr}{N}\sum_{x=0}^{n-1} \frac{\frac{(r-1)!}{x!(r-1-x)!}\frac{(N-r)!}{(n-1-x)!(N-r-n+1+x)!}}{\binom{N-1}{n-1}} \\ &= \frac{nr}{N}\sum_{x=0}^{n-1} \frac{\binom{r-1}{x}\binom{N-r}{n-1-x}}{\binom{N-1}{n-1}}. \end{aligned}$$

However,

$$\sum_{x=0}^{n-1} \frac{\binom{r-1}{x}\binom{N-r}{n-1-x}}{\binom{N-1}{n-1}} = \sum_{x=0}^{n-1} \frac{\binom{r-1}{x}\binom{(N-1)-(r-1)}{(n-1)-x}}{\binom{N-1}{n-1}} = 1$$

because the summand is the pmf of $X \sim \text{hyper}(N-1, n-1, r-1)$ and we sum over the support of this random variable; i.e., $x = 0, 1, ..., n-1$. Thus, the result. $\square$

Note: To derive $V(Y)$, it is easier to first calculate the **second factorial moment**

$$E[Y(Y-1)] = \sum_{y=0}^{n} y(y-1)\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}.$$

Recall that

$$E[Y(Y-1)] = E(Y^2 - Y) = E(Y^2) - E(Y) \implies E(Y^2) = E[Y(Y-1)] + E(Y).$$

Interesting: If the population size $N \to \infty$ so that $r/N \to p \in (0,1)$, note that

$$E(Y) = n\left(\frac{r}{N}\right) \to np,$$

the mean of the $b(n,p)$ distribution. Similarly,

$$V(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right) \to np(1-p),$$

which is the variance of the $b(n,p)$ distribution. Neither result is surprising given the result on pp 65 (notes); i.e., if the hyper(N,n,r) pmf converges to the $b(n,p)$ pmf as $N \to \infty$ and $r/N \to p$, then the corresponding moments should converge as well.

3.8 Poisson distribution

Setting: Suppose we count the number of "occurrences" in a continuous interval of time (or space). A **Poisson process** enjoys the following properties:

1. the number of occurrences in non-overlapping intervals are independent random variables
2. the probability of an occurrence in a sufficiently short interval is proportional to the length of the interval
3. the probability of 2 or more occurrences in a sufficiently short interval is zero.

Suppose a counting process satisfies the three conditions above. Define

$$Y = \text{ the number of occurrences in a } \textbf{unit interval} \text{ of time (or space).}$$

Our goal is to find an expression for $p_Y(y) = P(Y = y)$, the pmf of Y.

Derivation: Partition the unit interval $[0,1]$ into n subintervals, each of size $1/n$.

- If n is sufficiently large (i.e., much larger than y), then we can approximate the probability y events occur in the unit interval by finding the probability that exactly one event (occurrence) occurs in exactly y of the subintervals.
- By Property (2), we know that the probability of one event in any one subinterval is **proportional** to the subinterval's length, say λ/n, where λ is the proportionality constant.
- By Property (3), the probability of more than one occurrence in any subinterval is zero (for n large).

- Consider the occurrence/non-occurrence of an event in each subinterval as a **Bernoulli trial**. By Property (1), we have a sequence of n Bernoulli trials, each with probability of "success" $p = \lambda/n$. Thus, a binomial (approximate) calculation gives

$$P(Y = y) \approx \binom{n}{y}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y}.$$

To improve the approximation for $P(Y = y)$, we let n grow large without bound; i.e., let $n \to \infty$. We have

$$\begin{aligned}
\lim_{n\to\infty} P(Y=y) &= \lim_{n\to\infty}\binom{n}{y}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} \\
&= \lim_{n\to\infty}\frac{n!}{y!(n-y)!}\,\lambda^y\left(\frac{1}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^n\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y \\
&= \lim_{n\to\infty}\underbrace{\frac{n(n-1)\cdots(n-y+1)}{n^y}}_{a_n}\underbrace{\frac{\lambda^y}{y!}}_{b_n}\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{c_n}\underbrace{\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y}_{d_n}.
\end{aligned}$$

Now, the limit of the product is the product of the limits:

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty}\frac{n(n-1)\cdots(n-y+1)}{n^y} = 1 \qquad \lim_{n\to\infty} b_n = \lim_{n\to\infty}\frac{\lambda^y}{y!} = \frac{\lambda^y}{y!}$$

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \qquad \lim_{n\to\infty} d_n = \lim_{n\to\infty}\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y = 1.$$

We have shown that

$$\lim_{n\to\infty} P(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

We say that Y follows a **Poisson distribution** with parameter λ. Shorthand notation is $Y \sim \text{Poisson}(\lambda)$.

PMF: The pmf of $Y \sim \text{Poisson}(\lambda)$ is

$$p_Y(y) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Q: Is the Poisson pmf $p_Y(y)$ valid?
A: Clearly, $0 \leq p_Y(y) \leq 1$, for each $y = 0, 1, 2, ...$ Do the probabilities $p_Y(y)$ sum to 1? We have

$$\sum_{y=0}^{\infty}\frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{-\lambda}e^{\lambda} = 1.$$

Recall that $\sum_{y=0}^{\infty}\lambda^y/y!$ is the McLaurin series expansion of e^{λ}. $\square$

MGF: The mgf of $Y \sim \text{Poisson}(\lambda)$ is

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda} \underbrace{\sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!}}_{= \exp(\lambda e^t)} = e^{-\lambda} e^{\lambda e^t} = \exp[\lambda(e^t - 1)]. \ \square$$

Mean/Variance: The mean and variance of $Y \sim \text{Poisson}(\lambda)$ are

$$\begin{aligned} E(Y) &= \lambda \\ V(Y) &= \lambda. \end{aligned}$$

Proof. The first derivative of $m_Y(t)$ with respect to t is

$$m_Y'(t) = \frac{d}{dt} m_Y(t) = \frac{d}{dt} \exp[\lambda(e^t - 1)] = \lambda e^t \exp[\lambda(e^t - 1)].$$

Thus,

$$E(Y) = \frac{d}{dt} m_Y(t) \bigg|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] = \lambda.$$

To find $V(Y)$, we can find the second moment $E(Y^2)$ and then use the variance computing formula. The second derivative of $m_Y(t)$ with respect to t is

$$\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \underbrace{\lambda e^t \exp[\lambda(e^t - 1)]}_{m_Y'(t)} = \lambda e^t \exp[\lambda(e^t - 1)] + (\lambda e^t)^2 \exp[\lambda(e^t - 1)].$$

Thus,

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t) \bigg|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] + (\lambda e^0)^2 \exp[\lambda(e^0 - 1)] = \lambda + \lambda^2.$$

Finally,

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \lambda + \lambda^2 - \lambda^2 = \lambda. \ \square \end{aligned}$$

Note: WMS show how to derive $E(Y)$ directly by writing

$$E(Y) = \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!}$$

and then manipulating this sum. This is easy to do. Note that

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

Letting $x = y - 1$ in the last sum, we get

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

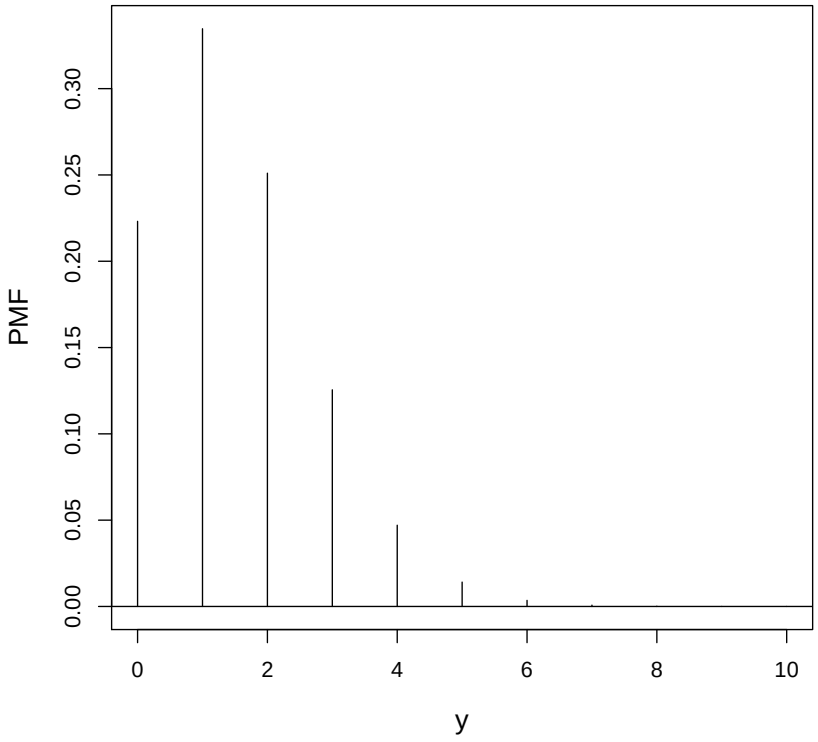

Figure 3.11: Pmf of $Y \sim \text{Poisson}(\lambda = 1.5)$ in Example 3.19.

To derive $V(Y)$, we could calculate the second factorial moment $E[Y(Y-1)]$ and then use the fact that

$$E[Y(Y-1)] = E(Y^2 - Y) = E(Y^2) - E(Y) \implies E(Y^2) = E[Y(Y-1)] + E(Y).$$

However, in this case, it is just as easy to calculate $E(Y^2)$ directly. Note that

$$E(Y^2) = \sum_{y=0}^{\infty} y^2 \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^{\infty} y^2 \frac{\lambda^y e^{-\lambda}}{y!} = \lambda \sum_{y=1}^{\infty} y \frac{\lambda^{y-1} e^{-\lambda}}{(y-1)!}$$

Letting $x = y - 1$ in the last sum, we get

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

where $X \sim \text{Poisson}(\lambda)$. Therefore, $E(Y^2) = \lambda(\lambda + 1) = \lambda^2 + \lambda$, which is the same as what we got by finding $E(Y^2)$ using the mgf of Y.

Example 3.19. In a certain region in the northeast US, the number of severe weather events per year Y is assumed to have a Poisson distribution with mean $\lambda = 1.5$. The pmf of $Y \sim \text{Poisson}(\lambda = 1.5)$ is shown in Figure 3.11 above.

Q: What is the probability there are four or more severe weather events in a given year?
A: We want to find $P(Y \geq 4)$. Work directly with the Poisson pmf; first note that

$$\begin{aligned} P(Y \leq 3) &= P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) \\ &= \frac{(1.5)^0 e^{-1.5}}{0!} + \frac{(1.5)^1 e^{-1.5}}{1!} + \frac{(1.5)^2 e^{-1.5}}{2!} + \frac{(1.5)^3 e^{-1.5}}{3!} \\ &\approx 0.223 + 0.335 + 0.251 + 0.126 = 0.935. \end{aligned}$$

By the complement rule,

$$P(Y \geq 4) = 1 - P(Y \leq 3) \approx 1 - 0.935 = 0.065.$$

```
> 1-ppois(3,1.5)
[1] 0.06564245
```

Q: A company buys a policy to insure its revenue in the event of severe weather that shuts down business. The policy pays nothing for the first such weather event of the year and \$10,000 for each one thereafter, until the end of the year. Calculate the expected amount paid to the company under this policy during a one-year period.
A: First note that if $Y = 0$ or $Y = 1$, then the company receives nothing according to the policy. It is only when there are 2 or more severe weather events does a payout occur, and this payout is \$10,000 for each event. Therefore, the payout when viewed as a function of Y is given by

$$g(Y) = \begin{cases} 0, & Y = 0, 1 \\ 10000(Y-1), & Y = 2, 3, 4, ... \end{cases}$$

and we want to calculate $E[g(Y)]$. From the definition of mathematical expectation, we have

$$\begin{aligned} E[g(Y)] &= \sum_{y=0}^{\infty} g(y) \frac{(1.5)^y e^{-1.5}}{y!} \\ &= 0 \times \frac{(1.5)^0 e^{-1.5}}{0!} + 0 \times \frac{(1.5)^1 e^{-1.5}}{1!} + \sum_{y=2}^{\infty} 10000(y-1) \frac{(1.5)^y e^{-1.5}}{y!} \\ &= 10000 \left[\sum_{y=0}^{\infty} (y-1) \frac{(1.5)^y e^{-1.5}}{y!} - (1-1) \times \frac{(1.5)^1 e^{-1.5}}{1!} - (0-1) \times \frac{(1.5)^0 e^{-1.5}}{0!} \right]. \end{aligned}$$

Note that

$$\sum_{y=0}^{\infty} (y-1) \frac{(1.5)^y e^{-1.5}}{y!} = E(Y-1) = E(Y) - 1 = 1.5 - 1 = 0.5.$$

Therefore,

$$E[g(Y)] = 10000(0.5 - 0 + e^{-1.5}) \approx 7231.30.$$

The expected payout to the company during a one-year period is \$7,231.30. □

4 Continuous Random Variables and their Probability Distributions

4.1 Introduction

Recall: The last chapter dealt with discrete random variables. A discrete random variable Y can assume a finite or (at most) a countable number of values. The probability mass function (pmf) of a discrete random variable

$$p_Y(y) = P(Y = y)$$

specifies how to assign probability to each support point $y \in R$, a countable set.

Preview: Continuous random variables have supports R that are not countable. Instead, the support of a continuous random variable Y is an interval like $R = \{y : 0 \leq y \leq 1\}$, $R = \{y : 0 < y < \infty\}$, $R = \{y : -\infty < y < \infty\}$, etc. Therefore, probabilities of events involving continuous random variables must be assigned in a different way.

Note: Before we introduce continuous random variables and their distributions formally, we start by introducing a new function that describes the probability distribution of any random variable (discrete, continuous, or any combination thereof).

4.2 Cumulative distribution functions

Terminology: The **cumulative distribution function (cdf)** of a random variable Y is the function

$$F_Y(y) = P(Y \leq y), \quad \text{for all } y \in \mathbb{R}.$$

The cdf $F_Y(y)$ is defined for all $y \in \mathbb{R} = (-\infty, \infty)$; not just for those values of y in the support. Every random variable, discrete or continuous, has a cdf. Note also that $F_Y(y) = P(Y \leq y)$ is a probability for any value of y. Therefore, the cdf of any random variable has domain $\mathbb{R}$ and range $[0, 1]$; i.e., $F_Y : \mathbb{R} \to [0, 1]$.

Example 4.1. Suppose $Y \sim b(n = 3, p = 0.4)$; i.e., Y has a binomial distribution with $n = 3$ trials and probability of success $p = 0.4$. Here are all the (nonzero) probabilities provided by the pmf:

$$\begin{aligned}
P(Y = 0) &= \binom{3}{0}(0.4)^0(0.6)^3 = 0.216 \\
P(Y = 1) &= \binom{3}{1}(0.4)^1(0.6)^2 = 0.432 \\
P(Y = 2) &= \binom{3}{2}(0.4)^2(0.6)^1 = 0.288 \\
P(Y = 3) &= \binom{3}{3}(0.4)^3(0.6)^0 = 0.064.
\end{aligned}$$

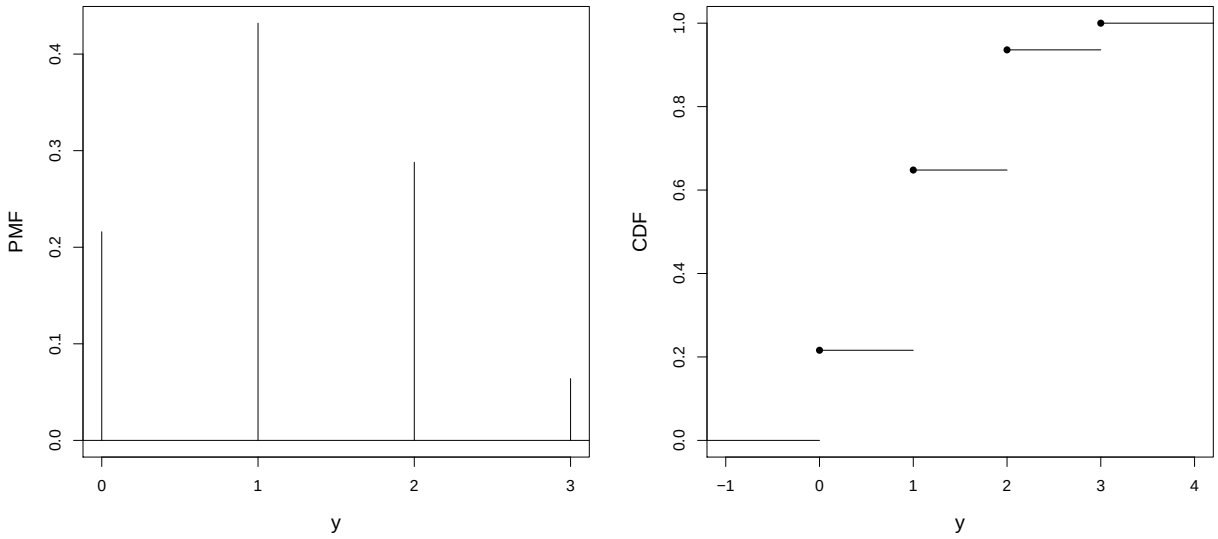

Figure 4.1: Pmf (left) and cdf (right) of $Y \sim b(n = 3, p = 0.4)$ in Example 4.1.

The pmf and cdf of Y are shown side by side in Figure 4.1 above. The cdf of Y is given by

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ 0.216, & 0 \leq y < 1 \\ 0.648, & 1 \leq y < 2 \\ 0.936, & 2 \leq y < 3 \\ 1, & y \geq 3, \end{cases}$$

which is a **step function**. The probabilities in $F_Y(y)$ are calculated as follows:

$$\begin{aligned} F_Y(0) = P(Y \leq 0) &= P(Y = 0) = 0.216 \\ F_Y(1) = P(Y \leq 1) &= P(Y = 0) + P(Y = 1) = 0.216 + 0.432 = 0.648 \\ F_Y(2) = P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) = 0.216 + 0.432 + 0.288 = 0.936 \\ F_Y(3) = P(Y \leq 3) &= P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) \\ &= 0.216 + 0.432 + 0.288 + 0.064 = 1. \end{aligned}$$

Note: The cdf $F_Y(y)$ in this example takes a "step" at the support points $y \in \{0, 1, 2, 3\}$ and stays constant otherwise. The height of the step at a particular point y is equal to $p_Y(y) = P(Y = y)$, the probability associated with that point. □

Important: The cdf of a random variable is an important function−both theoretically and practically.

- A random variable's cdf completely determines its probability distribution. In other words, if two different random variables have the same cdf, then the random variables have the same probability distribution.

- Statistical software (like R) catalogues the cdfs of well-known distributions so probabilities associated with these models can be calculated easily. For example, the R code

```
> pbinom(2,3,0.4)
[1] 0.936
```

calculates $F_Y(2) = P(Y \leq 2)$ in Example 4.1.

Result: The function $F_Y : \mathbb{R} \to [0, 1]$ is a cdf if and only if these conditions hold:

1. $\lim_{y \to -\infty} F_Y(y) = 0$ and $\lim_{y \to \infty} F_Y(y) = 1$.
2. $F_Y(y)$ is a nondecreasing function of y; i.e.,
$$y_1 \leq y_2 \Longrightarrow F_Y(y_1) \leq F_Y(y_2),$$
for all $y_1, y_2 \in \mathbb{R}$. **Note:** If $F_Y(y)$ is differentiable, then you can show $F_Y(y)$ is nondecreasing by showing $F'_Y(y) \geq 0$ for all $y \in \mathbb{R}$.
3. $F_Y(y)$ is right-continuous; i.e.,
$$\lim_{y \to y_0^+} F_Y(y) = F_Y(y_0),$$
for all $y_0 \in \mathbb{R}$.

4.3 Continuous random variables

Terminology: A random variable Y is said to be **continuous** if its cdf $F_Y(y)$ is a continuous function of y. Mathematically, this means

$$\lim_{y \to y_0} F_Y(y) = F_Y(y_0),$$

for all $y_0 \in \mathbb{R}$.

Remark: This definition highlights the salient difference between discrete and continuous random variables:

$$\begin{aligned} Y \text{ discrete} &\iff F_Y(y) \text{ is a step function} \\ Y \text{ continuous} &\iff F_Y(y) \text{ is continuous.} \end{aligned}$$

Recall that the height of a **discrete** cdf's "step" at any value $y \in R$ gives the probability $P(Y = y)$. Because a continuous cdf has no discontinuous steps, this means that (strictly positive) probabilities are not assigned to specific values of y in continuous distributions. This can be proven rigorously and we do this now.

Result: If Y is a continuous random variable with cdf $F_Y(y)$, then

$$P(Y = y) = 0, \quad \text{for all } y \in \mathbb{R}.$$

Proof. Suppose $\epsilon > 0$ so that $\{Y = y\} \subseteq \{y - \epsilon < Y \leq y\}$. By the monotonicity rule of probability (pp 10, notes) and Axiom 3,

$$P(Y = y) \leq P(y - \epsilon < Y \leq y) = P(Y \leq y) - P(Y \leq y - \epsilon) = F_Y(y) - F_Y(y - \epsilon).$$

Because probabilities are nonnegative (Axiom 1), we have

$$\begin{aligned} 0 \leq P(Y = y) &\leq \lim_{\epsilon \to 0} P(y - \epsilon < Y \leq y) \\ &= \lim_{\epsilon \to 0} [F_Y(y) - F_Y(y - \epsilon)] \\ &= F_Y(y) - \lim_{\epsilon \to 0} F_Y(y - \epsilon) \\ &= F_Y(y) - F_Y(y) = 0. \end{aligned}$$

Note that $\lim_{\epsilon \to 0} F_Y(y - \epsilon) = F_Y(y)$ because $F_Y(y)$ is continuous by assumption. Therefore, we have shown

$$0 \leq P(Y = y) \leq 0$$

which implies $P(Y = y) = 0$. Because ϵ was arbitrary, we are done. □

Summary: Discrete random variables Y have positive probability assigned to support points $y \in R$. Continuous random variables do not.

Example 4.2. The length of time until failure (in 100s of hours) for a transistor is a random variable Y with cumulative distribution function

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ 1 - e^{-y^2}, & y \geq 0. \end{cases}$$

Show that $F_Y(y)$ is a valid cdf.

Proof. The end behavior requirements are met:

$$\lim_{y \to -\infty} F_Y(y) = \lim_{y \to -\infty} 0 = 0$$

and

$$\lim_{y \to \infty} F_Y(y) = \lim_{y \to \infty} (1 - e^{-y^2}) = 1 - \lim_{y \to \infty} e^{-y^2} = 1 - 0 = 1.$$

A graph of $F_Y(y)$ is shown in Figure 4.2 (next page). Clearly, $F_Y(y)$ is nondecreasing. We can also show this mathematically by noting

$$F_Y'(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}(1 - e^{-y^2}) = 0 - (-2y)e^{-y^2} = 2ye^{-y^2} \geq 0,$$

for all $y \geq 0$. Because $F_Y'(y) \geq 0$, this means $F_Y(y)$ is nondecreasing. Finally, $F_Y(y)$ is a continuous function (see Figure 4.2) so it is clearly right-continuous. □

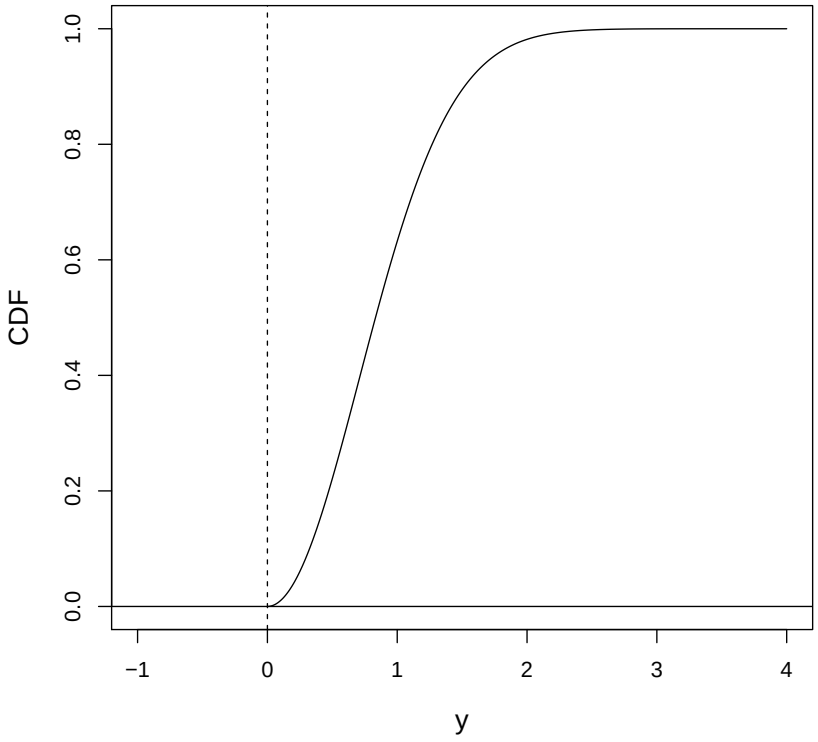

Figure 4.2: Cdf of Y in Example 4.2.

Terminology: Suppose Y is a continuous random variable with cdf $F_Y(y)$. The **probability density function** (**pdf**) for Y, denoted by $f_Y(y)$, is given by

$$f_Y(y) = \frac{d}{dy}F_Y(y),$$

provided that $(d/dy)F_Y(y)$ exists. If $f_Y(y)$ is a continuous function, then

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt.$$

This follows from the Fundamental Theorem of Calculus. These are important facts that describe how the pdf and cdf of a continuous random variable are related.

Remark: Every continuous random variable that we will discuss in this course has a pdf. It is common for students to think of a pdf $f_Y(y)$ as the "continuous analogue" of a pmf $p_Y(y)$ in the discrete case. It is fine to do this; however, probabilities in continuous distributions are not determined by calculating values of $f_Y(y)$ for $y \in R$. Instead, probabilities are determined by integration as we will see shortly. One can think of the pdf of a continuous random variable Y as a **theoretical model** for a population of measurements. This conceptualization is illustrated in the next example.

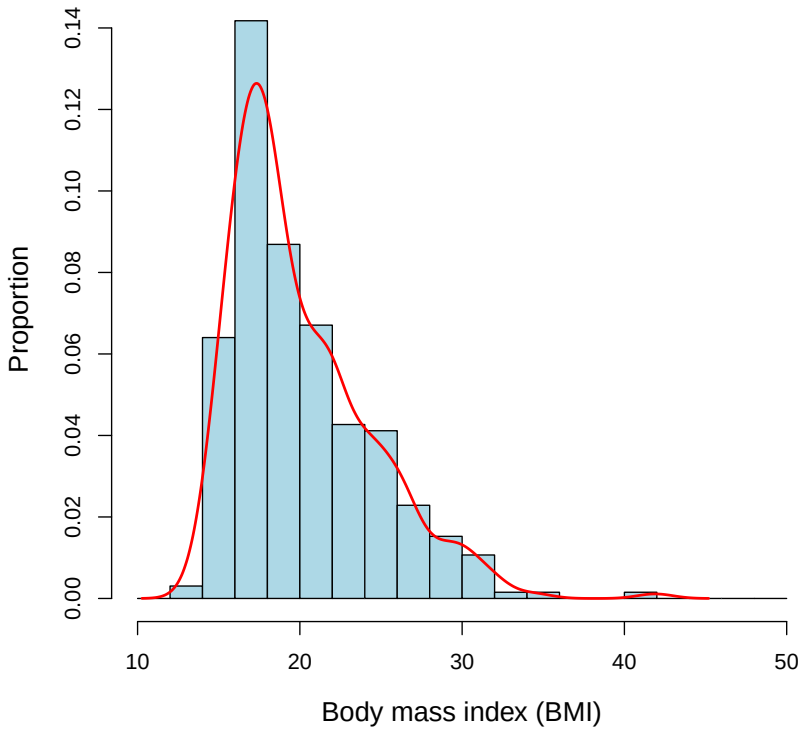

Figure 4.3: Histogram of $n = 328$ BMI measurements for fourth-grade children in Augusta. An estimate of the (population) pdf is shown as a smooth curve.

Example 4.3. In an observational study examining aspects related to childhood obesity, Baxter *and others* (2012) measured the body mass index (BMI) of $n = 328$ fourth-grade children sampled from a large public school district in Augusta, GA. A **histogram** of the data is shown in Figure 4.3 (above) along with a smooth curve that "approximates" the data. The smooth curve in this example is an estimate of what the pdf of Y is, where Y denotes the BMI of a child in this population of fourth-grade children in Augusta. This curve serves as a theoretical model for the entire population of children. □

Properties: The pdf of a continuous random variable Y has the following properties:

1. $f_Y(y) \geq 0$, for all $y \in \mathbb{R}$
2. The function $f_Y(y)$ integrates to one; i.e.,

$$\int_{\mathbb{R}} f_Y(y)dy = 1.$$

Compare these properties with those of a valid pmf $p_Y(y)$ in discrete case; see pp 38 (notes).

Example 4.4. Suppose Y is a continuous random variable with pdf

$$f_Y(y) = \begin{cases} c(4y - 2y^2), & 0 < y < 2 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the value of c that makes $f_Y(y)$ a valid pdf.
(b) Find the cdf of Y.

Solutions. (a) We can find c by using the fact $\int_{\mathbb{R}} f_Y(y)dy = 1$. Note that

$$\begin{aligned} 1 = \int_{\mathbb{R}} f_Y(y)dy &= \int_0^2 c(4y - 2y^2)dy \\ &= c\left(2y^2 - \frac{2}{3}y^3\right)\Big|_{y=0}^{2} = c\left(8 - \frac{16}{3}\right) = \frac{8}{3}c. \end{aligned}$$

Therefore, $c = 3/8$ and

$$f_Y(y) = \begin{cases} \frac{3}{8}(4y - 2y^2), & 0 < y < 2 \\ 0, & \text{otherwise.} \end{cases}$$

(b) The general expression for the cdf of Y is

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt,$$

which we must calculate for all $y \in \mathbb{R}$.

Case 1: When $y \leq 0$,

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y} 0dt = 0.$$

Case 2: When $0 < y < 2$,

$$\begin{aligned} F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt &= \int_{-\infty}^{0} 0dt + \int_0^y \frac{3}{8}(4t - 2t^2)dt \\ &= 0 + \frac{3}{8}\left(2t^2 - \frac{2}{3}t^3\right)\Big|_{t=0}^{y} = \frac{3}{8}\left(2y^2 - \frac{2}{3}y^3\right). \end{aligned}$$

Case 3: When $y \geq 2$,

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{0} 0dt + \underbrace{\int_0^2 \frac{3}{8}(4t - 2t^2)dt}_{=1} + \int_2^y 0dt = 1.$$

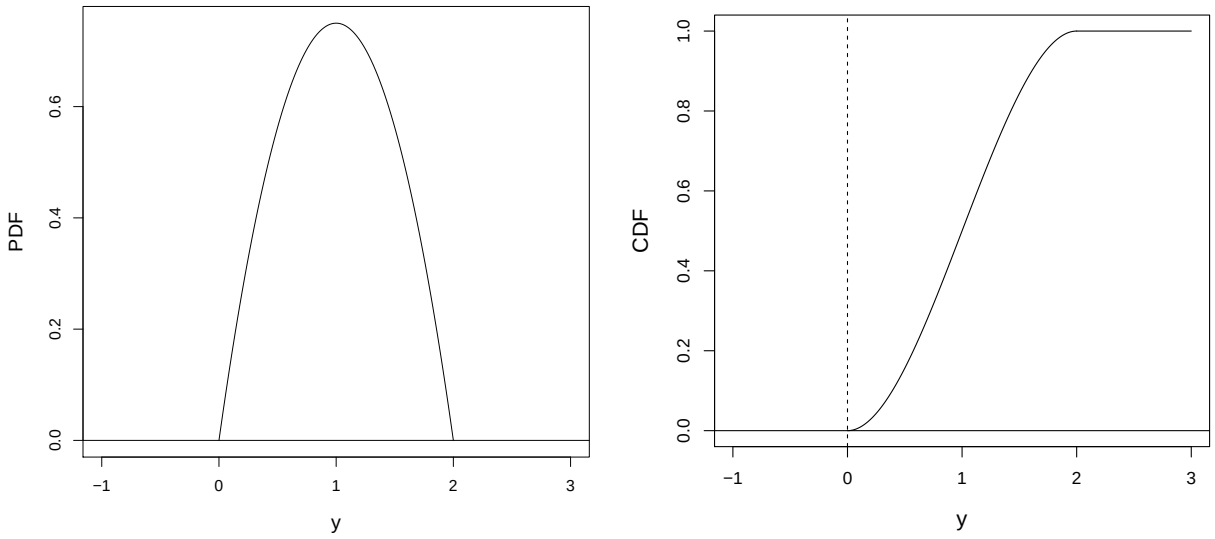

Figure 4.4: Pdf (left) and cdf (right) of Y in Example 4.4.

Summarizing, the cdf of Y is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ \frac{3}{8}\left(2y^2 - \frac{2}{3}y^3\right), & 0 < y < 2 \\ 1, & y \geq 2. \end{cases}$$

The pdf and cdf of Y are shown side by side in Figure 4.4 above. $\square$

Result: Suppose Y is a **continuous** random variable with pdf $f_Y(y)$. The probability of the event $\{Y \in B\}$ is found by integrating the pdf $f_Y(y)$ over the set B; i.e.,

$$P(Y \in B) = \int_B f_Y(y)dy.$$

For example, in Example 4.4,

$$\begin{aligned} P(Y \leq 0.5) = \int_0^{0.5} f_Y(y)dy &= \int_0^{0.5} \frac{3}{8}(4y - 2y^2)dy \\ &= \frac{3}{8}\left(2y^2 - \frac{2}{3}y^3\right)\Bigg|_0^{0.5} \\ &= \frac{3}{8}\left[2(0.5)^2 - \frac{2}{3}(0.5)^3\right] = F_Y(0.5) \approx 0.156. \end{aligned}$$

This calculation is depicted in Figure 4.5 (next page).

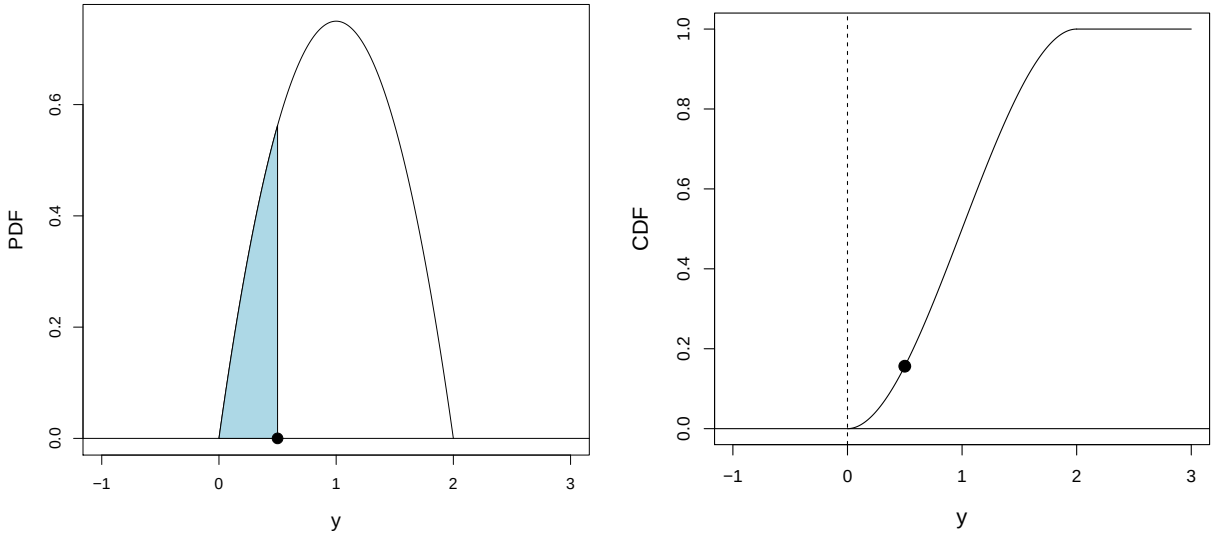

Figure 4.5: Pdf (left) and cdf (right) of Y in Example 4.4. Left: The shaded area equals $P(Y \leq 0.5) = \int_0^{0.5} f_Y(y)dy \approx 0.156$. Right: $P(Y \leq 0.5) = F_Y(0.5) \approx 0.156$.

Result: If Y is a **continuous** random variable with cdf $F_Y(y)$ and pdf $f_Y(y)$, then for any $a < b$,

$$P(a < Y < b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a \leq Y \leq b)$$

and each one equals

$$F_Y(b) - F_Y(a) = \int_a^b f_Y(y)dy.$$

Discussion: Instead of offering a rigorous proof of this result, we use intuition.

- The four probabilities above are the same because $P(Y = a) = 0$ and $P(Y = b) = 0$; recall that in continuous models, we assign zero probability to specific values. Therefore, in continuous distributions, the endpoints in $P(a \leq Y \leq b)$ do not influence the probability. Of course, this is **not** true in discrete distributions; i.e., the endpoints could be support points (which have positive probability).
- That $F_Y(b) - F_Y(a) = \int_a^b f_Y(y)dy$ is essentially an application of the Fundamental Theorem of Calculus.

Example 4.5. Suppose Y is a continuous random variable with cdf

$$F_Y(y) = \frac{1}{1 + e^{-y}}, \quad -\infty < y < \infty.$$

(a) Find the pdf of Y.
(b) Calculate $P(-2 < Y < 2)$ using the cdf and the pdf.

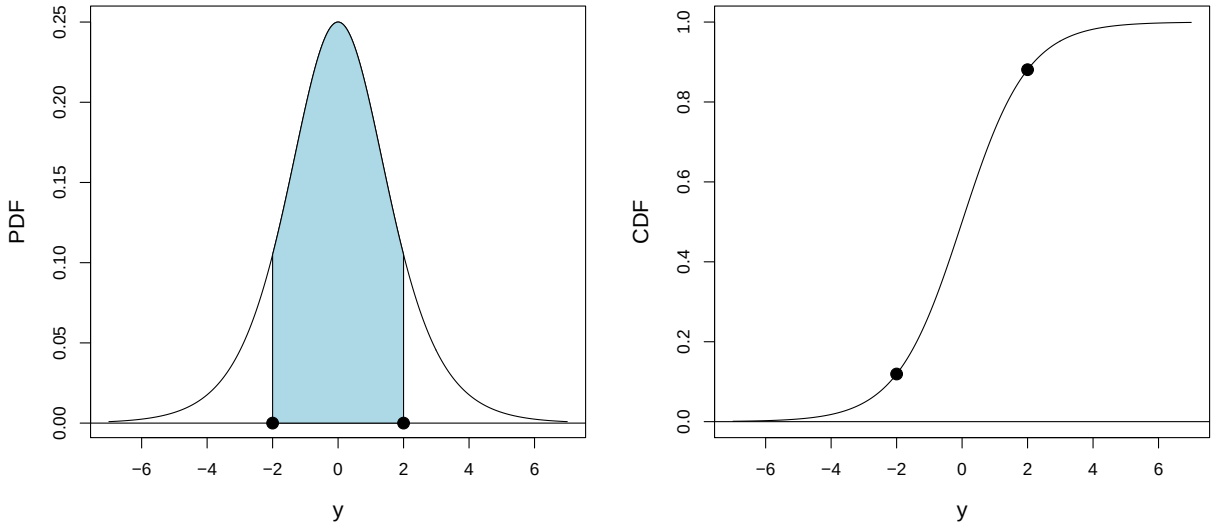

Figure 4.6: Pdf (left) and cdf (right) of Y in Example 4.5. Left: The shaded area equals $P(-2 < Y < 2) = \int_{-2}^{2} f_Y(y)dy \approx 0.762$. Right: $P(-2 < Y < 2) = F_Y(2) - F_Y(-2) \approx 0.762$.

Solutions. (a) The pdf of Y is the derivative of $F_Y(y)$; i.e.,

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}\left(\frac{1}{1+e^{-y}}\right) \\ &= (-1)(1+e^{-y})^{-2} \underbrace{\times \frac{d}{dy}(1+e^{-y})}_{\text{chain rule}} = \frac{e^{-y}}{(1+e^{-y})^2}. \end{aligned}$$

Therefore, the pdf of Y is

$$f_Y(y) = \begin{cases} \dfrac{e^{-y}}{(1+e^{-y})^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

(b) Using the cdf, we have

$$\begin{aligned} P(-2 < Y < 2) &= F_Y(2) - F_Y(-2) \\ &= \frac{1}{1+e^{-2}} - \frac{1}{1+e^{2}} \approx 0.881 - 0.119 = 0.762. \end{aligned}$$

Using the pdf, we have

$$P(-2 < Y < 2) = \int_{-2}^{2} f_Y(y)dy = \int_{-2}^{2} \frac{e^{-y}}{(1+e^{-y})^2}dy$$

To do this integral, let

$$u = 1 + e^{-y} \implies du = -e^{-y}dy.$$

With this u-substitution, we have

$$\int_{-2}^{2} \frac{e^{-y}}{(1+e^{-y})^2} dy = \int_{1+e^2}^{1+e^{-2}} -\frac{1}{u^2} du = \left(\frac{1}{u}\right) \Bigg|_{1+e^2}^{1+e^{-2}}$$
$$= \frac{1}{1+e^{-2}} - \frac{1}{1+e^2} \approx 0.881 - 0.119 = 0.762.$$

These calculations are depicted in Figure 4.6 (see previous page). □

Terminology: Suppose Y is a **continuous** random variable with cdf $F_Y(y)$ and pdf $f_Y(y)$. The p**th quantile** $(0 < p < 1)$ of Y, denoted by ϕ_p, is the smallest value that satisfies

$$P(Y \leq \phi_p) = \int_{-\infty}^{\phi_p} f_Y(y) dy = F_Y(\phi_p) = p.$$

Note: If $F_Y(y)$ is a strictly increasing function of y (which it often is), then the inverse $F_Y^{-1}(y)$ exists and hence

$$F_Y(\phi_p) = p \iff \phi_p = F_Y^{-1}(p).$$

Some authors prefer to call ϕ_p the $100p$**th percentile** of Y.

- For example, the $p = 0.5$ quantile (50th percentile) is called the **median** of Y.

Example 4.6. The amount of loss/damage (in millions of dollars) due to catastrophic weather is modeled as a continuous random variable Y with cdf

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ 1 - \left(\frac{10}{10+y}\right)^3, & y \geq 0. \end{cases}$$

(a) Find $\phi_{0.5}$, the median of Y.
(b) Find the pdf of Y and identify where $\phi_{0.5}$ falls.

Solutions. (a) We can set $F_Y(\phi_{0.5})$ equal to $p = 0.5$ and solve for $\phi_{0.5}$. That is,

$$0.5 \stackrel{\text{set}}{=} F_Y(\phi_{0.5}) = 1 - \left(\frac{10}{10+\phi_{0.5}}\right)^3 \implies \left(\frac{10}{10+\phi_{0.5}}\right)^3 = 0.5$$
$$\implies \frac{10}{10+\phi_{0.5}} = (0.5)^{1/3}$$
$$\implies \phi_{0.5} = \frac{10}{(0.5)^{1/3}} - 10 \approx 2.59921.$$

Therefore, the median loss is approximately \$2.6 million.

(b) For $y > 0$, the pdf of Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left[1 - \left(\frac{10}{10+y}\right)^3\right] = \frac{3000}{(10+y)^4}.$$

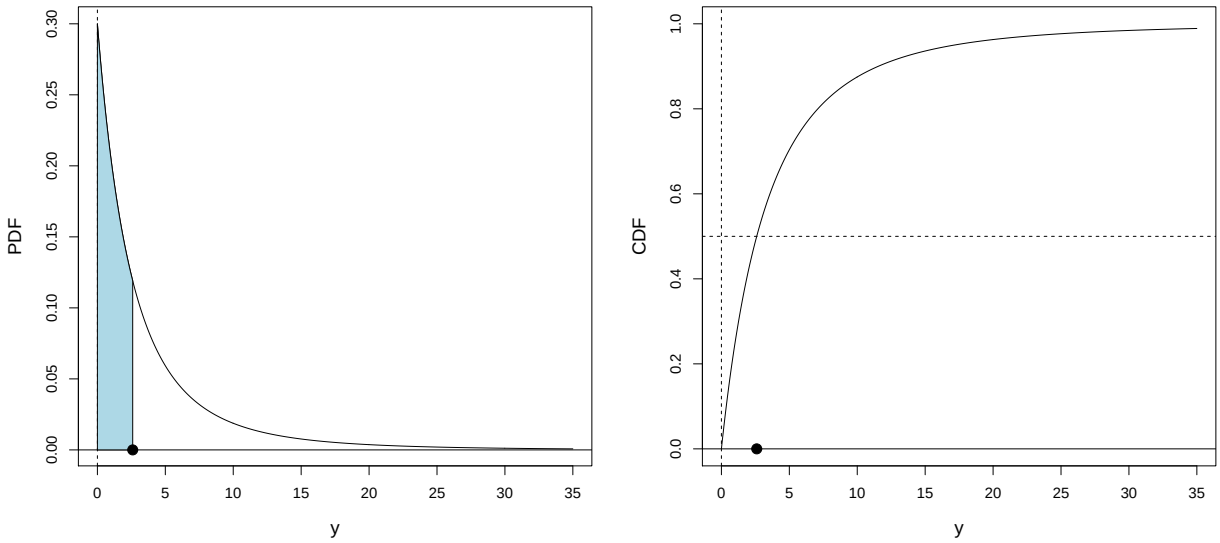

Figure 4.7: Pdf (left) and cdf (right) of Y in Example 4.6. Left: The shaded area equals 0.5. Right: A horizontal line at 0.5 has been added. In both figures, the median $\phi_{0.5} \approx 2.6$ is shown with a solid circle.

Therefore, the pdf of Y is

$$f_Y(y) = \begin{cases} \dfrac{3000}{(10+y)^4}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

The pdf and cdf of Y are shown side by side in Figure 4.7 above. $\square$

Remark: It makes perfect sense to talk about "quantiles" with discrete random variables. However, there are some potential problems with the definition that ϕ_p solves

$$F_Y(\phi_p) = P(Y \leq \phi_p) = p.$$

The reason is there may be no values of ϕ_p that satisfy this equation, or there could be an infinite number of values that solve it. For example, in Example 4.1 (see pp 73-74, notes),

- there is no value of $\phi_{0.5}$ that satisfies $F_Y(\phi_{0.5}) = 0.5$.
- there are an infinite number of values that solve $F_Y(\phi_{0.648}) = 0.648$; i.e., every number in $[1, 2)$.

We therefore have to alter the definition slightly to cover discrete distributions. The authors define the pth quantile ϕ_p more generally as the smallest value satisfying

$$F_Y(\phi_p) = P(Y \leq \phi_p) \geq p.$$

When Y is continuous and $F_Y(y)$ is strictly increasing, this more general definition reduces to what we gave earlier.

4.4 Mathematical expectation

Terminology: Suppose Y is a continuous random variable with pdf $f_Y(y)$. The **expected value** (or **mean**) of Y is

$$E(Y) = \int_{\mathbb{R}} y f_Y(y) dy.$$

We interpret $E(Y)$ in the same way as we did when Y was discrete (see pp 43, notes).

Technical note: For $E(Y)$ to exist, we need the integral above to converge absolutely; i.e.,

$$\int_{\mathbb{R}} |y| f_Y(y) dy < \infty.$$

Otherwise, we say that $E(Y)$ does not exist.

Terminology: Suppose Y is a continuous random variable with pdf $f_Y(y)$. The expected value of $g(Y)$ is

$$E[g(Y)] = \int_{\mathbb{R}} g(y) f_Y(y) dy,$$

provided that this integral converges absolutely. Otherwise, we say that $E[g(Y)]$ does not exist.

Properties: The expectation operator $E(\cdot)$ enjoys the same properties as in the discrete case (see pp 46-47, notes); i.e.,

1. $E(c) = c$, for any constant $c \in \mathbb{R}$
2. $E[cg(Y)] = cE[g(Y)]$
3. For real functions $g_1, g_2, ..., g_k$,

$$E\left[\sum_{j=1}^{k} g_j(Y)\right] = \sum_{j=1}^{k} E[g_j(Y)].$$

Exercise: Prove each of these results in the continuous case.

Terminology: Suppose Y is a continuous random variable with mean $E(Y) = \mu$. The **variance** of Y is

$$\sigma^2 = V(Y) = E[(Y-\mu)^2] = \int_{\mathbb{R}} (y-\mu)^2 f_Y(y) dy,$$

provided that this integral exists. The variance computing formula still applies, that is,

$$V(Y) = E(Y^2) - [E(Y)]^2.$$

The **standard deviation** of Y is the (positive) square root of the variance.

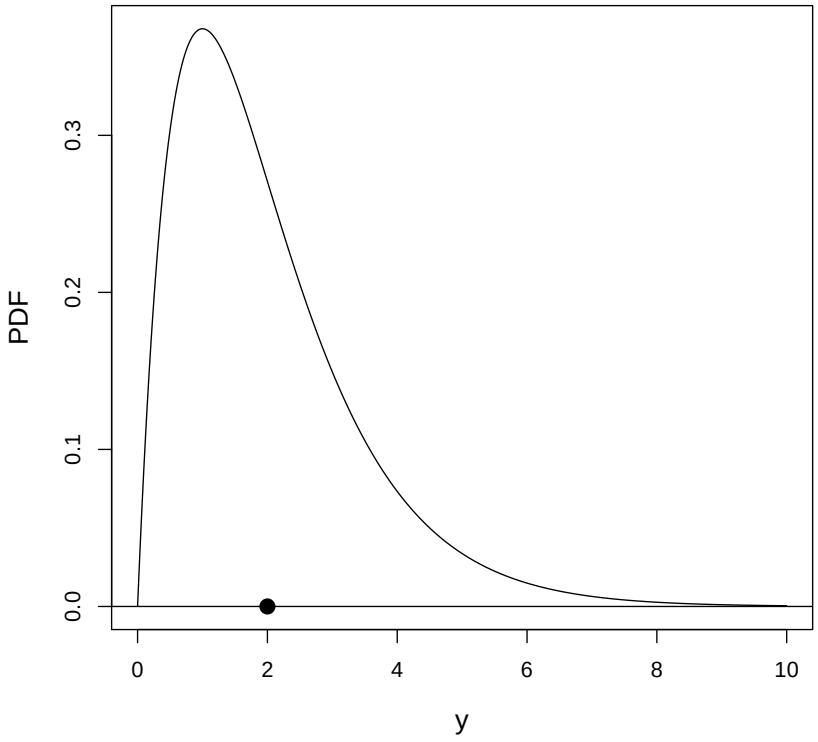

Figure 4.8: Pdf of Y in Example 4.7. The mean $E(Y) = 2$ is shown with a solid circle.

Example 4.7. The lifetime of an electrical component (in years) is modeled as a continuous random variable Y with pdf

$$f_Y(y) = \begin{cases} ye^{-y}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the mean and variance of Y.

Solutions. The mean of Y is

$$E(Y) = \int_{\mathbb{R}} y f_Y(y) dy = \int_0^\infty y \times ye^{-y} dy = \int_0^\infty y^2 e^{-y} dy.$$

This integral can be computed using integration by parts. Let

$$\begin{aligned} u &= y^2 \qquad & du &= 2ydy \\ dv &= e^{-y} \qquad & v &= -e^{-y}. \end{aligned}$$

With these selections,

$$\int_0^\infty y^2 e^{-y} dy = \underbrace{-y^2 e^{-y}\Big|_0^\infty}_{=\,0} + 2\underbrace{\int_0^\infty ye^{-y} dy}_{=\,1} = 2.$$

Therefore, the mean lifetime is $E(Y) = 2$ years. To find $V(Y)$, we could use the definition and calculate

$$V(Y) = E[(Y-\mu)^2] = \int_{\mathbb{R}} (y-\mu)^2 f_Y(y)dy = \int_0^\infty (y-2)^2 \times ye^{-y}dy,$$

or we could use the variance computing formula

$$V(Y) = E(Y^2) - [E(Y)]^2.$$

The second moment

$$E(Y^2) = \int_{\mathbb{R}} y^2 f_Y(y)dy = \int_0^\infty y^2 \times ye^{-y}dy = \int_0^\infty y^3 e^{-y}dy.$$

We use integration by parts again. Let

$$\begin{aligned} u &= y^3 \qquad & du &= 3y^2dy \\ dv &= e^{-y} \qquad & v &= -e^{-y}. \end{aligned}$$

With these selections,

$$\int_0^\infty y^3 e^{-y}dy = \underbrace{-y^3e^{-y}\Big|_0^\infty}_{=0} + 3\underbrace{\int_0^\infty y^2e^{-y}dy}_{=E(Y)=2} = 6.$$

Therefore,

$$V(Y) = E(Y^2) - [E(Y)]^2 = 6 - 4 = 2. \ \square$$

R: The `integrate` function in R can be helpful to calculate "messy" integrals or simply to check your work. Here is the code to find the first and second moments in Example 4.7:

```
# Calculate E(Y)
integrand <- function(y){y^2*exp(-y)}
integrate(integrand,lower=0,upper=Inf)
2 with absolute error < 7.1e-05

# Calculate E(Y^2)
integrand.2 <- function(y){y^3*exp(-y)}
integrate(integrand.2,lower=0,upper=Inf)
6 with absolute error < 2.6e-06
```

The `integrate` function in R uses numerical methods to calculate integrals within a certain level of "error."

Exercise: Calculate $E(Y)$ and $V(Y)$ in Examples 4.4, 4.5, and 4.6. Do this "by hand" first and then use R to check your work.

Terminology: Suppose Y is a continuous random variable with pdf $f_Y(y)$. The **moment-generating function (mgf)** of Y is

$$m_Y(t) = E(e^{tY}) = \int_{\mathbb{R}} e^{ty} f_Y(y) dy,$$

provided this expectation is finite for all t in an open neighborhood about $t = 0$; i.e., $\exists b > 0$ such that $E(e^{tY}) < \infty \; \forall t \in (-b, b)$. If no such $b > 0$ exists, then the moment generating function of Y does not exist.

Recall: If Y is a random variable with mgf $m_Y(t)$, then

$$E(Y^k) = m_Y^{(k)}(0),$$

where

$$m_Y^{(k)}(0) = \frac{d^k}{dt^k} m_Y(t) \bigg|_{t=0}.$$

Recall that this is also how we used the mgf to calculate moments like $E(Y)$ and $E(Y^2)$ in the discrete case.

Example 4.8. A continuous random variable Y is said to have an **exponential distribution** with parameter $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the mgf of Y.

Solution. The mgf of Y is

$$\begin{aligned} m_Y(t) = E(e^{tY}) = \int_{\mathbb{R}} e^{ty} f_Y(y) dy = \int_0^\infty e^{ty} \, \frac{1}{\beta} e^{-y/\beta} dy &= \frac{1}{\beta} \int_0^\infty e^{-y\left(\frac{1}{\beta} - t\right)} dy \\ &= \frac{1}{\beta} \left[-\frac{1}{\frac{1}{\beta} - t} \, e^{-y\left(\frac{1}{\beta} - t\right)} \bigg|_{y=0}^{\infty} \right] \\ &= \frac{1}{1 - \beta t} \left[e^{-y\left(\frac{1}{\beta} - t\right)} \bigg|_{\infty}^{0} \right] \\ &= \frac{1}{1 - \beta t} \left[1 - \lim_{y \to \infty} e^{-y\left(\frac{1}{\beta} - t\right)} \right]. \end{aligned}$$

Note that

$$\begin{aligned} \lim_{y \to \infty} e^{-y\left(\frac{1}{\beta} - t\right)} &= 0, \qquad \text{if } \frac{1}{\beta} - t > 0 \\ \lim_{y \to \infty} e^{-y\left(\frac{1}{\beta} - t\right)} &= +\infty, \qquad \text{if } \frac{1}{\beta} - t < 0. \end{aligned}$$

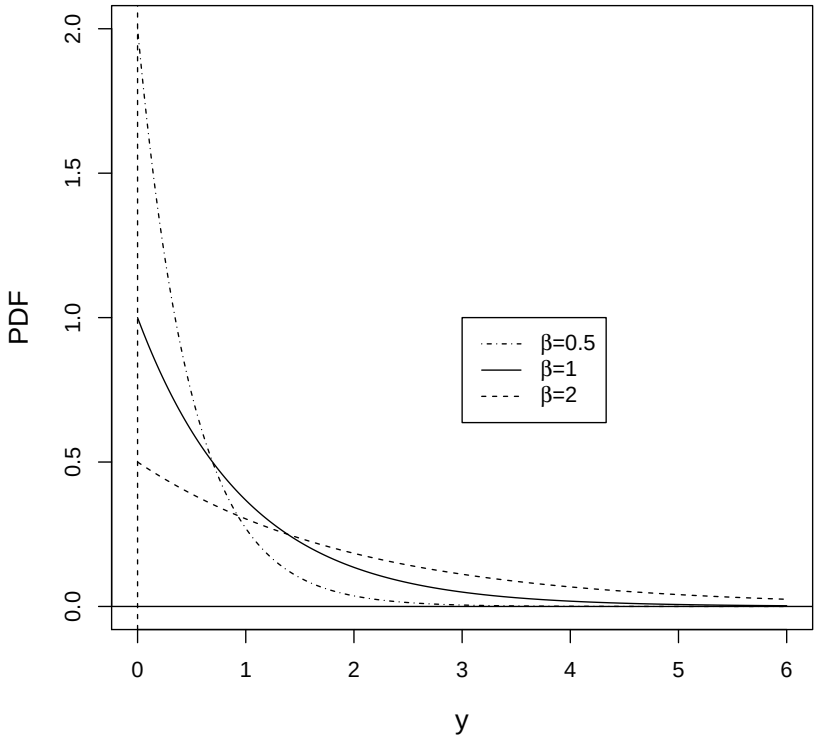

Figure 4.9: Exponential pdfs when $\beta = 0.5$, $\beta = 1$, and $\beta = 2$.

Therefore, provided that

$$\frac{1}{\beta} - t > 0 \iff t < \frac{1}{\beta},$$

the mgf of Y exists and is given by

$$m_Y(t) = \frac{1}{1-\beta t}.$$

Note that $\exists b > 0$ (e.g., $b = 1/\beta$) such that $m_Y(t) = E(e^{tY}) < \infty \; \forall t \in (-b, b)$. $\square$

Remark: The exponential distribution is widely used in engineering and actuarial science applications when modeling "time to event" random variables; e.g., the time until part failure, the time until a claim is made, etc. Figure 4.9 above shows the exponential pdf when $\beta = 0.5$, $\beta = 1$, and $\beta = 2$. All exponential pdfs have the same shape but decay with different scales.

Exercise: Use the mgf above to show that when Y has an exponential distribution with parameter $\beta > 0$, the mean and variance are given by $E(Y) = \beta$ and $V(Y) = \beta^2$, respectively. You can also show this by calculating $E(Y)$ and $E(Y^2)$ directly (i.e., not using the mgf), so do it both ways.

4.5 Uniform distribution

Terminology: A random variable Y is said to have a **uniform distribution** from θ_1 to θ_2 if its pdf is given by

$$f_Y(y) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1}, & \theta_1 < y < \theta_2 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \mathcal{U}(\theta_1, \theta_2)$. This is a valid pdf because $f_Y(y) \geq 0$ for all $y \in \mathbb{R}$ and

$$\int_{\mathbb{R}} f_Y(y)dy = \int_{\theta_1}^{\theta_2} \left(\frac{1}{\theta_2 - \theta_1}\right) dy = \left(\frac{y}{\theta_2 - \theta_1}\right)\Big|_{\theta_1}^{\theta_2} = \frac{\theta_2 - \theta_1}{\theta_2 - \theta_1} = 1.$$

CDF: The cdf of $Y \sim \mathcal{U}(\theta_1, \theta_2)$ is given by

$$F_Y(y) = \begin{cases} 0, & y \leq \theta_1 \\ \dfrac{y - \theta_1}{\theta_2 - \theta_1}, & \theta_1 < y < \theta_2 \\ 1, & y \geq \theta_2. \end{cases}$$

The form of this cdf makes sense; note that the pdf $f_Y(y)$ is a **constant** function of y. Its cdf (which is calculated by anti-differentiation) is a **linear** function of y.

Example 4.9. We observe a dart player throwing darts at a board. Let Y denote the angle of inclination from the horizontal axis drawn through the "bullseye." In this example, we might assume $Y \sim \mathcal{U}(0, 2\pi)$. The pdf of Y is

$$f_Y(y) = \begin{cases} \dfrac{1}{2\pi}, & 0 < y < 2\pi \\ 0, & \text{otherwise.} \end{cases}$$

The cdf of Y is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ \dfrac{y}{2\pi}, & 0 < y < 2\pi \\ 1, & y \geq 2\pi. \end{cases}$$

The pdf and cdf of Y are shown side by side in Figure 4.10 (see next page).

Q: What is the probability a throw lands in the first quadrant (formed by drawing horizontal and vertical axes through the bullseye)?
A: A first-quadrant throw results if and only the event $\{0 < Y < \pi/2\}$ occurs. Therefore,

$$P\left(0 < Y < \frac{\pi}{2}\right) = \int_0^{\pi/2} \frac{1}{2\pi}\, dy = \frac{\pi/2}{2\pi} = \frac{1}{4}.$$

Note that this answer also equals $F_Y(\pi/2) = P(Y \leq \pi/2)$. $\square$

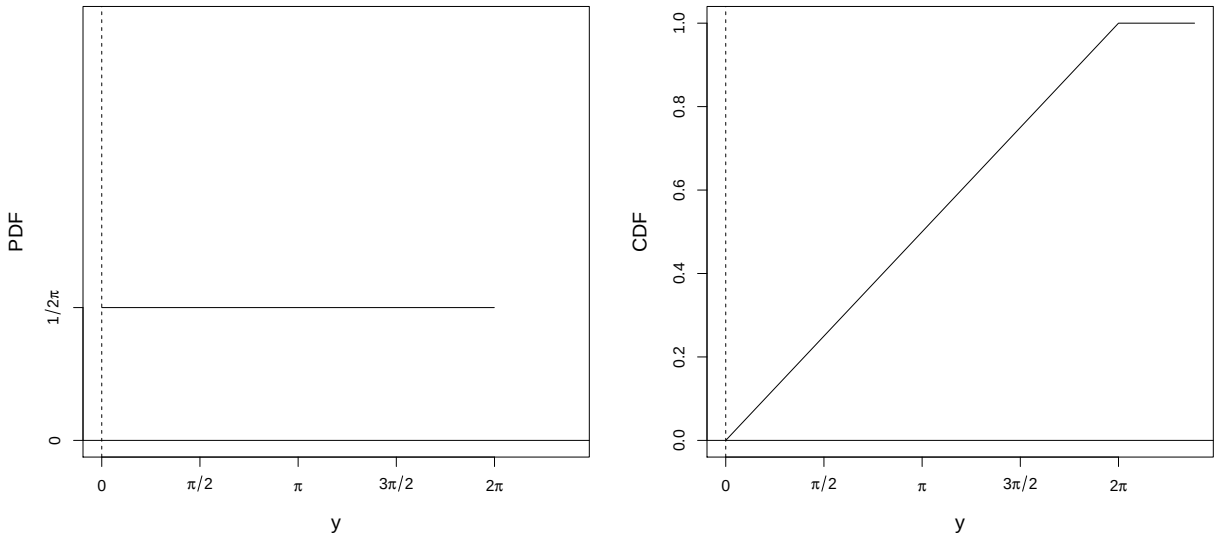

Figure 4.10: Pdf (left) and cdf (right) of $Y \sim \mathcal{U}(0, 2\pi)$ in Example 4.9.

Mean/Variance: If $Y \sim \mathcal{U}(\theta_1, \theta_2)$, then

$$\begin{aligned} E(Y) &= \frac{\theta_1 + \theta_2}{2} \\ V(Y) &= \frac{(\theta_2 - \theta_1)^2}{12}. \end{aligned}$$

MGF: The mgf of $Y \sim \mathcal{U}(\theta_1, \theta_2)$ is

$$m_Y(t) = \begin{cases} \dfrac{e^{\theta_2 t} - e^{\theta_1 t}}{t(\theta_2 - \theta_1)}, & t \neq 0 \\ 1, & t = 0. \end{cases}$$

Exercise: Verify these expressions above.

4.6 Normal distribution

Terminology: A random variable Y is said to have a **normal distribution** if its pdf is given by

$$f_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \mathcal{N}(\mu, \sigma^2)$. There are two parameters in the normal distribution: the mean μ and the variance σ^2. The standard deviation σ is the (positive) square root of the variance.

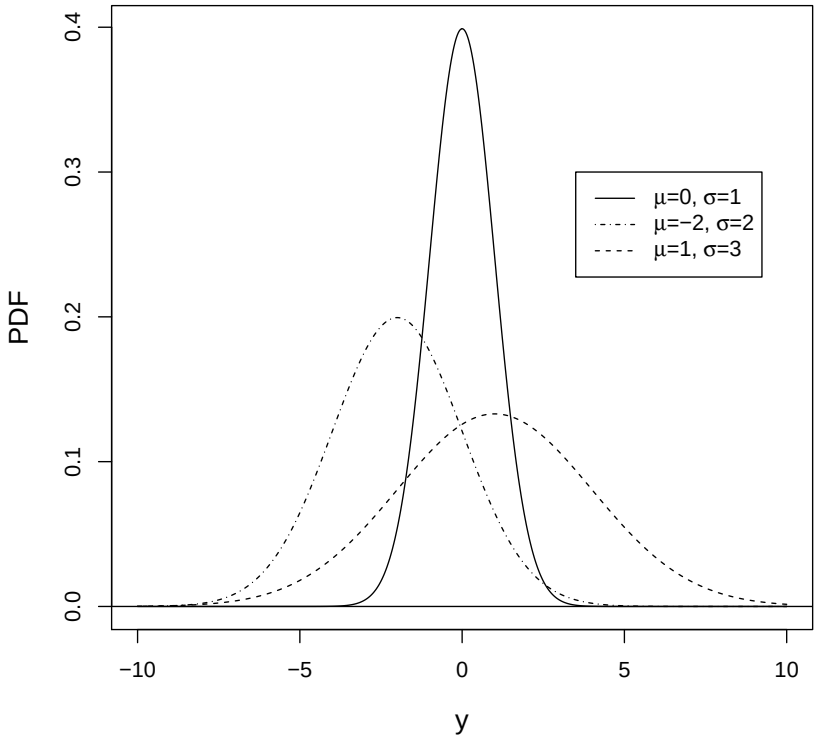

Figure 4.11: $\mathcal{N}(\mu = 0, \sigma^2 = 1)$, $\mathcal{N}(\mu = -2, \sigma^2 = 4)$, and $\mathcal{N}(\mu = 1, \sigma^2 = 9)$ pdfs.

Fact: The $\mathcal{N}(\mu, \sigma^2)$ pdf $f_Y(y)$ is **symmetric** about the mean μ; i.e.,

$$f_Y(\mu - a) = f_Y(\mu + a),$$

for all $a \in \mathbb{R}$.

Fact: The $\mathcal{N}(\mu, \sigma^2)$ pdf $f_Y(y)$ has **points of inflection** at $y = \mu \pm \sigma$.

Q: Is $f_Y(y)$ a valid pdf?
A: Clearly, $f_Y(y) \geq 0$ for all $y \in \mathbb{R}$. However, showing $\int_{\mathbb{R}} f_Y(y)dy = 1$ is not trivial. The reason why is that the antiderivative of $f_Y(y)$ does not exist in closed form. Define

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy.$$

We want to show that $I = 1$. Let

$$z = \frac{y - \mu}{\sigma} \implies dz = \frac{1}{\sigma} dy.$$

With this change of variable, the integral I above becomes

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Because $I > 0$, it suffices to show that $I^2 = 1$. Note that

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy\right) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\left(\frac{x^2+y^2}{2}\right)\right] dxdy. \end{aligned}$$

Switch to polar coordinates. Let

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

so that

$$x^2 + y^2 = r^2\cos^2\theta + r^2\sin^2\theta = r^2$$

and $dxdy = rdrd\theta$; i.e., the Jacobian of the transformation from (x, y) space to (r, θ) space. We have

$$\begin{aligned} I^2 = \frac{1}{2\pi}\int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} e^{-r^2/2}\, rdrd\theta &= \frac{1}{2\pi}\int_{\theta=0}^{2\pi}\left(\int_{r=0}^{\infty} re^{-r^2/2} dr\right) d\theta \\ &= \frac{1}{2\pi}\int_{\theta=0}^{2\pi}\left(-e^{-r^2/2}\,\Big|_{r=0}^{\infty}\right) d\theta = \frac{1}{2\pi}\int_{\theta=0}^{2\pi} d\theta = 1. \end{aligned}$$

Therefore, the $\mathcal{N}(\mu, \sigma^2)$ pdf $f_Y(y)$ is valid. $\square$

CDF: The cdf of $Y \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt,$$

for all $y \in \mathbb{R}$. This integral does not exist in closed form so it is of limited practical utility. The R function `pnorm` will calculate this probability upon request.

Mean/Variance: If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} E(Y) &= \mu \\ V(Y) &= \sigma^2. \end{aligned}$$

MGF: The mgf of $Y \sim \mathcal{N}(\mu, \sigma^2)$ is

$$m_Y(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Proof. Using the definition of the mgf, we have

$$m_Y(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty}\, \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{ty-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy.$$

Define $b = ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$, the exponent in the last integral. Note that

$$\begin{aligned} b = ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 &= ty - \frac{1}{2\sigma^2}(y^2 - 2\mu y + \mu^2) \\ &= -\frac{1}{2\sigma^2}(y^2 - 2\mu y - 2\sigma^2 ty + \mu^2) \\ &= -\frac{1}{2\sigma^2}\big[\underbrace{y^2 - 2(\mu + \sigma^2 t)y}_{\text{complete the square}} + \mu^2\big] \\ &= -\frac{1}{2\sigma^2}\Big[y^2 - 2(\mu+\sigma^2 t)y + \underbrace{(\mu+\sigma^2 t)^2 - (\mu+\sigma^2 t)^2}_{\text{add and subtract}} + \mu^2\Big] \\ &= -\frac{1}{2\sigma^2}\{[y - (\mu+\sigma^2 t)]^2\} + \frac{1}{2\sigma^2}[(\mu+\sigma^2 t)^2 - \mu^2] \\ &= -\frac{1}{2\sigma^2}(y-a)^2 + \frac{1}{2\sigma^2}(\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2) \\ &= -\frac{1}{2\sigma^2}(y-a)^2 + \underbrace{\mu t + \sigma^2 t^2/2}_{=\ c,\ \text{say}}, \end{aligned}$$

where $a = \mu + \sigma^2 t$. Note that $c = \mu t + \sigma^2 t^2/2$ is free of y. Therefore,

$$\begin{aligned} m_Y(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^b dy &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(y-a)^2 + c}\, dy \\ &= e^c \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-a)^2}}_{\mathcal{N}(a,\sigma^2)\ \text{pdf}} dy \ = \ e^c, \end{aligned}$$

because the $\mathcal{N}(a, \sigma^2)$ pdf integrates to 1 over $\mathbb{R} = (-\infty, \infty)$. Finally, note that

$$e^c = \exp(c) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right). \ \square$$

Exercise: Use the mgf $m_Y(t)$ to show $E(Y) = \mu$ and $V(Y) = \sigma^2$.

Terminology: A random variable Z is said to have a **standard normal distribution** if its pdf is given by

$$f_Z(z) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, & -\infty < z < \infty \\ 0, & \text{otherwise.} \end{cases}$$

A standard normal random variable Z arises when $\mu = 0$ and $\sigma^2 = 1$. Shorthand notation is $Z \sim \mathcal{N}(0, 1)$. An important result is that

$$Y \sim \mathcal{N}(\mu, \sigma^2) \implies Z = \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

In other words, any normal random variable Y can be "converted" into a standard normal random variable by a result known as **standardization**.

Proof. Let $Z = (Y - \mu)/\sigma$. The cdf of Z is given by

$$F_Z(z) = P(Z \leq z) = P\left(\frac{Y-\mu}{\sigma} \leq z\right) = P(Y \leq \sigma z + \mu) = F_Y(\sigma z + \mu).$$

Therefore, the pdf of Z is

$$\begin{aligned} f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}F_Y(\sigma z + \mu) &= f_Y(\sigma z + \mu) \times \frac{d}{dz}(\sigma z + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\sigma z + \mu - \mu)^2} \times \sigma = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \end{aligned}$$

which is the pdf of $\mathcal{N}(0,1)$ random variable. □

Implication: Because any normal random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$ can be transformed into a standard normal random variable $Z \sim \mathcal{N}(0,1)$, we can always write

$$\begin{aligned} P(a < Y < b) = P\left(\frac{a-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right), \end{aligned}$$

where $F_Z(\cdot)$ is the cdf of Z; i.e.,

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

This integral does not exist in closed form. However, probability tables exist that catalogue its value for different values of z (which are determined using numerical integration methods). For example, see Table 4 (pp 848, WMS), which catalogues values of $1 - F_Z(z)$. Before computing packages like R, these tables were needed. However, they are now somewhat outdated; e.g., the R command `pnorm(y,`μ`,`σ`)` calculates the cdf of any $\mathcal{N}(\mu, \sigma^2)$ random variable at the value y.

Example 4.10. The World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the systolic blood pressure (SBP) of American males (aged 18 and over). SBP is measured in millimeters of mercury (mmHg). Let Y denote the SBP of an individual selected from this population.

(a) An SBP of 90 mmHg or less is generally considered to be "low." Find $P(Y < 90)$.
(b) Find $\phi_{0.8}$, the $p = 0.8$ quantile (80th percentile) of this distribution.

Solutions. (a) We have

$$P(Y < 90) = \int_{-\infty}^{90} \underbrace{\frac{1}{\sqrt{2\pi}(15)} e^{-\frac{1}{2}\left(\frac{y-125}{15}\right)^2}}_{\mathcal{N}(125,15^2)\text{ pdf}} dy = F_Y(90),$$

where $F_Y(\cdot)$ denotes the $\mathcal{N}(125, 15^2)$ cdf. In R, this is calculated as

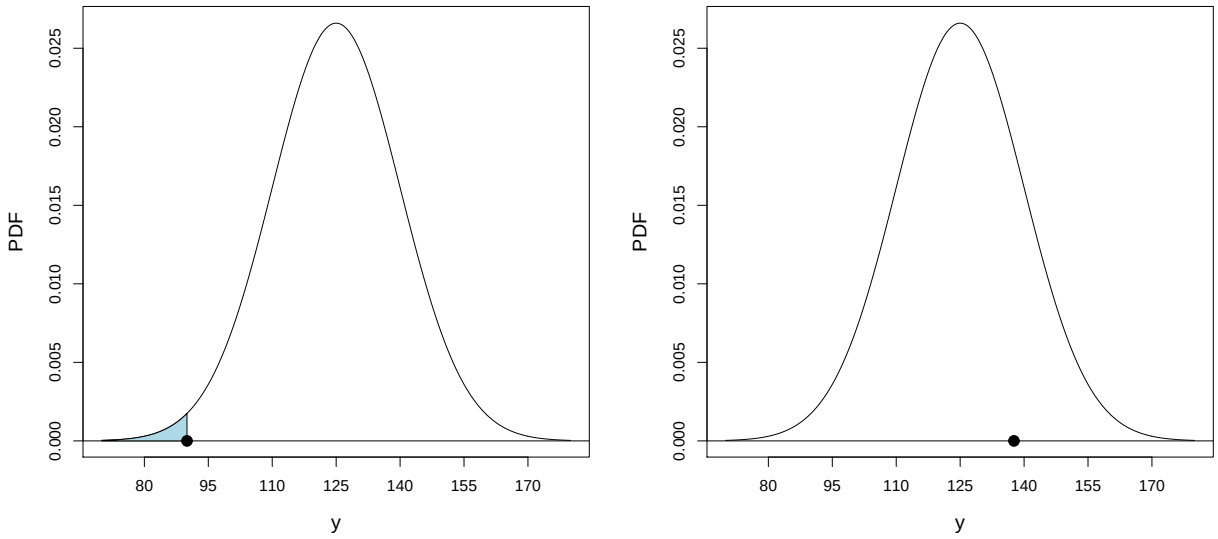

Figure 4.12: $\mathcal{N}(125, 15^2)$ pdf in Example 4.10. Left: The probability $P(Y < 90) \approx 0.0098$ is shown shaded. Right: The $p = 0.8$ quantile (80th percentile) $\phi_{0.8} \approx 137.6$ is shown with a solid circle.

```
> pnorm(90,125,15)
[1] 0.009815329
```

Therefore, the probability a randomly selected individual from this population has low SBP is about 0.01. Figure 4.12 (left) above shows this probability on the $\mathcal{N}(125, 15^2)$ pdf.

(b) The `qnorm` function in R calculates quantiles from any normal distribution. Here,

```
> qnorm(0.8,125,15)
[1] 137.6243
```

Therefore, 80 percent of this population has a SBP below 137.6 mmHg. Twenty percent of the population has a SBP above this value; see Figure 4.12 (right) above. □

4.7 The gamma family of distributions

Note: There are three popular "named distributions" in the gamma family:

- the exponential distribution
- the gamma distribution
- the χ^2 distribution.

We were introduced to the exponential distribution in Example 4.8 (pp 88, notes).

4.7.1 Exponential distribution

Terminology: A random variable Y is said to have an **exponential distribution** with parameter $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta}e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim$ exponential(β). The value β determines the scale of the distribution, so it is called a **scale parameter**. This is a valid pdf because $f_Y(y) \geq 0$, for all $y \in \mathbb{R}$ and

$$\int_{\mathbb{R}} f_Y(y)dy = \int_0^\infty \frac{1}{\beta}e^{-y/\beta}dy = \frac{1}{\beta}\left(-\beta e^{-y/\beta}\Big|_0^\infty\right) = e^{-0/\beta} - \lim_{y\to\infty} e^{-y/\beta} = 1 - 0 = 1.$$

CDF: The cdf of $Y \sim$ exponential(β) is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\beta}, & y > 0. \end{cases}$$

Exercise: Verify this expression. Note that the cdf of $Y \sim$ exponential(β) exists in closed form.

Mean/Variance: If $Y \sim$ exponential(β), then

$$\begin{aligned} E(Y) &= \beta \\ V(Y) &= \beta^2. \end{aligned}$$

MGF: The mgf of $Y \sim$ exponential(β) is

$$m_Y(t) = \frac{1}{1-\beta t}, \quad \text{for } t < \frac{1}{\beta}.$$

Example 4.11. "Time to event" studies are common in medical applications. One recent study involved patients with venous ulcers (also known as leg ulcers). A short-stretch bandage was applied to each patient's infected leg area, and investigators recorded

$$Y = \text{the time (in days) until the leg ulcer was completely healed.}$$

Suppose Y has an exponential distribution with mean $\beta = 190$.

(a) Calculate $P(Y > 100)$; i.e., the probability the ulcer takes longer than 100 days to heal.
(b) Find $\phi_{0.9}$, the $p = 0.9$ quantile (90th percentile) of the distribution of Y.

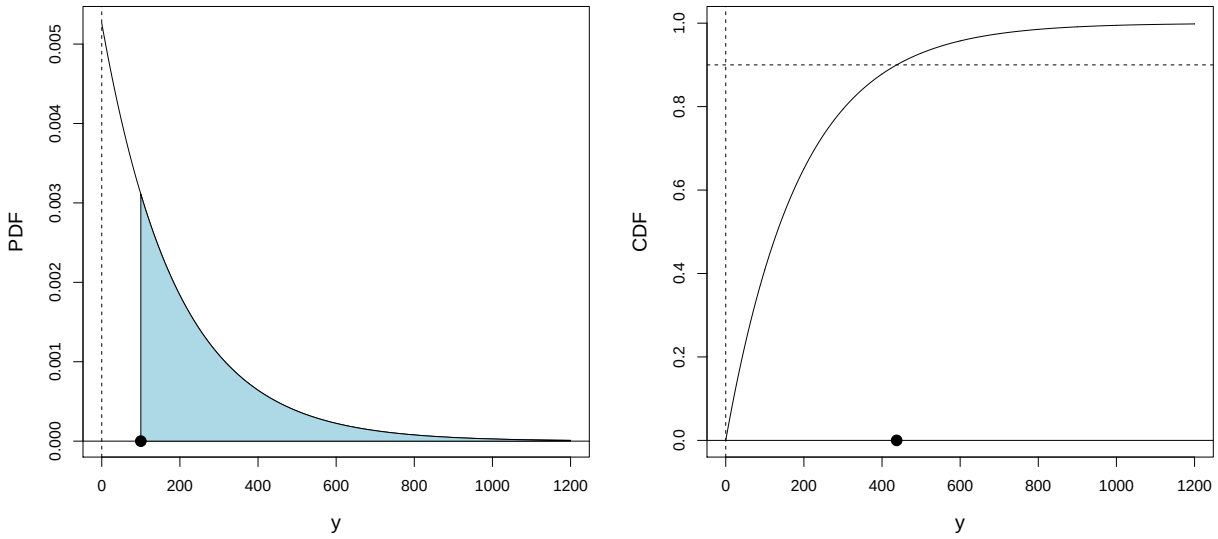

Figure 4.13: Pdf and cdf of $Y \sim$ exponential($\beta = 190$) in Example 4.11. Left: The probability $P(Y > 100) \approx 0.591$ is shown shaded on the pdf. Right: The $p = 0.9$ quantile (90th percentile) $\phi_{0.9} \approx 437.5$ is shown with a solid circle on the cdf. A horizontal line at 0.9 has been added.

Solutions. (a) Using the pdf of Y, we have

$$\begin{aligned} P(Y > 100) = \int_{100}^{\infty} \frac{1}{190} e^{-y/190} dy &= \frac{1}{190}\left(-190 e^{-y/190}\Big|_{100}^{\infty}\right) \\ &= e^{-100/190} - \underbrace{\lim_{y\to\infty} e^{-y/190}}_{=\,0} \approx 0.591. \end{aligned}$$

Figure 4.13 (left) above shows this probability on the exponential($\beta = 190$) pdf. Note that because the cdf $F_Y(y) = P(Y \leq y) = 1 - e^{-y/\beta}$ is in closed form, we could have computed

$$\begin{aligned} P(Y > 100) = 1 - P(Y \leq 100) &= 1 - F_Y(100) \\ &= 1 - (1 - e^{-100/190}) \approx 0.591 \end{aligned}$$

directly.

(b) The $p = 0.90$ quantile $\phi_{0.9}$ solves $F_Y(\phi_{0.9}) = P(Y \leq \phi_{0.9}) = 0.9$; see Figure 4.13 (right) above. We have

$$1 - e^{-\phi_{0.9}/190} \stackrel{\text{set}}{=} 0.9 \implies e^{-\phi_{0.9}/190} = 0.1 \implies -\frac{\phi_{0.9}}{190} = \ln(0.1)$$

Solving for $\phi_{0.9}$ gives

$$\phi_{0.9} = -190 \ln(0.1) \approx 437.5 \text{ days.}$$

Therefore, 90 percent of the patients' healing times will be less than 437.5 days (10 percent of the healing times will be greater than this). □

Memoryless Property: Suppose $Y \sim$ exponential(β) and let r and s be positive constants. Then

$$P(Y > r + s|Y > r) = P(Y > s).$$

That is, given that Y has exceeded r, the probability Y exceeds $r + s$ (i.e., an additional s units) is the same as if we were to look at Y unconditionally lasting until time s. The exponential distribution is the only continuous distribution that has this property.

Exponential/Poisson connection: Suppose we observe occurrences according to a Poisson process with rate $\lambda = 1/\beta$ and define

$$W = \text{ the time until the } \textbf{first} \text{ occurrence.}$$

The random variable W has an exponential distribution with mean β.
Proof. Clearly, W is a continuous random variable with nonnegative support. Thus, for $w > 0$, the cdf of W is

$$\begin{aligned} F_W(w) = P(W \leq w) &= 1 - P(W > w) \\ &= 1 - P(\{\text{no occurrences in } [0, w]\}) \\ &= 1 - \frac{e^{-\lambda w}(\lambda w)^0}{0!} = 1 - e^{-\lambda w}. \end{aligned}$$

Substituting $\lambda = 1/\beta$, we have $F_W(w) = 1 - e^{-w/\beta}$, the cdf of an exponential random variable with mean β. Thus, the result follows. □

Example 4.12. Customers arrive at a checkout counter according to a Poisson process with mean $\lambda = 10$ per hour. What is the probability it will take longer than 15 minutes for the first customer to arrive? Note that 15 minutes = 0.25 hour.

Solution. Let W denote the time until the first arrival; we know $W \sim$ exponential($\beta = 1/10$). Therefore,

$$\begin{aligned} P(W > 0.25) = 1 - P(W \leq 0.25) &= 1 - F_W(0.25) \quad \longleftarrow \quad (F_W \text{ is the cdf of } W) \\ &= 1 - (1 - e^{-0.25/(1/10)}) \\ &\approx 0.082. \ \square \end{aligned}$$

4.7.2 Gamma distribution

Terminology: A random variable Y is said to have a **gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim$ gamma(α, β).

Note: The gamma distribution is indexed by two parameters:

$$\begin{aligned}\alpha &= \textbf{shape} \text{ parameter} \\ \beta &= \textbf{scale} \text{ parameter.}\end{aligned}$$

Gamma function: For $\alpha > 0$, define the function

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}du.$$

The gamma function satisfies certain properties:

1. $\Gamma(1) = 1$
2. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$
3. $\Gamma(1/2) = \sqrt{\pi}$.

Note that if $\alpha \in \mathbb{N} = \{1, 2, 3, ..., \}$, then second (recursive) property implies

$$\Gamma(\alpha) = (\alpha-1)!$$

Important: When $\alpha = 1$, the gamma(α, β) distribution reduces to the exponential(β) distribution; note that

$$\frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta} \overset{\alpha=1}{=} \frac{1}{\Gamma(1)\beta^1}y^{1-1}e^{-y/\beta} = \frac{1}{\beta}e^{-y/\beta},$$

which is the exponential(β) pdf. Therefore, one can think of the gamma distribution as a generalization of the exponential. By introducing an extra parameter, α, we can get the pdf to assume more flexible shapes; see Figure 4.14 (next page).

Q: Is $f_Y(y)$ a valid pdf?
A: Clearly, $f_Y(y) \geq 0$ for all $y \in \mathbb{R}$. To show that $f_Y(y)$ integrates to 1, let

$$u = \frac{y}{\beta} \implies du = \frac{1}{\beta}dy.$$

We have

$$\begin{aligned}\int_{\mathbb{R}} f_Y(y)dy = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}dy &= \frac{1}{\Gamma(\alpha)}\int_0^\infty \frac{1}{\beta^\alpha}(\beta u)^{\alpha-1}e^{-\beta u/\beta} \times \beta du \\ &= \frac{1}{\Gamma(\alpha)}\int_0^\infty u^{\alpha-1}e^{-u}du \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1.\end{aligned}$$

Therefore, $f_Y(y)$ is a valid pdf. $\square$

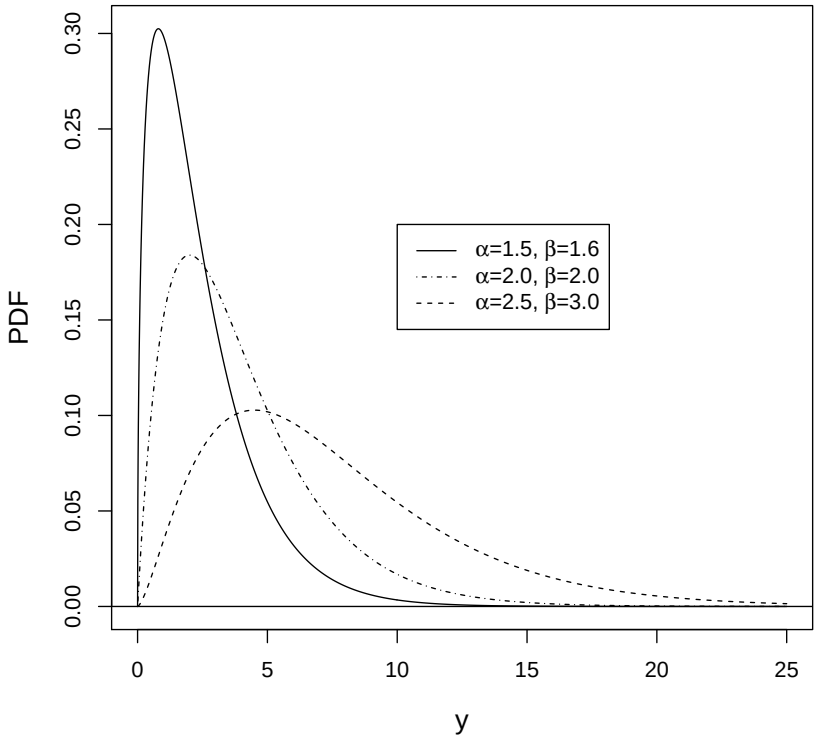

Figure 4.14: Gamma($\alpha = 1.5, \beta = 1.6$), gamma($\alpha = 2, \beta = 2$), and gamma($\alpha = 2.5, \beta = 3$) pdfs.

Important: Upon closer inspection, we see the nonzero part of the gamma(α, β) pdf

$$f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}$$

consists of two parts:

- the **kernel** of the pdf: $y^{\alpha-1}e^{-y/\beta}$
- the **constant** out front: $1/\Gamma(\alpha)\beta^\alpha$.

The kernel is the "guts" of the formula, while the constant out front is simply the "right quantity" that makes $f_Y(y)$ a valid pdf; i.e., the constant that makes $f_Y(y)$ integrate to 1. As such,

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy = 1 \quad \Longrightarrow \quad \int_0^\infty y^{\alpha-1} e^{-y/\beta} dy \; = \; \Gamma(\alpha)\beta^\alpha.$$

This result is extremely useful and will be used repeatedly. For example,

$$\int_0^\infty y^4 e^{-y/3} dy = \int_0^\infty y^{5-1} e^{-y/3} dy \; = \; \Gamma(5)3^5 = 4! \times 243 = 5832.$$

Mean/Variance: If $Y \sim \text{gamma}(\alpha, \beta)$, then

$$\begin{aligned} E(Y) &= \alpha\beta \\ V(Y) &= \alpha\beta^2. \end{aligned}$$

MGF: The mgf of $Y \sim \text{gamma}(\alpha, \beta)$ is

$$m_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, \quad \text{for } t < \frac{1}{\beta}.$$

Note that these formulas reduce to those for the exponential distribution when $\alpha = 1$; see pp 97 (notes). To see why the mgf formula is correct, note that

$$m_Y(t) = E(e^{tY}) = \int_{\mathbb{R}} e^{ty} f_Y(y)dy = \int_0^{\infty} e^{ty} \times \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta} dy.$$

In the integrand, write

$$e^{ty}e^{-y/\beta} = e^{ty-y/\beta} = e^{-y[(1/\beta)-t]} = e^{-y/[(1/\beta)-t]^{-1}} = e^{-y/\gamma},$$

where

$$\gamma = [(1/\beta) - t]^{-1}.$$

Therefore, $m_Y(t)$ can be written as

$$\frac{\gamma^{\alpha}}{\gamma^{\alpha}} \int_0^{\infty} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\gamma} dy = \frac{\gamma^{\alpha}}{\beta^{\alpha}} \underbrace{\int_0^{\infty} \frac{1}{\Gamma(\alpha)\gamma^{\alpha}} y^{\alpha-1} e^{-y/\gamma} dy}_{=1 \text{ (see note below)}} = \frac{1}{\beta^{\alpha}} \left(\frac{1}{\frac{1}{\beta} - t}\right)^{\alpha}$$

$$= \frac{1}{\beta^{\alpha}} \left(\frac{\beta}{1-\beta t}\right)^{\alpha} = \left(\frac{1}{1-\beta t}\right)^{\alpha}.$$

Note: The integral above is equal to 1 because the integrand is the gamma(α, γ) pdf and the integral is over $(0, \infty)$. However, for this to be true, γ cannot be negative or else the integral diverges. We must require

$$\gamma = [(1/\beta) - t]^{-1} > 0 \iff t < \frac{1}{\beta}.$$

Note that $\exists b > 0$ (e.g., $b = 1/\beta$) such that $m_Y(t) = E(e^{tY}) < \infty \; \forall t \in (-b, b)$. $\square$

Exercise: Verify the formulas for $E(Y)$ and $V(Y)$ above. You can do this by using the mgf or by using the definition of mathematical expectation directly.

Example 4.13. The time to death (Y, measured in days) for patients with a serious type of advanced tongue cancer follows a gamma distribution with $\alpha = 2.7$ and $\beta = 100$. What is the probability a patient with this type of cancer will live longer than one year? Note that 1 year is 365 days.

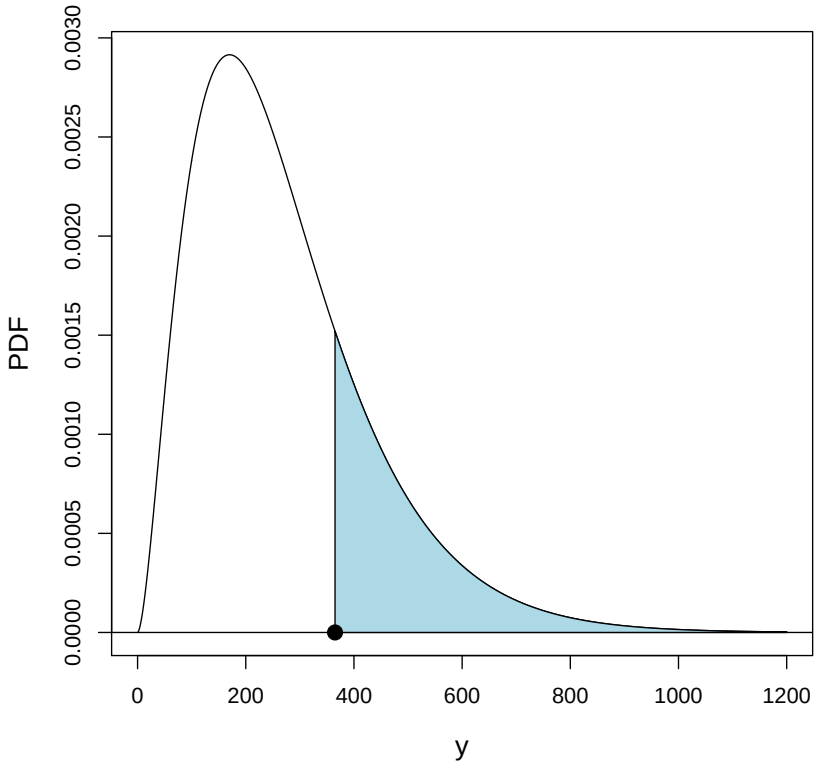

Figure 4.15: Pdf of $Y \sim$ gamma($\alpha = 2.7, \beta = 100$) in Example 4.13. The probability $P(Y > 365) \approx 0.236$ is shown shaded.

Solution. We want to compute $P(Y > 365)$, where $Y \sim$ gamma($\alpha = 2.7, \beta = 100$); i.e.,

$$P(Y > 365) = \int_{365}^{\infty} \frac{1}{\Gamma(2.7)100^{2.7}}\, y^{2.7-1}e^{-y/100}dy.$$

This integral must be calculated numerically because the gamma(α, β) cdf does not exist in closed form (unless $\alpha = 1$). Note that

$$P(Y > 365) = 1 - P(Y \leq 365) = 1 - F_Y(365),$$

where $F_Y(\cdot)$ is the gamma($\alpha = 2.7, \beta = 100$) cdf. The R function `pgamma` calculates the gamma cdf.

```
> 1-pgamma(365,2.7,1/100)
[1] 0.2355346
```

The pdf of $Y \sim$ gamma($\alpha = 2.7, \beta = 100$) is shown in Figure 4.15 above. □

Gamma/Poisson connection: Suppose we observe occurrences according to a Poisson process with rate $\lambda = 1/\beta$ and define

$$W = \text{ the time until the } \alpha\textbf{th} \text{ occurrence.}$$

The random variable $W \sim \text{gamma}(\alpha, \beta)$.

Proof. Clearly, W is a continuous random variable with nonnegative support. Thus, for $w > 0$, the cdf of W is

$$\begin{aligned} F_W(w) = P(W \leq w) &= 1 - P(W > w) \\ &= 1 - P(\{\text{fewer than } \alpha \text{ occurrences in } [0, w]\}) \\ &= 1 - \sum_{j=0}^{\alpha-1} \frac{(\lambda w)^j e^{-\lambda w}}{j!}. \end{aligned}$$

The pdf of W, for $w > 0$, is given by

$$\begin{aligned} f_W(w) = \frac{d}{dw} F_W(w) &= \lambda e^{-\lambda w} - e^{-\lambda w} \underbrace{\sum_{j=1}^{\alpha-1} \left[\frac{j(\lambda w)^{j-1}\lambda}{j!} - \frac{(\lambda w)^j \lambda}{j!} \right]}_{\text{telescoping sum}} \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \right] \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\lambda w}, \end{aligned}$$

which is the pdf of $W \sim \text{gamma}(\alpha, \beta)$, where $\beta = 1/\lambda$. □

Remark: In Example 4.12 (pp 99, notes), the time until the first customer arrives ($\alpha = 1$) follows an exponential distribution with mean $\beta = 1/10$.

- The time until the second customer arrives follows a gamma($\alpha = 2, \beta = 1/10$) distribution.
- The time until the third customer arrives follows a gamma($\alpha = 3, \beta = 1/10$) distribution, and so on.

4.7.3 χ^2 distribution

Terminology: A random variable Y is said to have a χ^2 **distribution** with $\nu > 0$ degrees of freedom if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\frac{\nu}{2}) 2^{\nu/2}} y^{(\nu/2)-1} e^{-y/2}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \chi^2(\nu)$. In practice, $\nu > 0$ is usually an integer but it doesn't have to be. Note that a $\chi^2(\nu)$ distribution is a gamma(α, β) distribution where

$$\alpha = \frac{\nu}{2} \quad \text{and} \quad \beta = 2.$$

Mean/Variance: If $Y \sim \chi^2(\nu)$, then

$$\begin{aligned} E(Y) &= \nu \\ V(Y) &= 2\nu. \end{aligned}$$

MGF: The mgf of $Y \sim \chi^2(\nu)$ is

$$m_Y(t) = \left(\frac{1}{1-2t}\right)^{\nu/2}, \quad \text{for } t < \frac{1}{2}.$$

Note: The $\chi^2(\nu)$ cdf does not exist in closed form; probabilities and quantiles associated with the $\chi^2(\nu)$ distribution can be calculated in R using the `pchisq` and `qchisq` functions, respectively.

4.8 Beta distribution

Terminology: A random variable Y is said to have a **beta distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, y^{\alpha-1}(1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \text{beta}(\alpha, \beta)$. Note that the support of a beta random variable Y is $R = \{y : 0 < y < 1\}$.

Remark: The nonzero part of the beta pdf is sometimes written as

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} = \frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1},$$

where the constant

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

In analysis, $B(\alpha, \beta)$ is called the **beta function**. Therefore, the nonzero part of the beta(α, β) pdf consists of two parts:

- the **kernel** of the pdf: $y^{\alpha-1}(1-y)^{\beta-1}$
- the **constant** out front: $1/B(\alpha, \beta)$.

The kernel is the "guts" of the formula, while the constant out front is simply the "right quantity" that makes $f_Y(y)$ a valid pdf; i.e., the constant that makes $f_Y(y)$ integrate to 1.

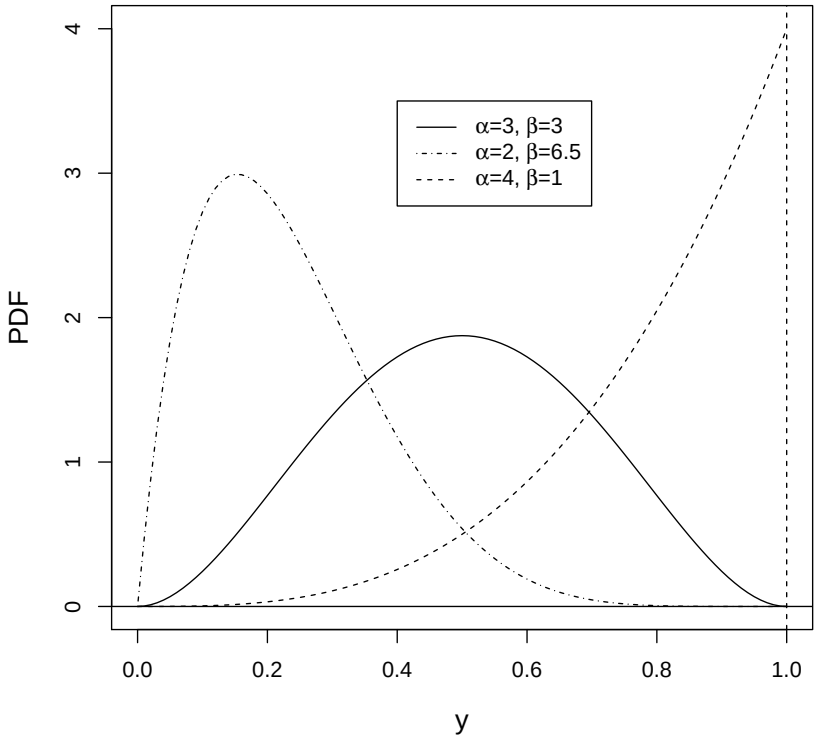

Figure 4.16: Beta($\alpha = 3, \beta = 3$), beta($\alpha = 2, \beta = 6.5$), and beta($\alpha = 4, \beta = 1$) pdfs.

Remark: The pdf of $Y \sim \text{beta}(\alpha, \beta)$ is very flexible; i.e., the pdf $f_Y(y)$ can assume many shapes over $R = \{y : 0 < y < 1\}$; see Figure 4.16 above. For example,

1. $\alpha = \beta \Longrightarrow f_Y(y)$ is symmetric about $y = 1/2$
   - $\alpha = \beta = 1 \Longrightarrow Y \sim \mathcal{U}(0, 1)$
2. $\alpha > \beta \Longrightarrow f_Y(y)$ is skewed left
3. $\alpha < \beta \Longrightarrow f_Y(y)$ is skewed right.

CDF/MGF: The cdf of $Y \sim \text{beta}(\alpha, \beta)$, for $0 < y < 1$, can be written as

$$F_Y(y) = P(Y \leq y) = \int_0^y \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1} dt,$$

which is called the **incomplete beta function**. However, this can not be simplified further in general (it can for certain values of α and/or β). Probabilities and quantiles associated with the beta(α, β) distribution can be calculated in R using the `pbeta` and `qbeta` functions, respectively. The mgf of $Y \sim \text{beta}(\alpha, \beta)$ exists, but its form is not very friendly.

Mean/Variance: If $Y \sim \text{beta}(\alpha, \beta)$, then

$$\begin{aligned} E(Y) &= \frac{\alpha}{\alpha+\beta} \\ V(Y) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

Proof. We will derive $E(Y)$ only. From the definition of expected value, we have

$$\begin{aligned} E(Y) = \int_{\mathbb{R}} y f_Y(y) dy &= \int_0^1 y \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \underbrace{y^{(\alpha+1)-1}(1-y)^{\beta-1}}_{\text{beta}(\alpha+1,\beta)\ \text{kernel}} dy. \end{aligned}$$

The last integrand is a beta kernel with parameters $\alpha+1$ and β. Because integration is over $R = \{y : 0 < y < 1\}$, the integral

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

Therefore,

$$\begin{aligned} E(Y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

To derive $V(Y)$, first find $E(Y^2)$ by using similar calculations. Then use the variance computing formula $V(Y) = E(Y^2) - [E(Y)]^2$ and simplify. □

Example 4.14. A filling station is supplied with gasoline once per day. Its daily volume in sales (Y, measured in 100,000s of gallons) is a random variable with pdf

$$f_Y(y) = \begin{cases} 5(1-y)^4, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find $E(Y)$, the mean daily volume in sales.
(b) What must the filling station's capacity be in order to have the probability of the supply being exhausted in a given day be 0.01?

Solutions. This is a beta$(1, 5)$ pdf. To see why note that the constant

$$\frac{\Gamma(1+5)}{\Gamma(1)\Gamma(5)} = \frac{\Gamma(6)}{\Gamma(5)} = \frac{5\Gamma(5)}{\Gamma(5)} = 5.$$

The kernel

$$(1-y)^4 = y^{\alpha-1}(1-y)^{\beta-1},$$

where $\alpha = 1$ and $\beta = 5$. Therefore, $Y \sim \text{beta}(1, 5)$. The pdf of Y is shown in Figure 4.17 (see next page).

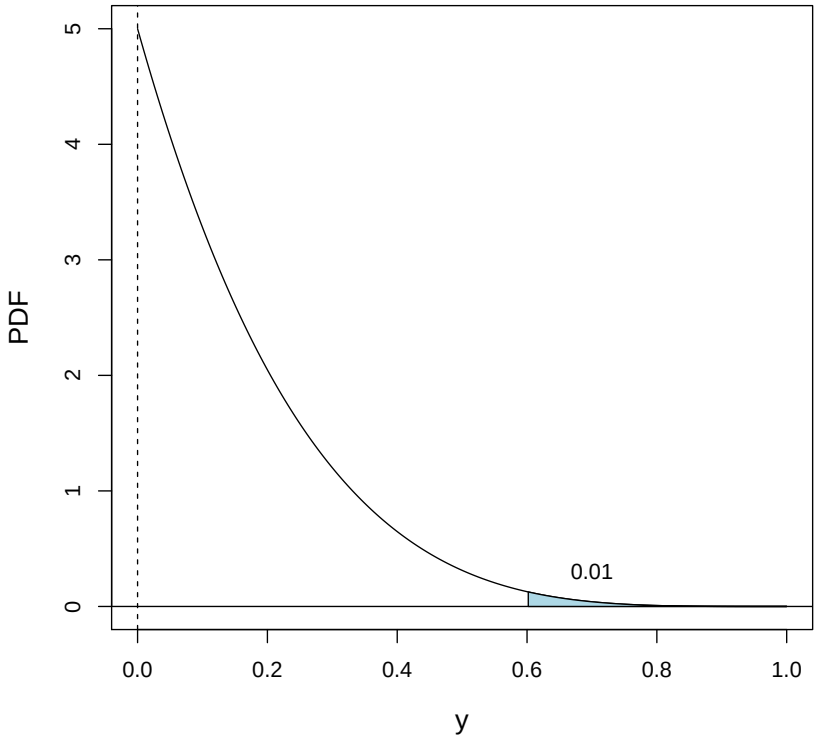

Figure 4.17: Pdf of $Y \sim \text{beta}(\alpha = 1, \beta = 5)$ in Example 4.14. The right tail probability is 0.01.

(a) The mean of $Y \sim \text{beta}(\alpha = 1, \beta = 5)$ is

$$E(Y) = \frac{1}{1+5} = \frac{1}{6}.$$

Therefore, the mean daily volume in sales is 16,666.67 gallons.

(b) We want to find $\phi_{0.99}$, the $p = 0.99$ quantile (99th percentile) of Y. This quantile solves

$$0.99 \stackrel{\text{set}}{=} \int_0^{\phi_{0.99}} 5(1-y)^4 dy.$$

To do this integral, let $u = 1 - y \Longrightarrow du = -dy$ so that

$$0.99 \stackrel{\text{set}}{=} \int_1^{1-\phi_{0.99}} -5u^4 \, du = \int_{1-\phi_{0.99}}^1 5u^4 \, du = u^5\Big|_{1-\phi_{0.99}}^1 = 1 - (1-\phi_{0.99})^5.$$

Solving for $\phi_{0.99}$ gives

$$\phi_{0.99} = 1 - (0.01)^{1/5} \approx 0.60189.$$

Therefore, the capacity would have to be set at 60,189 gallons. □

4.9 Tchebysheff's Inequality

Remark: When we calculate probabilities associated with a random variable Y, we usually do so under the assumption that Y follows a certain distribution, say $Y \sim \text{Poisson}(\lambda = 1.5)$ or $Y \sim \mathcal{N}(\mu = 125, \sigma^2 = 15^2)$. However, in some situations, we may not know what the distribution of Y is, or we may be unwilling to elicit Y's distribution for fear of making a bad choice. In these instances, we cannot calculate probabilities exactly, but we may be able to place bounds on how large or small these probabilities are.

Markov's Inequality: Suppose Y is a random variable with

- $P(Y \geq 0) = 1$; i.e., Y has positive support
- $P(Y = 0) < 1$; i.e., Y is not degenerate at $y = 0$.

For any $r > 0$,

$$P(Y \geq r) \leq \frac{E(Y)}{r}.$$

Proof. Suppose Y is continuous with pdf $f_Y(y)$. The expected value of Y is

$$\begin{aligned} E(Y) = \int_0^\infty y f_Y(y)dy &\geq \int_r^\infty y f_Y(y)dy \\ &\geq \int_r^\infty r f_Y(y)dy = r\int_r^\infty f_Y(y)dy = rP(Y \geq r). \end{aligned}$$

If Y is discrete with pmf $p_Y(y)$, the proof is analogous; simply replace $f_Y(y)$ with $p_Y(y)$ and replace integrals with sums. □

Note: The probability $P(Y \geq r)$ is a **right-tail probability**. Often r is a value out in the right tail of the pmf/pdf of Y. Markov's Inequality places an upper bound on how large this probability can be, and this upper bound holds regardless of what the distribution of Y is (as long as the two conditions above are satisfied).

Illustration: In Example 4.13 (pp 102-103, notes), we assumed the time to death $Y \sim \text{gamma}(\alpha = 2.7, \beta = 100)$ and calculated

$$P(Y > 365) = \int_{365}^\infty \frac{1}{\Gamma(2.7)100^{2.7}}\, y^{2.7-1}e^{-y/100}dy \approx 0.236.$$

If the gamma$(\alpha = 2.7, \beta = 100)$ distribution is the correct model for Y, then this is the correct answer for $P(Y > 365)$. On the other hand, the Markov upper bound on this probability, which assumes only that $E(Y) = 270$, is given by

$$P(Y > 365) \leq \frac{270}{365} \approx 0.740.$$

This example illustrates how conservative the Markov upper bound can be.

Tchebysheff's Inequality: Suppose Y is a random variable with mean $E(Y) = \mu$ and variance $V(Y) = \sigma^2 > 0$. For any $k > 0$,

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

Proof. Rewrite the event

$$\{|Y - \mu| \geq k\sigma\} = \{(Y - \mu)^2 \geq k^2\sigma^2\}.$$

This is justified because $|Y - \mu|$, k, and σ are all nonnegative. Therefore, we can write

$$P(|Y - \mu| \geq k\sigma) = P((Y - \mu)^2 \geq k^2\sigma^2).$$

Note that $(Y - \mu)^2$ is a nonnegative random variable and is not degenerate at 0 (because $\sigma^2 > 0$). Therefore, we can apply Markov's Inequality to the RHS with $r = k^2\sigma^2$ to get

$$P((Y - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(Y - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}. \ \square$$

Note: Tchebysheff's Inequality can be written equivalently as

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

This is true because the event $\{|Y - \mu| < k\sigma\}$ is the complement of $\{|Y - \mu| \geq k\sigma\}$. Writing Tchebysheff's Inequality in this way is helpful. Note that

$$|Y - \mu| < k\sigma \iff -k\sigma < Y - \mu < k\sigma \iff \mu - k\sigma < Y < \mu + k\sigma.$$

Therefore,

$$P(\mu - k\sigma < Y < \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

A statistical interpretation of $P(\mu - k\sigma < Y < \mu + k\sigma)$ is "the proportion of individuals in the population within k standard deviations of the mean μ."

Illustration: Let's calculate $P(\mu - k\sigma < Y < \mu + k\sigma)$ when $Y \sim \mathcal{N}(\mu, \sigma^2)$ and compare it to the lower bound conferred by Tchebysheff's Inequality.

k	Probability	Normal	Tchebysheff's lower bound
1	$P(\mu - \sigma < Y < \mu + \sigma)$	0.6827	0
2	$P(\mu - 2\sigma < Y < \mu + 2\sigma)$	0.9545	$3/4 = 0.75$
3	$P(\mu - 3\sigma < Y < \mu + 3\sigma)$	0.9973	$8/9 \approx 0.89$
4	$P(\mu - 4\sigma < Y < \mu + 4\sigma)$	0.9999	$15/16 \approx 0.94$

This illustrates how conservative the Tchebysheff bound can be. The lower bound conferred by Tchebysheff's Inequality must hold for every possible distribution (with the same mean and variance), so it is not surprising that lower (or upper) bounds are conservative.

5 Multivariate Probability Distributions

5.1 Introduction

Remark: In Chapters 3 and 4, we were interested in **univariate** random variables (of the discrete and continuous type, respectively). However, in many problems, there are two or more random variables of interest and the goal is to understand the probabilistic behavior of them together. For example,

- Researchers would like to use a student's pretest score Y_1 and his/her posttest score Y_2 to assess the effectiveness of an educational program.
- For a pool of high-risk drivers, actuaries want to describe the amount of financial loss due to collisions Y_1 and liability Y_2.
- In a clinical trial, physicians want to characterize the concentration of a drug in a patient's body Y as a function of the patient's body weight X.
- An electrical system consists of four components whose times to failure are denoted by Y_1, Y_2, Y_3, and Y_4. Engineers would like to describe the reliability of the system.

In each example, it is natural to posit a relationship between or among the random variables involved. This relationship can be described mathematically using a **joint probability distribution**. This distribution, in turn, allows us to make probability statements involving the random variables—just as univariate distributions allow us to do this with a single random variable.

5.2 Joint distributions for two random variables

5.2.1 The discrete case

Terminology: Suppose Y_1 and Y_2 are discrete random variables. We call $\mathbf{Y} = (Y_1, Y_2)$ a **discrete random vector**. The **joint probability mass function (pmf)** of Y_1 and Y_2 is

$$p_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2),$$

which is nonzero for all $(y_1, y_2) \in R$. The set $R \subset \mathbb{R}^2$ is the two-dimensional support of $\mathbf{Y} = (Y_1, Y_2)$. The joint pmf $p_{Y_1,Y_2}(y_1, y_2)$ has the following properties:

1. $0 \leq p_{Y_1,Y_2}(y_1, y_2) \leq 1$, for all y_1 and y_2
2. the sum of the probabilities over all y_1 and y_2 equals 1; i.e.,

$$\sum_{(y_1,y_2)\in R} \sum p_{Y_1,Y_2}(y_1, y_2) = 1.$$

Compare these properties with those of a valid pmf $p_Y(y)$ in the univariate discrete case; see pp 38 (notes).

Example 5.1. An actuary is interested in the number of tornadoes recorded in two Iowa counties on a per-year basis. Define

$$\begin{aligned} Y_1 &= \text{the number of tornados recorded each year in Lee County} \\ Y_2 &= \text{the number of tornados recorded each year in Van Buren County.} \end{aligned}$$

The joint pmf of Y_1 and Y_2 is given in the table below:

$p_{Y_1,Y_2}(y_1,y_2)$	$y_2=0$	$y_2=1$	$y_2=2$
$y_1=0$	0.64	0.08	0.04
$y_1=1$	0.12	0.06	0.02
$y_1=2$	0.02	0.01	0.01

In this example, the support of $\mathbf{Y} = (Y_1, Y_2)$ is

$$R = \{(0,0),(1,0),(2,0),(0,1),(1,1),(2,1),(0,2),(1,2),(2,2)\}.$$

The probabilities $p_{Y_1,Y_2}(y_1,y_2) = P(Y_1 = y_1, Y_2 = y_2)$ associated with each support point are the entries in the table.

(a) What is the probability there is no more than one tornado recorded in the two counties combined in a given year?
(b) What is the probability there are two tornadoes recorded in Lee County in a given year?

Solutions. In part (a), we want $P(Y_1 + Y_2 \leq 1)$. The support points which correspond to the event $\{Y_1 + Y_2 \leq 1\}$ are $(0,0), (1,0)$ and $(0,1)$. Thus,

$$\begin{aligned} P(Y_1 + Y_2 \leq 1) &= p_{Y_1,Y_2}(0,0) + p_{Y_1,Y_2}(1,0) + p_{Y_1,Y_2}(0,1) \\ &= 0.64 + 0.12 + 0.08 = 0.84. \end{aligned}$$

In part (b), we want $P(Y_1 = 2)$. The support points which correspond to the event $\{Y_1 = 2\}$ are $(2,0)$, $(2,1)$, and $(2,2)$. Thus,

$$\begin{aligned} P(Y_1 = 2) &= p_{Y_1,Y_2}(2,0) + p_{Y_1,Y_2}(2,1) + p_{Y_1,Y_2}(2,2) \\ &= 0.02 + 0.01 + 0.01 = 0.04. \ \square \end{aligned}$$

This example illustrates the following general result.

Result: Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a discrete random vector with joint pmf $p_{Y_1,Y_2}(y_1,y_2)$. The probability of the event $\{(Y_1,Y_2) \in B\}$ is found by adding the probabilities $p_{Y_1,Y_2}(y_1,y_2)$ for all $(y_1,y_2) \in B$; i.e.,

$$P((Y_1,Y_2) \in B) = \sum_{(y_1,y_2)\in B} \sum p_{Y_1,Y_2}(y_1,y_2).$$

This is analogous to how probabilities were calculated in discrete distributions for univariate random variables; see pp 39 (notes).

5.2.2 The continuous case

Terminology: Suppose Y_1 and Y_2 are continuous random variables. We call $\mathbf{Y} = (Y_1, Y_2)$ a **continuous random vector**. The **joint probability density function (pdf)** of Y_1 and Y_2 is denoted by

$$f_{Y_1,Y_2}(y_1, y_2).$$

The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which is strictly larger than zero over $R \subset \mathbb{R}^2$, the two-dimensional support of $\mathbf{Y} = (Y_1, Y_2)$. The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ has the following properties:

1. $f_{Y_1,Y_2}(y_1, y_2) \geq 0$, for all $(y_1, y_2) \in \mathbb{R}^2$
2. The function $f_{Y_1,Y_2}(y_1, y_2)$ integrates to one; i.e.,

$$\int\int_{\mathbb{R}^2} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 = 1.$$

Compare these properties with those of a valid pdf $f_Y(y)$ in the univariate continuous case; see pp 78 (notes).

Result: Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. The probability of the event $\{(Y_1, Y_2) \in B\}$ is found by integrating $f_{Y_1,Y_2}(y_1, y_2)$ over the set B; i.e.,

$$P((Y_1, Y_2) \in B) = \int\int_{(y_1,y_2)\in B} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2.$$

In other words, $P((Y_1, Y_2) \in B)$ is the volume under the joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ over the two-dimensional set B.

Example 5.2. Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} cy_1y_2, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the value of c that makes $f_{Y_1,Y_2}(y_1, y_2)$ a valid pdf.
(b) Calculate $P(Y_1 - Y_2 > \frac{1}{8})$.

Solutions. Whenever we have a problem like this, the first thing we do is make a detailed picture of what the two-dimensional support looks like; here,

$$R = \{(y_1, y_2) : 0 < y_2 < y_1 < 1\}.$$

This triangular region is shown in Figure 5.1 (see next page). The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which takes the value cy_1y_2 over this region (and equals zero, otherwise).

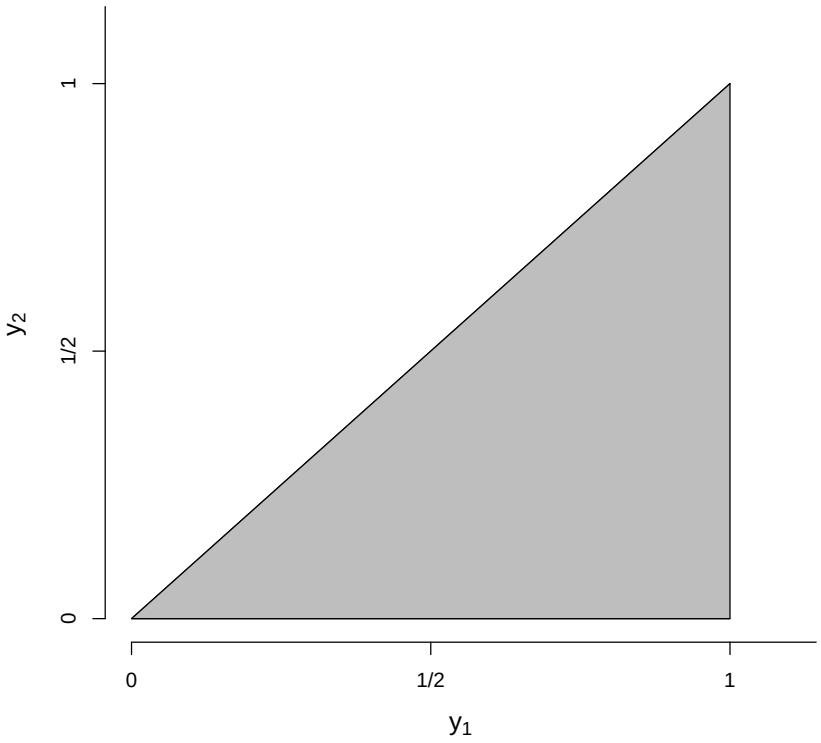

Figure 5.1: The support $R = \{(y_1, y_2) : 0 < y_2 < y_1 < 1\}$ in Example 5.2.

(a) To find the value of c, we use the fact that

$$\int\int_{\mathbb{R}^2} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 = 1 \implies \int_{y_2=0}^{1}\int_{y_1=y_2}^{1} cy_1y_2 \; dy_1dy_2 \stackrel{\text{set}}{=} 1.$$

We could also set up the double integral by integrating in the reverse order; i.e.,

$$\int_{y_1=0}^{1}\int_{y_2=0}^{y_1} cy_1y_2 \; dy_2dy_1 \stackrel{\text{set}}{=} 1.$$

We can solve either integral equation for c. Let's solve the second integral:

$$\int_{y_1=0}^{1}\int_{y_2=0}^{y_1} cy_1y_2 \, dy_2dy_1 = c\int_{y_1=0}^{1} y_1\left(\frac{y_2^2}{2}\bigg|_{y_2=0}^{y_1}\right)dy_1 = c\int_{y_1=0}^{1}\frac{y_1^3}{2}\,dy_1 = \frac{c}{2}\left(\frac{y_1^4}{4}\bigg|_{y_1=0}^{1}\right) = \frac{c}{8} \stackrel{\text{set}}{=} 1.$$

Therefore, $c = 8$ and the joint pdf is

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 8y_1y_2, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

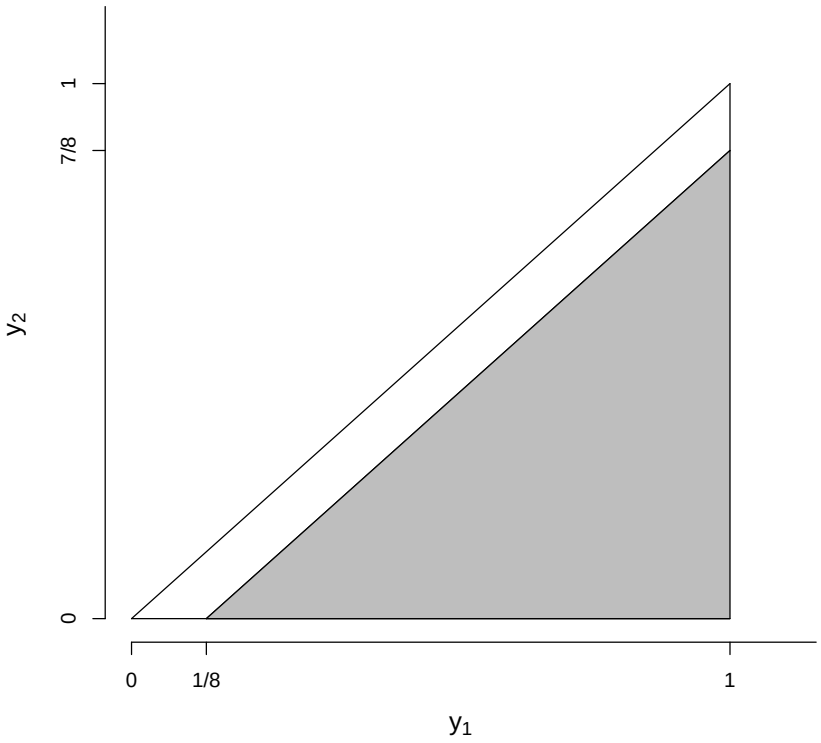

Figure 5.2: The set $B = \{(y_1, y_2) : 0 < y_2 < y_1 < 1,\ y_1 - y_2 > \frac{1}{8}\}$ in Example 5.2.

(b) To find $P(Y_1 - Y_2 > \frac{1}{8})$, we integrate $f_{Y_1,Y_2}(y_1, y_2)$ over the set

$$B = \left\{(y_1, y_2) : 0 < y_2 < y_1 < 1,\ y_1 - y_2 > \frac{1}{8}\right\},$$

shown in Figure 5.2 (see above). The boundary of B is determined as follows:

$$y_1 - y_2 = \frac{1}{8} \implies y_2 = y_1 - \frac{1}{8}.$$

Therefore,

$$P\left(Y_1 - Y_2 > \frac{1}{8}\right) = \int_{y_2=0}^{\frac{7}{8}} \int_{y_1=y_2+\frac{1}{8}}^{1} 8y_1y_2\ dy_1dy_2 \approx 0.698.$$

I calculated this double integral in R using the `integral2` function in the `pracma` package:

```
> library(pracma)
> joint.pdf <- function(y1,y2) 8*y1*y2
> y1min <- function(y2) y2+1/8
> integral2(joint.pdf,0,7/8,y1min,1)
```

```
$Q
[1] 0.6978353
$error
[1] 2.775558e-17
```

We could also calculate this by integrating in the reverse order (see Figure 5.2); i.e.,

$$P\left(Y_1 - Y_2 > \frac{1}{8}\right) = \int_{y_1=\frac{1}{8}}^{1} \int_{y_2=0}^{y_1-\frac{1}{8}} 8y_1y_2 \; dy_2dy_1 \approx 0.698.$$

```
> library(pracma)
> joint.pdf <- function(y1,y2) 8*y1*y2
> y2max <- function(y1) y1-1/8
> integral2(joint.pdf,1/8,1,0,y2max)
$Q
[1] 0.6978353
$error
[1] 1.752071e-16
```

Remark: When working with joint distributions for continuous random variables, constructing good pictures of the support and regions of integration is very helpful. Double integral limits are determined from good pictures. Students who do not take the time to construct good pictures usually get the wrong answer. □

5.3 Marginal distributions

5.3.1 The discrete case

Recall: In Example 5.1 (notes), we examined the joint distribution of

Y_1 = the number of tornados recorded each year in Lee County
Y_2 = the number of tornados recorded each year in Van Buren County.

The joint pmf of Y_1 and Y_2 was described in the following table:

$p_{Y_1,Y_2}(y_1,y_2)$	$y_2=0$	$y_2=1$	$y_2=2$
$y_1=0$	0.64	0.08	0.04
$y_1=1$	0.12	0.06	0.02
$y_1=2$	0.02	0.01	0.01

A joint pmf describes how two random variables are distributed jointly. We now discuss marginal distributions, which describe how random variables are distributed separately.

Terminology: Suppose Y_1 and Y_2 are discrete random variables with joint pmf $p_{Y_1,Y_2}(y_1, y_2)$. The **marginal pmf** of Y_1 is

$$p_{Y_1}(y_1) = \sum_{\text{all } y_2} p_{Y_1,Y_2}(y_1, y_2).$$

Similarly, the **marginal pmf** of Y_2 is

$$p_{Y_2}(y_2) = \sum_{\text{all } y_1} p_{Y_1,Y_2}(y_1, y_2).$$

In other words, to find the marginal pmf of one random variable, you take the joint pmf and sum over the possible values of the other random variable.

The table below shows how the marginal distributions are calculated in Example 5.1:

$p_{Y_1,Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$p_{Y_1}(y_1)$
$y_1 = 0$	0.64	0.08	0.04	0.76
$y_1 = 1$	0.12	0.06	0.02	0.20
$y_1 = 2$	0.02	0.01	0.01	0.04
$p_{Y_2}(y_2)$	0.78	0.15	0.07	

That is, the marginal pmf of Y_1 is

y_1	0	1	2
$p_{Y_1}(y_1)$	0.76	0.20	0.04

and the marginal pmf of Y_2 is

y_2	0	1	2
$p_{Y_2}(y_2)$	0.78	0.15	0.07

Note: Marginal pmfs are univariate pmfs—just like those we saw in Chapter 3. The marginal pmf of Y_1 describes how Y_1 varies on its own and similarly for the marginal pmf of Y_2.

5.3.2 The continuous case

Terminology: Suppose Y_1 and Y_2 are continuous random variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. The **marginal pdf** of Y_1 is

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2.$$

Similarly, the **marginal pdf** of Y_2 is

$$f_{Y_2}(y_2) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)\, dy_1.$$

In other words, to find the marginal pdf of one random variable, you take the joint pdf and integrate it over the other variable.

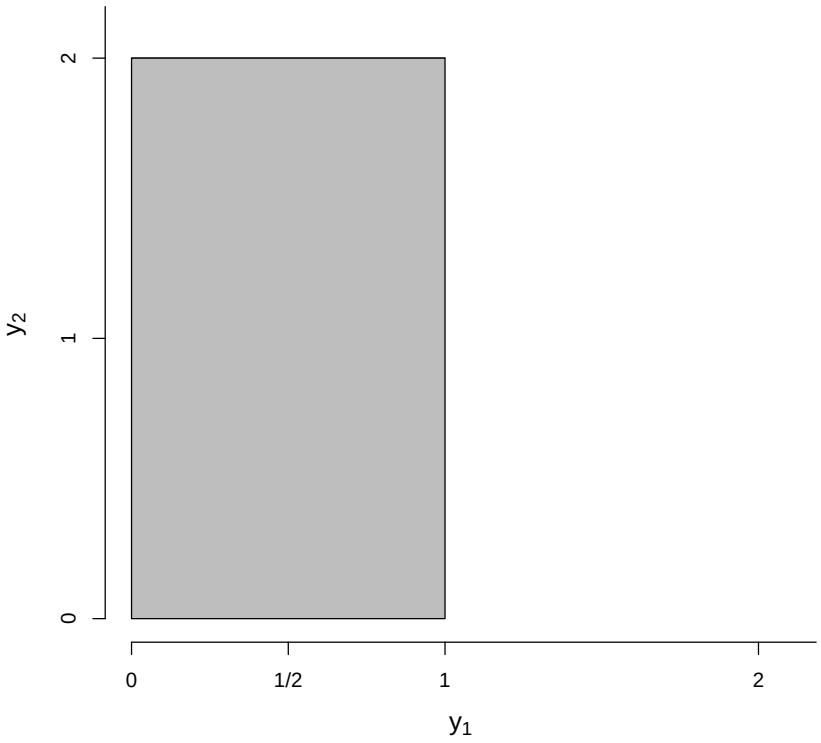

Figure 5.3: The support $R = \{(y_1, y_2) : 0 \leq y_1 \leq 1,\ 0 \leq y_2 \leq 2\}$ in Example 5.3.

Example 5.3. An insurance company insures a large number of drivers. Let Y_1 denote the company's losses under collision insurance and let Y_2 denote the company's losses under liability insurance. The joint pdf of Y_1 and Y_2 is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \dfrac{2y_1 + 2 - y_2}{4}, & 0 \leq y_1 \leq 1,\ 0 \leq y_2 \leq 2 \\ 0, & \text{otherwise.} \end{cases}$$

Note that the support $R = \{(y_1, y_2) : 0 \leq y_1 \leq 1,\ 0 \leq y_2 \leq 2\}$ is a rectangular set, shown in Figure 5.3 (above). The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which takes the value $(2y_1 + 2 - y_2)/4$ over this region (and equals zero, otherwise).

Let's find both marginal pdfs. The marginal pdf of Y_1 is nonzero when $0 \leq y_1 \leq 1$. For these values, the pdf is found as follows:

$$\begin{aligned} f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 &= \int_{y_2=0}^{2} \left(\frac{2y_1 + 2 - y_2}{4}\right) dy_2 \\ &= \frac{1}{4}\left[\left(2y_1y_2 + 2y_2 - \frac{y_2^2}{2}\right)\Big|_{y_2=0}^{2}\right] = \frac{1}{4}\left(4y_1 + 4 - 2\right) = y_1 + \frac{1}{2}. \end{aligned}$$

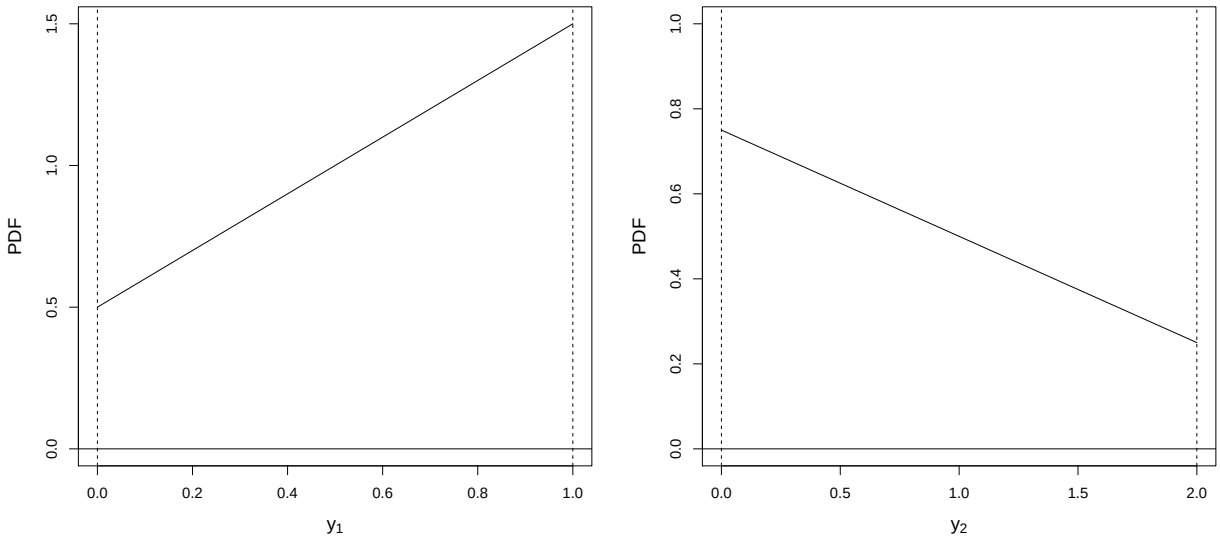

Figure 5.4: Marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ in Example 5.3. Note that both the horizontal and vertical axis scales are different in the two figures.

Therefore, the marginal pdf of Y_1 is given by

$$f_{Y_1}(y_1) = \begin{cases} y_1 + \frac{1}{2}, & 0 \leq y_1 \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

The marginal pdf of Y_2 is nonzero when $0 \leq y_2 \leq 2$. For these values, the pdf is found as follows:

$$\begin{aligned} f_{Y_2}(y_2) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)\, dy_1 &= \int_{y_1=0}^{1} \left(\frac{2y_1 + 2 - y_2}{4} \right) dy_1 \\ &= \frac{1}{4}\left[\left(y_1^2 + 2y_1 - y_1 y_2 \right) \Big|_{y_1=0}^{1} \right] = \frac{1}{4}\left(1 + 2 - y_2\right) = \frac{1}{4}(3 - y_2). \end{aligned}$$

Therefore, the marginal pdf of Y_2 is given by

$$f_{Y_2}(y_2) = \begin{cases} \frac{1}{4}(3 - y_2), & 0 \leq y_2 \leq 2 \\ 0, & \text{otherwise.} \end{cases}$$

Note: It is easy to verify that both $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ are valid (univariate) pdfs; i.e., both functions are nonnegative and integrate to one over their respective supports. Both marginal pdfs are shown in Figure 5.4 (see above).

Q: How would we calculate $P(Y_1 < 0.5)$, the probability the collision loss Y_1 is less than 0.5?

A: We could actually do this in two ways. Probably the easiest way is to just use the marginal pdf $f_{Y_1}(y_1)$. From Chapter 4, we know

$$P(Y_1 < 0.5) = \int_0^{0.5} f_{Y_1}(y_1)dy_1 = \int_0^{0.5} \left(y_1 + \frac{1}{2}\right) dy_1 = \left(\frac{y_1^2}{2} + \frac{y_1}{2}\right)\Big|_0^{0.5} = 0.375.$$

Note that we have calculated the **area** under $f_{Y_1}(y_1)$ over the set $\{y_1 : 0 < y_1 < 0.5\}$; see Figure 5.4 (left).

We could also find $P(Y_1 < 0.5)$ by using the joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. To see how, note that

$$\{Y_1 < 0.5\} = \{0 < Y_1 < 0.5,\ 0 < Y_2 < 2\}.$$

Therefore, we could calculate the **volume** under $f_{Y_1,Y_2}(y_1, y_2)$ over the two-dimensional set

$$B = \{(y_1, y_2) : 0 < y_1 < 0.5,\ 0 < y_2 < 2\};$$

see Figure 5.3. As a double integral, this equals

$$P(Y_1 < 0.5) = \int_{y_1=0}^{0.5}\int_{y_2=0}^{2} f_{Y_1,Y_2}(y_1, y_2)dy_2dy_1 = \int_{y_1=0}^{0.5} \underbrace{\int_{y_2=0}^{2} \left(\frac{2y_1 + 2 - y_2}{4}\right) dy_2}_{= f_{Y_1}(y_1)} dy_1 = 0.375. \square$$

```
> library(pracma)
> joint.pdf <- function(y1,y2) (2*y1+2-y2)/4
> integral2(joint.pdf,0,0.5,0,2)
$Q
[1] 0.375
$error
[1] 0
```

Example 5.4. An engineering system consists of two components whose lifetimes are denoted by Y_1 and Y_2 and whose joint pdf is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{1}{32}y_2^2 e^{-(y_1+y_2)/2}, & y_1 > 0,\ y_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note that the support $R = \{(y_1, y_2) : y_1 > 0,\ y_2 > 0\}$ is the entire first quadrant; see Figure 5.5 (next page). The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which takes the value $(1/32)y_2^2 e^{-(y_1+y_2)/2}$ over this region (and equals zero, otherwise).

Let's find both marginal pdfs. The marginal pdf of Y_1 is nonzero when $y_1 > 0$. For these values, the pdf is found as follows:

$$f_{Y_1}(y_1) = \int_{y_2=0}^{\infty} \frac{1}{32}y_2^2 e^{-(y_1+y_2)/2}dy_2 = \frac{1}{32}e^{-y_1/2} \underbrace{\int_{y_2=0}^{\infty} y_2^2 e^{-y_2/2}dy_2}_{= \Gamma(3)2^3=16} = \frac{1}{2}e^{-y_1/2}.$$

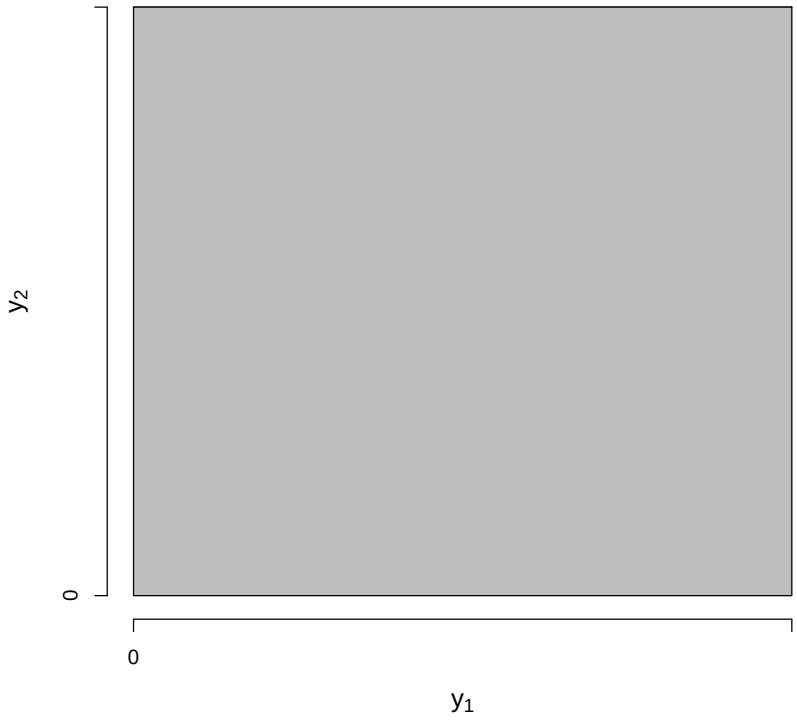

Figure 5.5: The support $R = \{(y_1, y_2) : y_1 > 0,\ y_2 > 0\}$ in Example 5.4; i.e., the entire first quadrant.

Note that in the last integral, the integrand

$$y_2^2 e^{-y_2/2} = y_2^{3-1} e^{-y_2/2}$$

is the kernel of the gamma$(3, 2)$ distribution, and the integral is over $(0, \infty)$. We know immediately that

$$\int_{y_2=0}^{\infty} y_2^2 e^{-y_2/2} dy_2 = \Gamma(3)2^3 = 16.$$

Therefore, the marginal pdf of Y_1 is given by

$$f_{Y_1}(y_1) = \begin{cases} \frac{1}{2}e^{-y_1/2}, & y_1 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $Y_1 \sim$ exponential(2). The marginal pdf of Y_2 is nonzero when $y_2 > 0$. For these values, the pdf is found as follows:

$$f_{Y_2}(y_2) = \int_{y_1=0}^{\infty} \frac{1}{32} y_2^2 e^{-(y_1+y_2)/2} dy_1 = \frac{1}{16} y_2^2 e^{-y_2/2} \underbrace{\int_{y_1=0}^{\infty} \frac{1}{2} e^{-y_1/2} dy_1}_{=1} = \frac{1}{16} y_2^2 e^{-y_2/2}.$$

Note that the last integral equals 1 because the integrand is the exponential(2) pdf and this pdf is being integrated over $(0, \infty)$. Therefore, the marginal pdf of Y_2 is given by

$$f_{Y_2}(y_2) = \begin{cases} \frac{1}{16} y_2^2 e^{-y_2/2}, & y_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Because

$$\frac{1}{16} y_2^2 e^{-y_2/2} = \frac{1}{\Gamma(3) 2^3} y_2^{3-1} e^{-y_2/2}$$

and the support of Y_2 is $\{y_2 : y_2 > 0\}$, we see immediately that $Y_2 \sim \text{gamma}(3, 2)$. $\square$

Interesting: In this example, note that the joint pdf can be written as the **product** of the two marginal pdfs; i.e.,

$$\frac{1}{32} y_2^2 e^{-(y_1+y_2)/2} \;=\; \frac{1}{2} e^{-y_1/2} \;\times\; \frac{1}{16} y_2^2 e^{-y_2/2};$$

i.e., $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$, for all $y_1 \in \mathbb{R}$ and $y_2 \in \mathbb{R}$. (Continuous) random variables Y_1 and Y_2 that have this property are said to be **independent**. More on this soon.

5.4 Conditional distributions

5.4.1 The discrete case

Terminology: Suppose Y_1 and Y_2 are discrete random variables with joint pmf $p_{Y_1,Y_2}(y_1, y_2)$ and marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$, respectively. The **conditional probability mass function (pmf)** of Y_1, given $Y_2 = y_2$, is given by

$$p_{Y_1|Y_2}(y_1|y_2) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)},$$

whenever $p_{Y_2}(y_2) > 0$. Similarly, the conditional probability mass function (pmf) of Y_2, given $Y_1 = y_1$, is

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)},$$

whenever $p_{Y_1}(y_1) > 0$.

Interpretation: The conditional pmf $p_{Y_1|Y_2}(y_1|y_2) = P(Y_1 = y_1|Y_2 = y_2)$ is a univariate pmf. It describes the distribution of Y_1 (i.e., how Y_1 varies) when Y_2 is fixed at the value y_2. Similarly, $p_{Y_2|Y_1}(y_2|y_1) = P(Y_2 = y_2|Y_1 = y_1)$ describes the distribution of Y_2 (i.e., how Y_2 varies) when Y_1 is fixed at the value y_1.

Recall: In Example 5.1 (notes), we examined the joint distribution of

Y_1 = the number of tornados recorded each year in Lee County
Y_2 = the number of tornados recorded each year in Van Buren County.

The joint pmf of Y_1 and Y_2, $p_{Y_1,Y_2}(y_1, y_2)$, was described in the table

$p_{Y_1,Y_2}(y_1,y_2)$	$y_2=0$	$y_2=1$	$y_2=2$	$p_{Y_1}(y_1)$
$y_1=0$	0.64	0.08	0.04	0.76
$y_1=1$	0.12	0.06	0.02	0.20
$y_1=2$	0.02	0.01	0.01	0.04
$p_{Y_2}(y_2)$	0.78	0.15	0.07	

The marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$ are in the margins of this table.

(a) Find $p_{Y_1|Y_2}(y_1|y_2=0)$, the conditional pmf of Y_1 when $Y_2=0$.
(b) Find $P(Y_1 \leq 1|Y_2=0)$.

Solutions. For part (a), the conditional pmf $p_{Y_1|Y_2}(y_1|y_2=0)$ describes the distribution of Y_1 when $Y_2=0$. This is a univariate pmf with three possible values of Y_1, namely, 0, 1, and 2. These conditional probabilities are calculated below:

$$\begin{aligned}
p_{Y_1|Y_2}(y_1=0|y_2=0) &= \frac{p_{Y_1,Y_2}(0,0)}{p_{Y_2}(0)} = \frac{0.64}{0.78} \approx 0.820 \\
p_{Y_1|Y_2}(y_1=1|y_2=0) &= \frac{p_{Y_1,Y_2}(1,0)}{p_{Y_2}(0)} = \frac{0.12}{0.78} \approx 0.154 \\
p_{Y_1|Y_2}(y_1=2|y_2=0) &= \frac{p_{Y_1,Y_2}(2,0)}{p_{Y_2}(0)} = \frac{0.02}{0.78} \approx 0.026.
\end{aligned}$$

We can display the conditional pmf of Y_1 given $Y_2=0$ in the following table:

y_1	0	1	2
$p_{Y_1|Y_2}(y_1|y_2=0)$	0.820	0.154	0.026

(b) Using the conditional pmf $p_{Y_1|Y_2}(y_1|y_2=0)$ in part (a), we have

$$P(Y_1 \leq 1|Y_2=0) = P(Y_1=0|Y_2=0) + P(Y_1=1|Y_2=0) = 0.820 + 0.154 = 0.974.$$

Note that this probability is different than had we calculated $P(Y_1 \leq 1)$ marginally; i.e.,

$$P(Y_1 \leq 1) = P(Y_1=0) + P(Y_1=1) = 0.76 + 0.20 = 0.96. \ \square$$

5.4.2 The continuous case

Terminology: Suppose Y_1 and Y_2 are continuous random variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ and marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. The **conditional probability density function (pdf)** of Y_1, given $Y_2 = y_2$, is given by

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)},$$

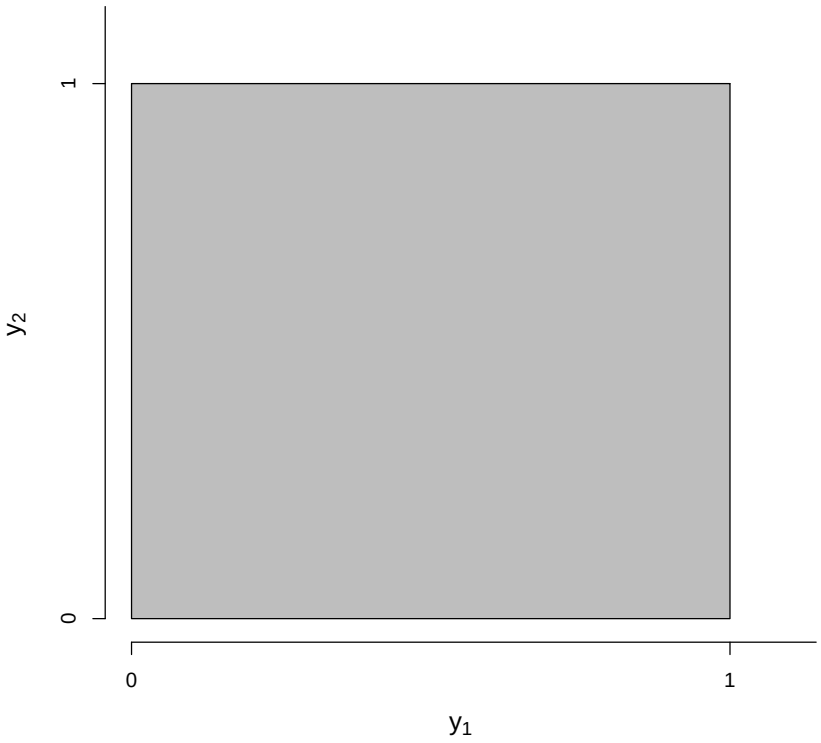

Figure 5.6: The support $R = \{(y_1, y_2) : 0 < y_1 < 1,\ 0 < y_2 < 1\}$ in Example 5.5.

whenever $f_{Y_2}(y_2) > 0$. Similarly, the conditional probability density function (pdf) of Y_2, given $Y_1 = y_1$, is

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)},$$

whenever $f_{Y_1}(y_1) > 0$.

Interpretation: The conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$ is a univariate pdf. It describes the distribution of Y_1 (i.e., how Y_1 varies) when Y_2 is fixed at the value y_2. Similarly, $f_{Y_2|Y_1}(y_2|y_1)$ describes the distribution of Y_2 (i.e., how Y_2 varies) when Y_1 is fixed at the value y_1.

Example 5.5. The demand for two products, Y_1 and Y_2, is modeled using the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{24}{11}\left(y_1^2 + \frac{y_1 y_2}{2}\right), & 0 < y_1 < 1,\ 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note that the support $R = \{(y_1, y_2) : 0 < y_1 < 1,\ 0 < y_2 < 1\}$ is the unit square, shown in Figure 5.6 (above). The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which takes the value $\frac{24}{11}(y_1^2 + y_1 y_2/2)$ over this region (and equals zero, otherwise).

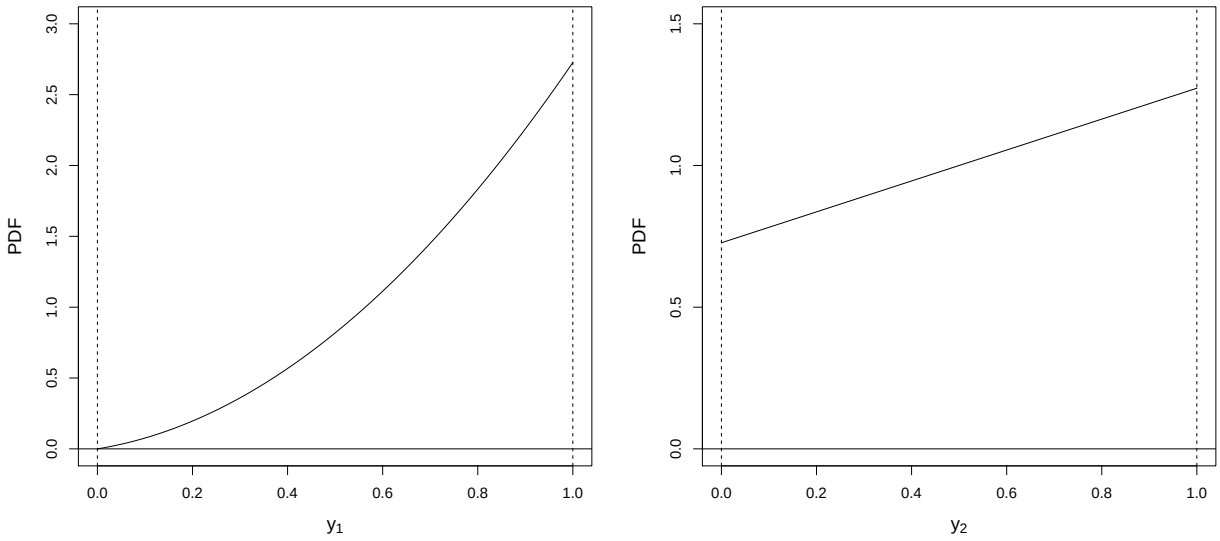

Figure 5.7: Marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ in Example 5.5.

(a) Find both marginal distributions.
(b) Find both conditional distributions.

Solutions. (a) Recall that to find the marginal distribution of one variable, we integrate the joint pdf over the other. Integration is easy here because the bivariate support is the unit square. The marginal pdf of Y_1 is nonzero when $0 < y_1 < 1$. For these values,

$$f_{Y_1}(y_1) = \int_{y_2=0}^{1} \frac{24}{11}\left(y_1^2 + \frac{y_1 y_2}{2}\right) dy_2 = \frac{24}{11}\left[\left(y_1^2 y_2 + \frac{y_1 y_2^2}{4}\right)\Big|_{y_2=0}^{1}\right] = \frac{24}{11}\left(y_1^2 + \frac{y_1}{4}\right) = \frac{6}{11}(y_1 + 4y_1^2).$$

The marginal pdf of Y_2 is also nonzero when $0 < y_2 < 1$. For these values,

$$f_{Y_2}(y_2) = \int_{y_1=0}^{1} \frac{24}{11}\left(y_1^2 + \frac{y_1 y_2}{2}\right) dy_1 = \frac{24}{11}\left[\left(\frac{y_1^3}{3} + \frac{y_1^2 y_2}{4}\right)\Big|_{y_1=0}^{1}\right] = \frac{24}{11}\left(\frac{1}{3} + \frac{y_2}{4}\right) = \frac{2}{11}(4 + 3y_2).$$

Summarizing, we have

$$f_{Y_1}(y_1) = \begin{cases} \frac{6}{11}(y_1 + 4y_1^2), & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} \frac{2}{11}(4 + 3y_2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ are shown side by side in Figure 5.7 (above). It is easy to show both pdfs are valid.

(b) The conditional pdf of Y_1 is nonzero when $0 < y_1 < 1$. For these values,

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)} = \frac{\frac{24}{11}\left(y_1^2 + \frac{y_1y_2}{2}\right)}{\frac{2}{11}(4+3y_2)} = \underbrace{\frac{12}{4+3y_2}\left[y_1^2 + \left(\frac{y_2}{2}\right)y_1\right]}_{\text{quadratic function of } y_1;\ y_2 \text{ fixed}}.$$

The conditional pdf of Y_2 is nonzero when $0 < y_2 < 1$. For these values,

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_1}(y_1)} = \frac{\frac{24}{11}\left(y_1^2 + \frac{y_1y_2}{2}\right)}{\frac{6}{11}(y_1+4y_1^2)} = \underbrace{\frac{4y_1^2}{y_1+4y_1^2} + \left(\frac{2y_1}{y_1+4y_1^2}\right)y_2}_{\text{linear function of } y_2;\ y_1 \text{ fixed}}.$$

Summarizing,

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \dfrac{12}{4+3y_2}\left[y_1^2 + \left(\dfrac{y_2}{2}\right)y_1\right], & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} \dfrac{4y_1^2}{y_1+4y_1^2} + \left(\dfrac{2y_1}{y_1+4y_1^2}\right)y_2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Important: Both $f_{Y_1|Y_2}(y_1|y_2)$ and $f_{Y_2|Y_1}(y_2|y_1)$ are univariate pdfs! The conditioning variable is regarded as fixed.

Q: How would we calculate $P(Y_1 > \frac{1}{2}|Y_2 = \frac{1}{3})$?
A: This is a conditional probability, so we calculate it using the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$. Specifically, when $y_2 = \frac{1}{3}$, the pdf $f_{Y_1|Y_2}(y_1|y_2)$ above reduces to

$$f_{Y_1|Y_2}(y_1|y_2 = 1/3) = \begin{cases} \dfrac{12}{5}y_1^2 + \dfrac{2}{5}y_1, & 0 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} P\left(Y_1 > \frac{1}{2}\middle|Y_2 = \frac{1}{3}\right) &= \int_{y_1=\frac{1}{2}}^{1} f_{Y_1|Y_2}(y_1|y_2 = 1/3)dy_1 \\ &= \int_{y_1=\frac{1}{2}}^{1} \left(\frac{12}{5}y_1^2 + \frac{2}{5}y_1\right)dy_1 \\ &= \left(\frac{12}{15}y_1^3 + \frac{2}{10}y_1^2\right)\Big|_{y_1=\frac{1}{2}}^{1} = \frac{12}{15} + \frac{2}{10} - \frac{12}{120} - \frac{2}{40} = 0.85. \end{aligned}$$

The conditional pdf $f_{Y_1|Y_2}(y_1|y_2 = 1/3)$ is shown in Figure 5.8 (see next page).

Exercise: Calculate $P(Y_2 \leq \frac{1}{2}|Y_1 = \frac{1}{4})$. *Hint:* Use $f_{Y_2|Y_1}(y_2|y_1)$ when $y_1 = \frac{1}{4}$.

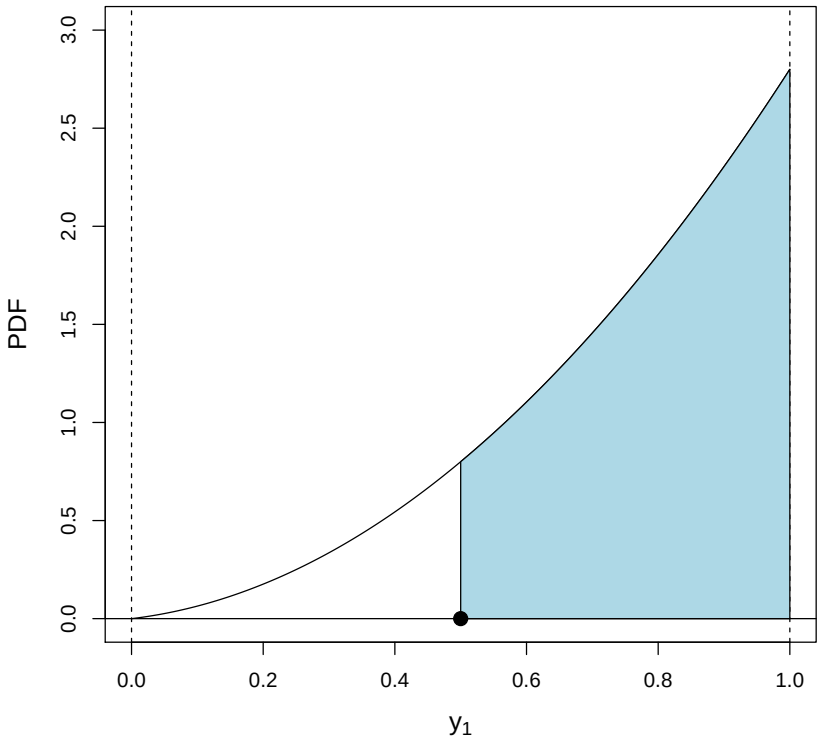

Figure 5.8: The conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$ in Example 5.5 when $y_2 = \frac{1}{3}$. The shaded area corresponds to $P(Y_1 > \frac{1}{2}|Y_2 = \frac{1}{3})$.

Remark: It is interesting to compare the marginal distribution of Y_1, which is described by

$$f_{Y_1}(y_1) = \begin{cases} \frac{6}{11}(y_1 + 4y_1^2), & 0 < y_1 < 1 \\ 0, & \text{otherwise,} \end{cases}$$

to the conditional distribution of Y_1 (when $Y_2 = \frac{1}{3}$), which is described by

$$f_{Y_1|Y_2}(y_1|y_2 = 1/3) = \begin{cases} \frac{12}{5}y_1^2 + \frac{2}{5}y_1, & 0 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Both pdfs describe the distribution of Y_1. However,

- The marginal pdf describes how Y_1 is distributed after removing the influence of Y_2.
- The conditional pdf describes how Y_1 is distributed after incorporating information about the value of Y_2.

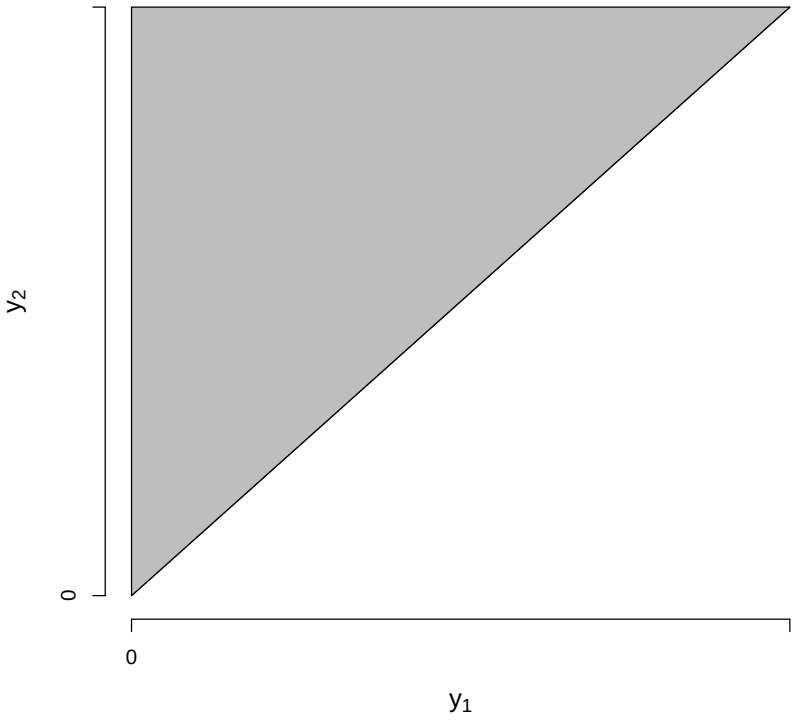

Figure 5.9: The support $R = \{(y_1, y_2) : 0 < y_1 < y_2 < \infty\}$ in Example 5.6.

Note that $f_{Y_1}(y_1) \neq f_{Y_1|Y_2}(y_1|y_2 = 1/3)$; i.e., knowledge of what Y_2 is changes how Y_1 is distributed. Incorporating this knowledge also changes how probabilities are assigned. Recall that we calculated

$$P\left(Y_1 > \frac{1}{2}\middle| Y_2 = \frac{1}{3}\right) = 0.85.$$

Had we ignored Y_2 and calculated $P(Y_1 > \frac{1}{2})$ marginally, we would get

$$P\left(Y_1 > \frac{1}{2}\right) = \int_{y_1=\frac{1}{2}}^{1} f_{Y_1}(y_1)dy_1 = \int_{y_1=\frac{1}{2}}^{1} \frac{6}{11}(y_1 + 4y_1^2)dy_1 \approx 0.841. \ \square$$

Example 5.6. Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} e^{-y_2}, & 0 < y_1 < y_2 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Note that the support $R = \{(y_1, y_2) : 0 < y_1 < y_2 < \infty\}$ is shown in Figure 5.9 (see above). The joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is a three-dimensional function which takes the value e^{-y_2} over this region (and equals zero, otherwise).

(a) Find both marginal distributions.
(b) Find both conditional distributions.

Solutions. (a) The marginal pdf of Y_1 is nonzero when $y_1 > 0$. For these values,

$$f_{Y_1}(y_1) = \int_{y_2=y_1}^{\infty} e^{-y_2} dy_2 = -e^{-y_2}\Big|_{y_2=y_1}^{\infty} = -\left(\lim_{y_2\to\infty} e^{-y_2} - e^{-y_1}\right) = e^{-y_1}.$$

The marginal pdf of Y_2 is also nonzero when $y_2 > 0$. For these values,

$$f_{Y_2}(y_2) = \int_{y_1=0}^{y_2} e^{-y_2} dy_1 = y_1 e^{-y_2}\Big|_{y_1=0}^{y_2} = y_2 e^{-y_2} - 0 = y_2 e^{-y_2}.$$

Summarizing, we have

$$f_{Y_1}(y_1) = \begin{cases} e^{-y_1}, & y_1 > 0, \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} y_2 e^{-y_2}, & y_2 > 0, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to show that both marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ are valid. In fact, both marginal distributions are well known; i.e.,

$$\begin{aligned} Y_1 &\sim \text{exponential}(1) \\ Y_2 &\sim \text{gamma}(2, 1). \end{aligned}$$

(b) The conditional pdf of Y_1 is nonzero when $0 < y_1 < y_2$, where y_2 is fixed. For these values,

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)} = \frac{e^{-y_2}}{y_2 e^{-y_2}} = \frac{1}{y_2}.$$

The conditional pdf of Y_2 is nonzero when $y_2 > y_1$, where $y_1 > 0$ is fixed. For these values,

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)} = \frac{e^{-y_2}}{e^{-y_1}} = e^{-(y_2-y_1)}.$$

Summarizing, we have

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \dfrac{1}{y_2}, & 0 < y_1 < y_2 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} e^{-(y_2-y_1)}, & y_2 > y_1 \\ 0, & \text{otherwise.} \end{cases}$$

Interesting: Recall that both conditional pdfs are univariate pdfs. In fact, $Y_1|Y_2 = y_2 \sim \mathcal{U}(0, y_2)$; i.e., conditional on $Y_2 = y_2$, the random variable Y_1 is uniformly distributed from 0 to y_2. The conditional pdf of Y_2 is that of a "shifted" exponential(1) pdf, where the fixed value of y_1 denotes the shift. □

5.5 Independence

Remark: Informally, we say two random variables are **independent** if the value of one random variable does not affect how we assign probabilities to events involving the other one. Recall that if two events A and B are independent, then

$$P(A \cap B) = P(A)P(B).$$

In terms of independent random variables Y_1 and Y_2, this translates into statements like

$$\underbrace{P(a < Y_1 < b,\ c < Y_2 < d)}_{\text{calculated using the joint pmf/pdf}} = \underbrace{P(a < Y_1 < b)P(c < Y_2 < d)}_{\text{calculated using marginal pmfs/pdfs}}.$$

In other words, probabilities calculated using joint distributions can be "broken down" into calculations using marginal distributions (which are often easier). We now describe formally what it means for random variables to be independent.

Terminology: Suppose Y_1 and Y_2 are discrete random variables with joint pmf $p_{Y_1,Y_2}(y_1, y_2)$ and marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$, respectively. We say Y_1 and Y_2 are **independent** if

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2}(y_2),$$

for all $(y_1, y_2) \in \mathbb{R}^2$. In other words, the joint pmf factors into the product of the marginal pmfs.

Recall: In Example 5.1 (notes), we examined the joint distribution of

Y_1 = the number of tornados recorded each year in Lee County
Y_2 = the number of tornados recorded each year in Van Buren County.

$p_{Y_1,Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$p_{Y_1}(y_1)$
$y_1 = 0$	0.64	0.08	0.04	0.76
$y_1 = 1$	0.12	0.06	0.02	0.20
$y_1 = 2$	0.02	0.01	0.01	0.04
$p_{Y_2}(y_2)$	0.78	0.15	0.07	

Note: For Y_1 and Y_2 to be independent, we would need

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2}(y_2)$$

to hold for all (y_1, y_2) in the support

$$R = \{(0,0), (1,0), (2,0), (0,1), (1,1), (2,1), (0,2), (1,2), (2,2)\}.$$

However, this condition does not even hold for the first value $(0, 0)$; i.e.,

$$0.64 = p_{Y_1,Y_2}(0,0) \neq p_{Y_1}(0)p_{Y_2}(0) = 0.76(0.78) = 0.5928.$$

Therefore, Y_1 and Y_2 are not independent.

Example 5.7. Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a discrete random vector with joint pmf

$$p_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \dfrac{y_1 y_2^2}{30}, & y_1 = 1, 2, 3,\ y_2 = 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

Here is the joint pmf of Y_1 and Y_2 written out in tabular form; the marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$ are in the margins.

$p_{Y_1,Y_2}(y_1, y_2)$	$y_2 = 1$	$y_2 = 2$	$p_{Y_1}(y_1)$
$y_1 = 1$	1/30	4/30	5/30
$y_1 = 2$	2/30	8/30	10/30
$y_1 = 3$	3/30	12/30	15/30
$p_{Y_2}(y_2)$	6/30	24/30	

It is easy to verify that

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1) p_{Y_2}(y_2)$$

for all (y_1, y_2) in the support

$$R = \{(1,1), (1,2), (2,1), (2,2), (3,1), (3,2)\}.$$

For example,

$$\frac{1}{30} = p_{Y_1,Y_2}(1,1) = p_{Y_1}(1) p_{Y_2}(1) = \frac{5}{30} \times \frac{6}{30}.$$

Therefore, Y_1 and Y_2 are independent. $\square$

Terminology: Suppose Y_1 and Y_2 are continuous random variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ and marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. We say Y_1 and Y_2 are **independent** if

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2),$$

for all $(y_1, y_2) \in \mathbb{R}^2$. In other words, the joint pdf factors into the product of the marginal pdfs.

Example 5.8. This past year was an encouraging year for biodiversity discovery, as scientists identified thousands of new species of life. For one newly discovered species, geneticists model

$$\begin{aligned} Y_1 &= \text{the percentage of the species possessing Trait 1} \\ Y_2 &= \text{the percentage of the species possessing Trait 2} \end{aligned}$$

using the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 12 y_1^3 (1 - y_2)^2, & 0 < y_1 < 1,\ 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Show that Y_1 and Y_2 are independent.

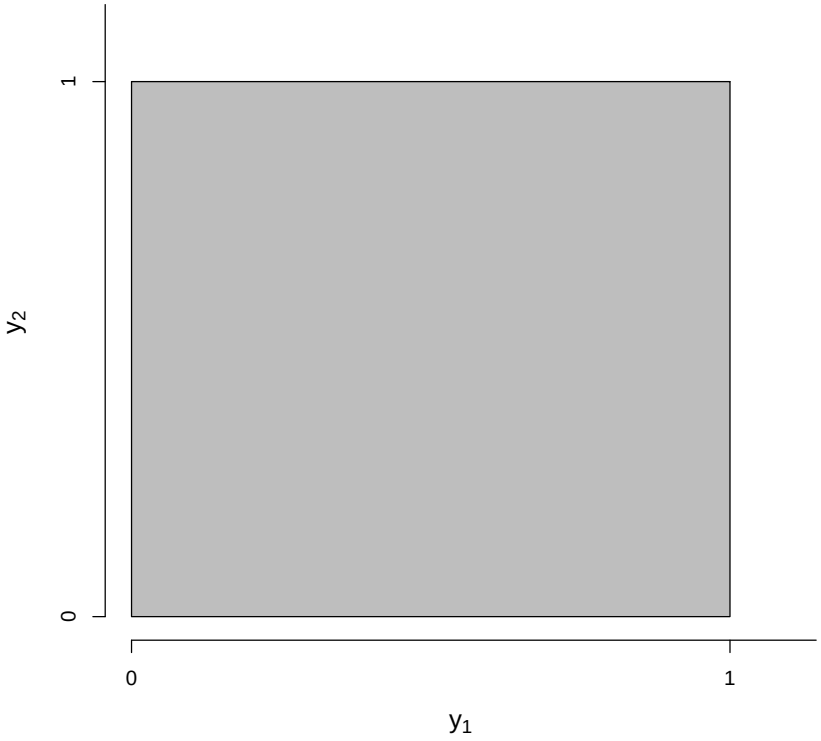

Figure 5.10: The support $R = \{(y_1, y_2) : 0 < y_1 < 1,\ 0 < y_2 < 1\}$ in Example 5.8.

Proof. It suffices to show $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$, where $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ are the marginal pdfs.

The marginal pdf of Y_1 is nonzero when $0 < y_1 < 1$. For these values,

$$\begin{aligned} f_{Y_1}(y_1) = \int_{y_2=0}^{1} 12y_1^3(1-y_2)^2 dy_2 &= 12y_1^3 \int_{y_2=0}^{1} (1-y_2)^2 dy_2 \\ &= -12y_1^3 \int_{u=1}^{0} u^2 du \qquad (u = 1 - y_2) \\ &= -12y_1^3 \left(0 - \frac{1}{3}\right) = 4y_1^3. \end{aligned}$$

The marginal pdf of Y_2 is also nonzero when $0 < y_2 < 1$. For these values,

$$\begin{aligned} f_{Y_2}(y_2) = \int_{y_1=0}^{1} 12y_1^3(1-y_2)^2 dy_1 &= 12(1-y_2)^2 \int_{y_2=0}^{1} y_1^3 dy_1 \\ &= 12(1-y_2)^2 \left(\frac{1}{4} - 0\right) = 3(1-y_2)^2. \end{aligned}$$

Summarizing, we have

$$f_{Y_1}(y_1) = \begin{cases} 4y_1^3, & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} 3(1-y_2)^2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see that the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = 12y_1^3(1-y_2)^2 = 4y_1^3 \times 3(1-y_2)^2 = f_{Y_1}(y_1)f_{Y_2}(y_2),$$

for all $(y_1, y_2) \in \mathbb{R}^2$. Therefore, Y_1 and Y_2 are independent. □

Implication: Suppose we wanted to calculate $P(Y_1 > \frac{1}{2}, Y_2 < \frac{1}{2})$ in Example 5.8. We could always calculate this using the joint pdf of Y_1 and Y_2; i.e.,

$$P\left(Y_1 > \frac{1}{2}, Y_2 < \frac{1}{2}\right) = \int_{y_1=\frac{1}{2}}^{1} \int_{y_2=0}^{\frac{1}{2}} 12y_1^3(1-y_2)^2 dy_2 dy_1.$$

However, because Y_1 and Y_2 are independent, we can write

$$P\left(Y_1 > \frac{1}{2}, Y_2 < \frac{1}{2}\right) = P\left(Y_1 > \frac{1}{2}\right) P\left(Y_2 < \frac{1}{2}\right)$$

and then make two simpler calculations using the marginal distributions above.

Example 5.9. In Example 5.6 (notes), we considered the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} e^{-y_2}, & 0 < y_1 < y_2 < \infty \\ 0, & \text{otherwise} \end{cases}$$

and derived the marginal pdfs to be

$$f_{Y_1}(y_1) = \begin{cases} e^{-y_1}, & y_1 > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_{Y_2}(y_2) = \begin{cases} y_2 e^{-y_2}, & y_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

In this example, we see that

$$f_{Y_1,Y_2}(y_1, y_2) = e^{-y_2} \neq e^{-y_1} \times y_2 e^{-y_2} = f_{Y_1}(y_1)f_{Y_2}(y_2).$$

Therefore, Y_1 and Y_2 are not independent. □

Remark: Upon closer inspection, it should be obvious that the random variables Y_1 and Y_2 in Example 5.9 are not independent. Take a look at the support:

$$R = \{(y_1, y_2) : 0 < y_1 < y_2 < \infty\}.$$

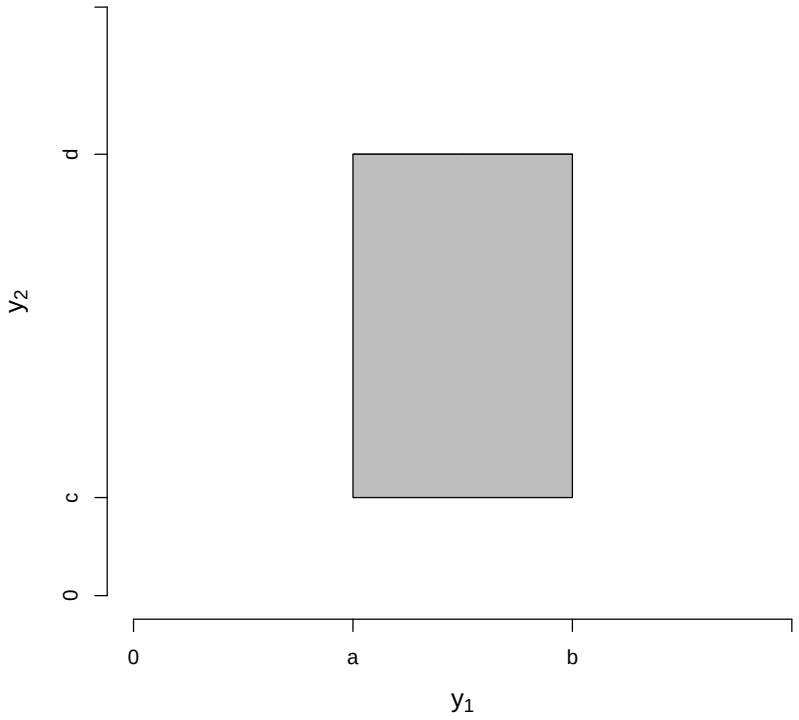

Figure 5.11: Bivariate support of the form $R = \{(y_1, y_2) : a \leq y_1 \leq b,\ c \leq y_2 \leq d\}$.

This support involves a **constraint** between y_1 and y_2, namely, y_1 is always smaller than y_2. Therefore, if someone tells you the value of Y_1, then you have information about what Y_2 is. Y_1 and Y_2 can not independent if this is true. Random variables whose joint pdf involves a constraint like the one in Example 5.9 can never be independent.

Result: Suppose Y_1 and Y_2 are continuous random variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$, which is strictly positive over the support set

$$R = \{(y_1, y_2) : a \leq y_1 \leq b,\ c \leq y_2 \leq d\},$$

where a, b, c, d are all constants; $f_{Y_1,Y_2}(y_1, y_2) = 0$, otherwise. Then Y_1 and Y_2 are independent if and only if we can write

$$f_{Y_1,Y_2}(y_1, y_2) = g(y_1)h(y_2),$$

for all $(y_1, y_2) \in \mathbb{R}^2$, where $g(y_1)$ is a nonnegative function of y_1 only and $h(y_2)$ is a nonnegative function of y_2 only.

Remark: The usefulness of this result is $g(y_1)$ and $h(y_2)$ can be *any* nonnegative functions of y_1 and y_2, respectively; they need not be pdfs. Note the support R cannot involve a constraint between y_1 and y_2 for this result to be applicable; e.g., see Figure 5.11 (above).

Proof. Suppose Y_1 and Y_2 are continuous random variables with joint pdf $f_{Y_1,Y_2}(y_1,y_2)$ and marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. Showing the necessity ($\Longrightarrow$) is easy. Suppose Y_1 and Y_2 are independent and take $g(y_1) = f_{Y_1}(y_1)$ and $h(y_2) = f_{Y_2}(y_2)$. Because $f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$, we have shown there exist nonnegative functions $g(y_1)$ and $h(y_2)$ satisfying $f_{Y_1,Y_2}(y_1,y_2) = g(y_1)h(y_2)$. Now for the sufficiency ($\Longleftarrow$). Suppose the factorization holds; i.e., suppose $f_{Y_1,Y_2}(y_1,y_2) = g(y_1)h(y_2)$, for all $(y_1,y_2) \in \mathbb{R}^2$, for nonnegative functions $g(y_1)$ and $h(y_2)$. Let

$$\int_{\mathbb{R}} g(y_1)dy_1 = c \quad \text{and} \quad \int_{\mathbb{R}} h(y_2)dy_2 = d.$$

Note that

$$cd = \int_{\mathbb{R}} g(y_1)dy_1 \int_{\mathbb{R}} h(y_2)dy_2 = \int_{\mathbb{R}}\int_{\mathbb{R}} g(y_1)h(y_2)dy_1dy_2 = \int\int_{\mathbb{R}^2} f_{Y_1,Y_2}(y_1,y_2)dy_1dy_2 = 1,$$

because the factorization $f_{Y_1,Y_2}(y_1,y_2) = g(y_1)h(y_2)$ holds by assumption. Furthermore,

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1,y_2)dy_2 \;\; = \;\; \int_{\mathbb{R}} g(y_1)h(y_2)dy_2 = dg(y_1).$$

An analogous argument shows $f_{Y_2}(y_2) = ch(y_2)$. Therefore, for all $(y_1,y_2) \in \mathbb{R}^2$, we have

$$f_{Y_1,Y_2}(y_1,y_2) = g(y_1)h(y_2) = dg(y_1)ch(y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2). \; \square$$

Note: The proof of this result in the discrete case is analogous; simply replace pdfs with pmfs and integrals with sums.

Example 5.10. Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \dfrac{cy_1^2e^{y_1}}{y_2}, & 0 < y_1 < 1,\ 1 < y_2 < 2 \\ 0, & \text{otherwise,} \end{cases}$$

where the constant satisfies $c^{-1} = (e-1)\ln 2$. We know immediately that Y_1 and Y_2 are independent because the support $R = \{(y_1,y_2) : 0 < y_1 < 1,\ 1 < y_2 < 2\}$ does not involve a constraint and

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{cy_1^2e^{y_1}}{y_2} = cy_1^2e^{y_1} \times \frac{1}{y_2} = g(y_1)h(y_2). \; \square$$

Example 5.11. Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \dfrac{1}{384}y_1^2y_2^4e^{-y_2-y_1/2}, & 0 < y_1 < \infty,\ 0 < y_2 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Again, we know immediately that Y_1 and Y_2 are independent because the support $R = \{(y_1,y_2) : 0 < y_1 < \infty,\ 0 < y_2 < \infty\}$ does not involve a constraint and

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{384}y_1^2y_2^4e^{-y_2-y_1/2} = \frac{1}{384}y_1^2e^{-y_1/2} \times y_2^4e^{-y_2} = g(y_1)h(y_2). \; \square$$

5.6 More on independence

Note: We now describe alternative ways to characterize independence of random variables Y_1 and Y_2, further implications of independence, and extensions to higher dimensions.

Terminology: Suppose that $\mathbf{Y} = (Y_1, Y_2)$ is a random vector−discrete or continuous. The **joint cumulative distribution function** (cdf) of (Y_1, Y_2) is

$$F_{Y_1,Y_2}(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2), \quad \text{for all } (y_1, y_2) \in \mathbb{R}^2.$$

As in the univariate case, a random vector's cdf completely determines its distribution.

- If Y_1 and Y_2 are discrete with joint pmf $p_{Y_1,Y_2}(y_1, y_2)$, then

$$F_{Y_1,Y_2}(y_1, y_2) = \sum_{t_2 \leq y_2} \sum_{t_1 \leq y_1} p_{Y_1,Y_2}(t_1, t_2).$$

- If Y_1 and Y_2 are continuous with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$, then

$$F_{Y_1,Y_2}(y_1, y_2) = \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} f_{Y_1,Y_2}(t_1, t_2) dt_1 dt_2$$

and

$$\frac{\partial^2 F_{Y_1,Y_2}(y_1, y_2)}{\partial y_1 \partial y_2} = f_{Y_1,Y_2}(y_1, y_2).$$

Result: Suppose Y_1 and Y_2 are random variables (discrete or continuous) with joint cdf $F_{Y_1,Y_2}(y_1, y_2)$. Then Y_1 and Y_2 are independent if and only if

$$F_{Y_1,Y_2}(y_1, y_2) = F_{Y_1}(y_1) F_{Y_2}(y_2),$$

for all $(y_1, y_2) \in \mathbb{R}^2$, where $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are the marginal cdfs of Y_1 and Y_2, respectively.

Example 5.12. In Example 5.11, we considered random variables Y_1 and Y_2 whose joint pdf was

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{1}{384} y_1^2 y_2^4 e^{-y_2 - y_1/2}, & 0 < y_1 < \infty,\ 0 < y_2 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Calculate $F_{Y_1,Y_2}(2, 3) = P(Y_1 \leq 2, Y_2 \leq 3)$.

Solution. One way to do this would be to simply work with the joint pdf and write

$$P(Y_1 \leq 2, Y_2 \leq 3) = \int_{y_1=0}^{2} \int_{y_2=0}^{3} \frac{1}{384} y_1^2 y_2^4 e^{-y_2 - y_1/2} dy_2 dy_1.$$

We could do this integral "by hand" or numerically in R.

A different way to calculate $F_{Y_1,Y_2}(2,3) = P(Y_1 \leq 2, Y_2 \leq 3)$ is to first note that Y_1 and Y_2 are independent (Example 5.11). Therefore,

$$P(Y_1 \leq 2, Y_2 \leq 3) = F_{Y_1,Y_2}(2,3) = F_{Y_1}(2)F_{Y_2}(3) = P(Y_1 \leq 2)P(Y_2 \leq 3).$$

Now, from the joint pdf (and looking at the support), you should be able to conclude that

$$\begin{aligned} Y_1 &\sim \text{gamma}(3,2) \\ Y_2 &\sim \text{gamma}(5,1). \end{aligned}$$

Therefore, $P(Y_1 \leq 2)$ and $P(Y_2 \leq 3)$ can be calculated separately from these marginal distributions; e.g.,

```
> pgamma(2,3,1/2)*pgamma(3,5,1)
[1] 0.01483462
```

Curiosity: What happens to **conditional distributions** under independence? Suppose Y_1 and Y_2 are continuous and recall the conditional pdf of Y_1 given $Y_2 = y_2$ is given by

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)},$$

wherever $f_{Y_2}(y_2) > 0$. If Y_1 and Y_2 are independent, then $f_{Y_1,Y_2}(y_1,y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$ and hence

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1}(y_1)f_{Y_2}(y_2)}{f_{Y_2}(y_2)} = f_{Y_1}(y_1).$$

In other words, the value of $Y_2 = y_2$ has no effect on the distribution of Y_1. Similarly,

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1}(y_1)f_{Y_2}(y_2)}{f_{Y_1}(y_1)} = f_{Y_2}(y_2).$$

The discrete conclusion is analogous; simply replace pdfs with pmfs.

Result: Suppose Y_1 and Y_2 are independent random variables (discrete or continuous). The random variables $U_1 = g(Y_1)$ and $U_2 = h(Y_2)$ are also independent. In other words, *functions of independent random variables are independent.*

Illustration: Suppose Y_1 and Y_2 are independent random variables. The following pairs of random variables are also independent:

$$\begin{aligned} Y_1^2 &\quad \text{and} \quad Y_2^3 \\ \sin Y_1 &\quad \text{and} \quad \ln Y_2 \\ \sqrt{Y_1 - 4} &\quad \text{and} \quad e^{-Y_2}. \end{aligned}$$

Proof. Suppose Y_1 and Y_2 are independent. Assume Y_1 and Y_2 are continuous. For any $u_1, u_2 \in \mathbb{R}$, define the sets

$$A_{u_1} = \{y_1 \in \mathbb{R} : g(y_1) \leq u_1\} \qquad \text{and} \qquad B_{u_2} = \{y_2 \in \mathbb{R} : h(y_2) \leq u_2\}.$$

The joint cdf of $U_1 = g(Y_1)$ and $U_2 = h(Y_2)$ is

$$\begin{aligned} F_{U_1,U_2}(u_1,u_2) = P(U_1 \leq u_1, U_2 \leq u_2) &= P(g(Y_1) \leq u_1, h(Y_2) \leq u_2) \\ &= P(Y_1 \in A_{u_1}, Y_2 \in B_{u_2}) = P(Y_1 \in A_{u_1})P(Y_2 \in B_{u_2}), \end{aligned}$$

because Y_1 and Y_2 are independent by assumption. Therefore, the joint pdf of $U_1 = g(Y_1)$ and $U_2 = h(Y_2)$ is

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} F_{U_1,U_2}(u_1,u_2) &= \frac{\partial^2}{\partial u_1 \partial u_2} P(Y_1 \in A_{u_1})P(Y_2 \in B_{u_2}) \\ &= \underbrace{\frac{d}{du_1} P(Y_1 \in A_{u_1})}_{\text{function of } u_1} \underbrace{\frac{d}{du_2} P(Y_2 \in A_{u_2})}_{\text{function of } u_2}. \end{aligned}$$

We have shown the joint pdf $f_{U_1,U_2}(u_1,u_2)$ factors into the product of two functions−one of which depends only on u_1 and the other which depends only on u_2. Thus, U_1 and U_2 are independent. The discrete case is analogous. □

Extension: Suppose $Y_1, Y_2, ..., Y_n$ are random variables. We call

$$\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$$

a random vector (which is n-dimensional). We say that $Y_1, Y_2, ..., Y_n$ are **mutually independent** if the joint cdf

$$\begin{aligned} F_{\mathbf{Y}}(y_1, y_2, ..., y_n) &= P(Y_1 \leq y_1, Y_2 \leq y_2, ..., Y_n \leq y_n) \\ &= P(Y_1 \leq y_1)P(Y_2 \leq y_2) \cdots P(Y_n \leq y_n) = F_{Y_1}(y_1)F_{Y_2}(y_2) \cdots F_{Y_n}(y_n). \end{aligned}$$

This is a generalization of the independence result for cdfs we stated in the two-dimensional ($n = 2$) case.

Remark: We can also describe independence in terms of joint pmfs/pdfs for random variables $Y_1, Y_2, ..., Y_n$. In the discrete case, the joint probability mass function (pmf)

$$p_{\mathbf{Y}}(y_1, y_2, ..., y_n) = P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n).$$

If $Y_1, Y_2, ..., Y_n$ are mutually independent, then the joint pmf

$$\begin{aligned} p_{\mathbf{Y}}(y_1, y_2, ..., y_n) &= P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) \\ &= P(Y_1 = y_1)P(Y_2 = y_2) \cdots P(Y_n = y_n) = \underbrace{p_{Y_1}(y_1)p_{Y_2}(y_2) \cdots p_{Y_n}(y_n)}_{\text{product of marginal pmfs}}, \end{aligned}$$

for all $\mathbf{y} = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$. In the continuous case, the joint cdf $F_{\mathbf{Y}}(y_1, y_2, ..., y_n)$ and the joint pdf $f_{\mathbf{Y}}(y_1, y_2, ..., y_n)$ are related through

$$F_{\mathbf{Y}}(y_1, y_2, ..., y_n) = \int_{-\infty}^{y_n} \cdots \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} f_{\mathbf{Y}}(t_1, t_2, ..., t_n) dt_1 dt_2 \cdots dt_n$$

and

$$\frac{\partial^n}{\partial y_1 \partial y_2 \cdots \partial y_n} F_{\mathbf{Y}}(y_1, y_2, ..., y_n) = f_{\mathbf{Y}}(y_1, y_2, ..., y_n),$$

provided this partial derivative exists. If $Y_1, Y_2, ..., Y_n$ are mutually independent, then the joint pdf

$$f_{\mathbf{Y}}(y_1, y_2, ..., y_n) = \underbrace{f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n)}_{\text{product of marginal pdfs}},$$

for all $\mathbf{y} = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$. These results will be used repeatedly in STAT 512/513.

Example 5.13. Suppose Y_1, Y_2, Y_3 are mutually independent Poisson random variables, where

$$\begin{aligned} Y_1 &\sim \text{Poisson}(\lambda_1 = 1) \\ Y_2 &\sim \text{Poisson}(\lambda_2 = 2) \\ Y_3 &\sim \text{Poisson}(\lambda_3 = 3). \end{aligned}$$

Find the joint pmf of $\mathbf{Y} = (Y_1, Y_2, Y_3)$.

Solution. Because Y_1, Y_2, Y_3 are mutually independent, the joint pmf is the product of the marginal pmfs. For each $y_i = 0, 1, 2, ...$, we have

$$p_{\mathbf{Y}}(y_1, y_2, y_3) = p_{Y_1}(y_1) p_{Y_2}(y_2) p_{Y_3}(y_3) = \frac{1^{y_1} e^{-1}}{y_1!} \times \frac{2^{y_2} e^{-2}}{y_2!} \times \frac{3^{y_3} e^{-3}}{y_3!} = \frac{e^{-6} 2^{y_2} 3^{y_3}}{y_1! y_2! y_3!}.$$

Summarizing,

$$p_{\mathbf{Y}}(y_1, y_2, y_3) = \begin{cases} \dfrac{e^{-6} 2^{y_2} 3^{y_3}}{y_1! y_2! y_3!}, & y_1 = 0, 1, 2, ...,\ y_2 = 0, 1, 2, ...,\ y_3 = 0, 1, 2, ..., \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.14. Suppose Y_1, Y_2, Y_3 are mutually independent exponential random variables, where

$$\begin{aligned} Y_1 &\sim \text{exponential}(\beta) \\ Y_2 &\sim \text{exponential}(\beta) \\ Y_3 &\sim \text{exponential}(\beta). \end{aligned}$$

Find the joint pdf of $\mathbf{Y} = (Y_1, Y_2, Y_3)$.

Solution. Because Y_1, Y_2, Y_3 are mutually independent, the joint pdf is the product of the marginal pdfs. For each $y_i > 0$, we have

$$\begin{aligned} f_{\mathbf{Y}}(y_1, y_2, y_3) &= f_{Y_1}(y_1) f_{Y_2}(y_2) f_{Y_3}(y_3) \\ &= \frac{1}{\beta} e^{-y_1/\beta} \times \frac{1}{\beta} e^{-y_2/\beta} \times \frac{1}{\beta} e^{-y_3/\beta} = \frac{1}{\beta^3} e^{-(y_1+y_2+y_3)/\beta}. \end{aligned}$$

Summarizing,

$$f_{\mathbf{Y}}(y_1, y_2, y_3) = \begin{cases} \dfrac{1}{\beta^3} e^{-(y_1+y_2+y_3)/\beta}, & y_1 > 0,\ y_2 > 0,\ y_3 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

5.7 Mathematical expectation

Terminology: Suppose $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is a random vector (discrete or continuous) and suppose g is a vector-valued function mapping vectors in n-dimensional space to the real number line; i.e., $g : \mathbb{R}^n \to \mathbb{R}$.

- If $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is discrete, then

$$E[g(Y_1, Y_2, ..., Y_n)] = \sum_{\text{all } y_n} \cdots \sum_{\text{all } y_2} \sum_{\text{all } y_1} g(y_1, y_2, ..., y_n) p_{\mathbf{Y}}(y_1, y_2, ..., y_n).$$

- If $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is continuous, then

$$E[g(Y_1, Y_2, ..., Y_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2, ..., y_n) f_{\mathbf{Y}}(y_1, y_2, ..., y_n) dy_1 dy_2 \cdots dy_n.$$

The usual existence issues arise. We need the sum/integral above to be absolutely convergent; otherwise, we say that $E[g(Y_1, Y_2, ..., Y_n)]$ does not exist.

Remark: Although these expressions apply for general random vectors $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, we are often interested in these formulas when $n = 2$. In other words, $\mathbf{Y} = (Y_1, Y_2)$ is a bivariate random vector, $g : \mathbb{R}^2 \to \mathbb{R}$, and

$$\begin{aligned} E[g(Y_1, Y_2)] &= \sum_{(y_1,y_2) \in R} \sum g(y_1, y_2) p_{Y_1,Y_2}(y_1, y_2) \qquad \text{(discrete case)} \\ E[g(Y_1, Y_2)] &= \int\int_{\mathbb{R}^2} g(y_1, y_2) f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 \qquad \text{(continuous case).} \end{aligned}$$

Recall: In Example 5.1 (notes), we examined the joint distribution of

$$\begin{aligned} Y_1 &= \text{the number of tornados recorded each year in Lee County} \\ Y_2 &= \text{the number of tornados recorded each year in Van Buren County.} \end{aligned}$$

The joint pmf of Y_1 and Y_2 was described in the following table:

$p_{Y_1,Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$
$y_1 = 0$	0.64	0.08	0.04
$y_1 = 1$	0.12	0.06	0.02
$y_1 = 2$	0.02	0.01	0.01

The expected value of the number of tornados in the two counties combined $(Y_1 + Y_2)$ is

$$\begin{aligned} E(Y_1 + Y_2) &= \sum_{(y_1,y_2) \in R} \sum (y_1 + y_2) p_{Y_1,Y_2}(y_1, y_2) \\ &= (0+0)(0.64) + (1+0)(0.12) + (2+0)(0.02) + (0+1)(0.08) + (1+1)(0.06) \\ &\quad +(2+1)(0.01) + (0+2)(0.04) + (1+2)(0.02) + (2+2)(0.01) \\ &= 0.57. \end{aligned}$$

Q: How could you calculate $V(Y_1 + Y_2)$, the variance of the total number of tornados?

A: One way you could do this is to use the variance computing formula for $Y_1 + Y_2$; i.e.,

$$V(Y_1 + Y_2) = E[(Y_1 + Y_2)^2] - [E(Y_1 + Y_2)]^2.$$

The second moment of $Y_1 + Y_2$ is

$$\begin{aligned} E[(Y_1 + Y_2)^2] &= \sum_{(y_1,y_2)\in R}\sum (y_1 + y_2)^2 p_{Y_1,Y_2}(y_1, y_2) \\ &= (0+0)^2(0.64) + (1+0)^2(0.12) + (2+0)^2(0.02) \\ &\quad +(0+1)^2(0.08) + (1+1)^2(0.06) + (2+1)^2(0.01) \\ &\quad +(0+2)^2(0.04) + (1+2)^2(0.02) + (2+2)^2(0.01) = 1.11. \end{aligned}$$

Therefore,

$$V(Y_1 + Y_2) = 1.11 - (0.57)^2 \approx 0.785.$$

Example 5.15. Suppose $\mathbf{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 6y_1^2 y_2, & 0 < y_1 < y_2,\ y_1 + y_2 < 2 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Calculate $E(Y_1Y_2)$.
(b) Calculate $E(Y_1)$.

Solutions. (a) Based on the support in Figure 5.12 (see next page), you should see that it is easier to integrate with respect to y_2 first. We have

$$\begin{aligned} E(Y_1Y_2) &= \int_{y_1=0}^{1}\int_{y_2=y_1}^{2-y_1} y_1y_2\ 6y_1^2y_2\ dy_2dy_1 \\ &= \int_{y_1=0}^{1}\int_{y_2=y_1}^{2-y_1} 6y_1^3y_2^2\ dy_2dy_1 \\ &= \int_{y_1=0}^{1} 6y_1^3\left[\left(\frac{y_2^3}{3}\right)\Big|_{y_2=y_1}^{2-y_1}\right] dy_1 \\ &= \int_{y_1=0}^{1} 2y_1^3\left[(2-y_1)^3 - y_1^3\right] dy_1 \\ &= \int_{y_1=0}^{1} 2y_1^3\left(8 - 12y_1 + 6y_1^2 - 2y_1^3\right) dy_1 \\ &= \int_{y_1=0}^{1} \left(16y_1^3 - 24y_1^4 + 12y_1^5 - 4y_1^6\right) dy_1 = \frac{16}{4} - \frac{24}{5} + \frac{12}{6} - \frac{4}{7} \approx 0.629. \end{aligned}$$

Note that had we elected to integrate in the y_1 direction first, we would have to compute

$$E(Y_1Y_2) = \underbrace{\int_{y_2=0}^{1}\int_{y_1=0}^{y_2} y_1y_2\ 6y_1^2y_2\ dy_1dy_2}_{\text{over lower triangle}} + \underbrace{\int_{y_2=1}^{2}\int_{y_1=0}^{2-y_2} y_1y_2\ 6y_1^2y_2\ dy_1dy_2}_{\text{over upper triangle}}.$$

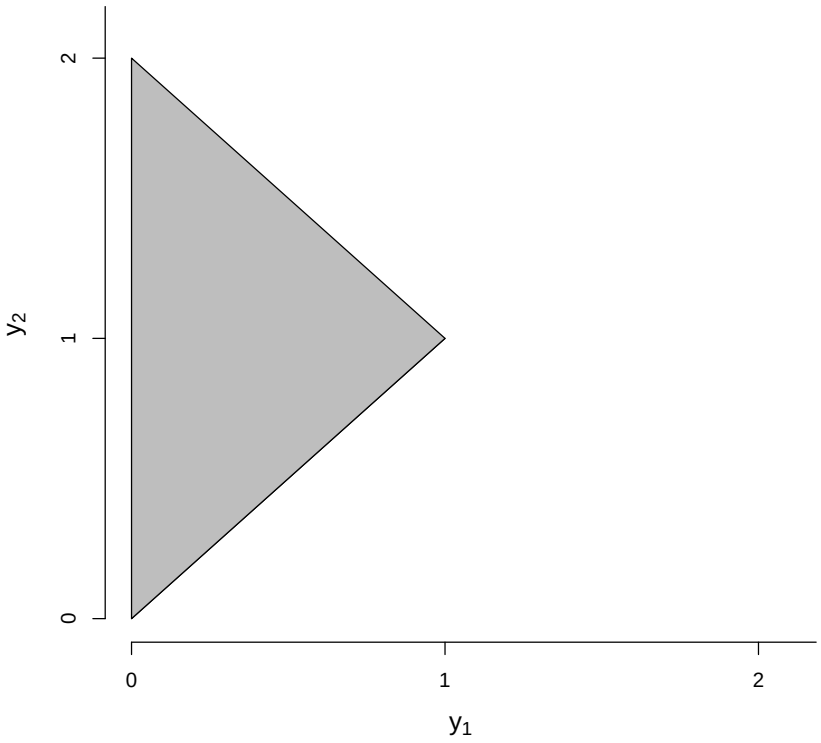

Figure 5.12: The support $R = \{(y_1, y_2) : 0 < y_1 < y_2,\ y_1 + y_2 < 2\}$ in Example 5.15. The lower boundary line is $y_2 = y_1$ and the upper is $y_2 = 2 - y_1$.

(b) We could calculate $E(Y_1)$ in two ways. Noting that $g(Y_1, Y_2) = Y_1$ is a function of Y_1 and Y_2, we could use the **joint pdf** and calculate

$$E(Y_1) = \int_{y_1=0}^{1} \int_{y_2=y_1}^{2-y_1} y_1\ 6y_1^2 y_2\ dy_2 dy_1$$

directly. The other way would be to first derive the **marginal pdf** $f_{Y_1}(y_1)$ and then calculate

$$E(Y_1) = \int_{y_1=0}^{1} y_1 f_{Y_1}(y_1) dy_1.$$

Essentially, this is what you are doing when you calculate $E(Y_1)$ using the joint pdf above anyway. Note that

$$E(Y_1) = \int_{y_1=0}^{1} \int_{y_2=y_1}^{2-y_1} y_1\ 6y_1^2 y_2\ dy_2 dy_1 \ = \ \int_{y_1=0}^{1} y_1 \underbrace{\int_{y_2=y_1}^{2-y_1} 6y_1^2 y_2\ dy_2}_{= f_{Y_1}(y_1)} dy_1.$$

For $0 < y_1 < 1$, the marginal pdf of Y_1 is

$$f_{Y_1}(y_1) = \int_{y_2=y_1}^{2-y_1} 6y_1^2 y_2\ dy_2 = 12y_1^2(1 - y_1) = \frac{\Gamma(5)}{\Gamma(3)\Gamma(2)} y_1^{3-1}(1 - y_1)^{2-1};$$

i.e., $Y_1 \sim \text{beta}(3, 2)$. Using what we know about the beta distribution, the mean is

$$E(Y_1) = \frac{\alpha}{\alpha + \beta} = \frac{3}{3+2} = 0.6. \ \square$$

Properties: The expectation operator $E(\cdot)$ enjoys the same properties when working with multivariate distributions of discrete and continuous random vectors $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$. For example, when $n = 2$, we have

1. $E(c) = c$, for any constant $c \in \mathbb{R}$
2. $E[cg(Y_1, Y_2)] = cE[g(Y_1, Y_2)]$
3. For real functions $g_1, g_2, ..., g_k$,

$$E\left[\sum_{j=1}^{k} g_j(Y_1, Y_2)\right] = \sum_{j=1}^{k} E[g_j(Y_1, Y_2)].$$

Result: Suppose Y_1 and Y_2 are **independent** random variables, and suppose $g(Y_1)$ and $h(Y_2)$ are functions of Y_1 and Y_2, respectively. Then

$$E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)],$$

provided that all expectations exist. In other words, *the expectation of the product is the product of the expectations.* This is only true when Y_1 and Y_2 are independent.

Proof. We'll prove this result in the continuous case. Suppose Y_1 and Y_2 are independent with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ and marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. Using the definition of mathematical expectation, we have

$$\begin{aligned} E[g(Y_1)h(Y_2)] &= \int\int_{\mathbb{R}^2} g(y_1)h(y_2)f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 \\ &= \int\int_{\mathbb{R}^2} g(y_1)h(y_2)f_{Y_1}(y_1)f_{Y_2}(y_2)dy_1dy_2 \\ &= \int_{\mathbb{R}} g(y_1)f_{Y_1}(y_1)dy_1 \int_{\mathbb{R}} h(y_2)f_{Y_2}(y_2)dy_2 \ = \ E[g(Y_1)]E[h(Y_2)]. \end{aligned}$$

The proof of this result in the discrete case is analogous; simply replace pdfs with pmfs and integrals with sums. $\square$

Example 5.16. In Example 5.4, we considered the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{1}{32}y_2^2 e^{-(y_1+y_2)/2}, & y_1 > 0, \ y_2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

for an engineering system with two components whose lifetimes are denoted by Y_1 and Y_2.

(a) Calculate $E(Y_1^3Y_2^2)$.
(b) Calculate the expected value of the ratio Y_1/Y_2; i.e.,

$$E\left(\frac{Y_1}{Y_2}\right).$$

Solutions. (a) We could appeal to the definition of mathematical expectation and calculate

$$E(Y_1^3Y_2^2) = \int_{y_1=0}^{\infty}\int_{y_2=0}^{\infty} y_1^3y_2^2\ \frac{1}{32}y_2^2e^{-(y_1+y_2)/2}\ dy_2dy_1$$

directly. However, we saw in Example 5.4 that

$$\begin{aligned} Y_1 &\sim \text{exponential}(2) \\ Y_2 &\sim \text{gamma}(3,2) \end{aligned}$$

and that Y_1 and Y_2 are independent. Therefore,

$$E(Y_1^3Y_2^2) = E(Y_1^3)E(Y_2^2).$$

The third moment of Y_1 is

$$E(Y_1^3) = \int_0^{\infty} y_1^3\ \frac{1}{2}e^{-y_1/2}dy_1 = \frac{1}{2}\ \Gamma(4)2^4 = 48.$$

Also, from the variance computing formula,

$$V(Y_2) = E(Y_2^2) - [E(Y_2)]^2 \implies E(Y_2^2) = V(Y_2) + [E(Y_2)]^2 = 3(2)^2 + [3(2)]^2 = 48.$$

Therefore,

$$E(Y_1^3Y_2^2) = E(Y_1^3)E(Y_2^2) = 48 \times 48 = 2304.$$

(b) Again, we could appeal to the definition of mathematical expectation and calculate

$$E\left(\frac{Y_1}{Y_2}\right) = \int_{y_1=0}^{\infty}\int_{y_2=0}^{\infty} \frac{y_1}{y_2}\ \frac{1}{32}y_2^2e^{-(y_1+y_2)/2}\ dy_2dy_1$$

directly. Alternatively, we could write

$$E\left(\frac{Y_1}{Y_2}\right) = E(Y_1)E\left(\frac{1}{Y_2}\right) = 2E\left(\frac{1}{Y_2}\right),$$

and calculate the first inverse moment $E(\frac{1}{Y_2})$ of the gamma$(3,2)$ distribution. Note that

$$E\left(\frac{1}{Y_2}\right) = \int_0^{\infty} \frac{1}{y_2}\ \frac{1}{\Gamma(3)2^3}y_2^2e^{-y_2/2}dy_2 = \frac{1}{\Gamma(3)2^3}\int_0^{\infty} y_2e^{-y_2/2}dy_2 = \frac{1}{\Gamma(3)2^3} \times \Gamma(2)2^2 = \frac{1}{4}.$$

Therefore,

$$E\left(\frac{Y_1}{Y_2}\right) = \frac{1}{2}. \ \square$$

5.8 Covariance, correlation, and the bivariate normal distribution

Recall: When the random variables Y_1 and Y_2 are **independent**, the value of one variable does not influence what the value of the other variable will be. This was illustrated precisely by examining conditional distributions; e.g.,

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)} \overset{Y_1 \perp\!\!\!\perp Y_2}{=} \frac{f_{Y_1}(y_1)f_{Y_2}(y_2)}{f_{Y_2}(y_2)} = f_{Y_1}(y_1).$$

In other words, conditioning on the value of Y_2 does not influence the distribution of Y_1. The notation $Y_1 \perp\!\!\!\perp Y_2$ is shorthand for "Y_1 and Y_2 are independent."

Note: When Y_1 and Y_2 are not independent, then

$$f_{Y_1|Y_2}(y_1|y_2) \neq f_{Y_1}(y_1)$$

which suggests that Y_1 and Y_2 are related in some way. The covariance and correlation are quantities that describe a certain type of relationship between random variables.

Terminology: Suppose Y_1 and Y_2 are random variables (discrete or continuous) with means $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$, respectively. The **covariance** of Y_1 and Y_2 is defined as

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)].$$

Note that

$$\begin{aligned}
\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] &= E(Y_1Y_2 - \mu_1Y_2 - \mu_2Y_1 + \mu_1\mu_2) \\
&= E(Y_1Y_2) - \mu_1E(Y_2) - \mu_2E(Y_1) + \mu_1\mu_2 \\
&= E(Y_1Y_2) - \mu_1\mu_2 - \mu_1\mu_2 + \mu_1\mu_2 \\
&= E(Y_1Y_2) - \mu_1\mu_2 \\
&= E(Y_1Y_2) - E(Y_1)E(Y_2).
\end{aligned}$$

This is called the **covariance computing formula**.

Interpretation: The covariance is a number that describes the linear relationship between Y_1 and Y_2:

- if $\text{Cov}(Y_1, Y_2) > 0$, then Y_1 and Y_2 are positively (linearly) related
- if $\text{Cov}(Y_1, Y_2) < 0$, then Y_1 and Y_2 are negatively (linearly) related
- if $\text{Cov}(Y_1, Y_2) = 0$, then Y_1 and Y_2 are not linearly related.

Remark: Students reading the last bullet might be tempted to infer that

$$\text{Cov}(Y_1, Y_2) = 0 \implies Y_1 \perp\!\!\!\perp Y_2.$$

Unfortunately, this is not true in general, as illustrated by the following example.

Example 5.17. Suppose $Y_1 \sim \mathcal{U}(-1, 1)$; i.e., the pdf of Y_1 is

$$f_{Y_1}(y_1) = \begin{cases} \frac{1}{2}, & -1 < y_1 < 1 \\ 0, & \text{otherwise,} \end{cases}$$

and let $Y_2 = Y_1^2$. Therefore, Y_1 and Y_2 are not independent; in fact, they are *perfectly related*. Despite this, we show $\text{Cov}(Y_1, Y_2) = 0$. Note that $E(Y_1Y_2) = E(Y_1Y_1^2) = E(Y_1^3)$. The third moment of Y_1 is

$$E(Y_1^3) = \int_{-1}^{1} y_1^3 \, \frac{1}{2} dy_1 = \frac{1}{2}\left(\frac{y_1^4}{4}\right)\Big|_{-1}^{1} = \frac{1}{2}\left(\frac{1}{4} - \frac{1}{4}\right) = 0.$$

Clearly, $E(Y_1) = 0$ because $f_{Y_1}(y_1)$ is symmetric about zero. Also,

$$E(Y_2) = E(Y_1^2) = \int_{-1}^{1} y_1^2 \, \frac{1}{2} dy_1 = \frac{1}{2}\left(\frac{y_1^3}{3}\right)\Big|_{-1}^{1} = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{3}\right) = \frac{1}{3}.$$

Therefore, from the covariance computing formula,

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = 0 - 0\left(\frac{1}{3}\right) = 0.$$

Note: This counterexample shows

$$\text{Cov}(Y_1, Y_2) = 0 \implies Y_1 \perp\!\!\!\perp Y_2.$$

The covariance describes **linear** relationships between two random variables. In Example 5.17, the random variables Y_1 and Y_2 are related (perfectly); it's just that the relationship is not linear. The covariance does not describe nonlinear relationships. □

Important: Although zero covariance does not necessarily imply independence, the converse is true; i.e.,

$$Y_1 \perp\!\!\!\perp Y_2 \implies \text{Cov}(Y_1, Y_2) = 0.$$

Proof. If Y_1 and Y_2 are independent, then $E(Y_1Y_2) = E(Y_1)E(Y_2)$. Therefore,

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = E(Y_1)E(Y_2) - E(Y_1)E(Y_2) = 0. \; \square$$

The contrapositive of this statement (which is also true) is

$$\text{Cov}(Y_1, Y_2) \neq 0 \implies Y_1 \text{ and } Y_2 \text{ are dependent.}$$

Example 5.18. Gasoline is stocked in a tank at the beginning of each week and then sold to customers during the week. Define

Y_1 = proportion of the tank available for sale after it is stocked
Y_2 = proportion sold during the week.

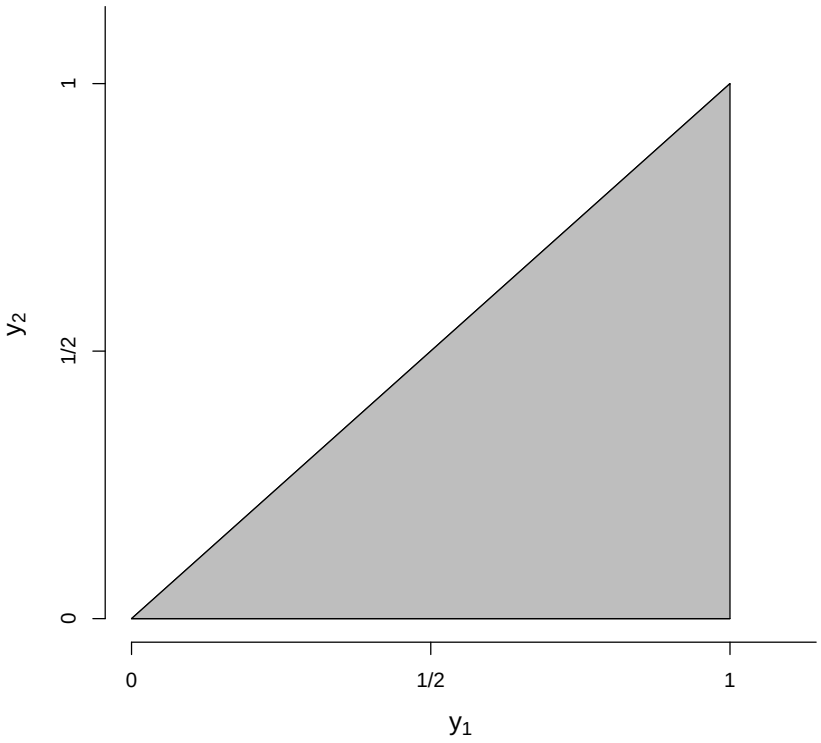

Figure 5.13: The support $R = \{(y_1, y_2) : 0 < y_2 < y_1 < 1\}$ in Example 5.18.

Suppose the joint pdf for Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Calculate the covariance of Y_1 and Y_2.

Solution. It is easiest to use the covariance computing formula. First,

$$E(Y_1Y_2) = \int_{y_1=0}^{1} \int_{y_2=0}^{y_1} y_1y_2 \; 3y_1 \; dy_2dy_1 = \int_{y_1=0}^{1} \int_{y_2=0}^{y_1} 3y_1^2y_2 \; dy_2dy_1 = \frac{3}{10}.$$

To get $E(Y_1)$ and $E(Y_2)$, we could calculate

$$\begin{aligned} E(Y_1) &= \int_{y_1=0}^{1} \int_{y_2=0}^{y_1} y_1 \; 3y_1 \; dy_2dy_1 \\ E(Y_2) &= \int_{y_1=0}^{1} \int_{y_2=0}^{y_1} y_2 \; 3y_1 \; dy_2dy_1 \end{aligned}$$

using the joint pdf. Another option is to derive the marginal pdfs of Y_1 and Y_2 first and then get $E(Y_1)$ and $E(Y_2)$ from those. This might seem like an unnecessary step (i.e., deriving the

marginal pdfs), but in some problems knowing the marginal distributions can make future calculations easier or even automatic. The marginal pdf of Y_1 is nonzero when $0 < y_1 < 1$. For these values,

$$f_{Y_1}(y_1) = \int_{y_2=0}^{y_1} 3y_1 dy_2 = 3y_1 \int_{y_2=0}^{y_1} dy_2 = 3y_1 \left(y_2 \Big|_{y_2=0}^{y_1} \right) = 3y_1^2.$$

The marginal pdf of Y_2 is also nonzero when $0 < y_2 < 1$. For these values,

$$f_{Y_2}(y_2) = \int_{y_1=y_2}^{1} 3y_1 dy_1 = \left(\frac{3y_1^2}{2} \Big|_{y_1=y_2}^{1} \right) = \frac{3}{2}(1 - y_2^2).$$

Summarizing, we have

$$f_{Y_1}(y_1) = \begin{cases} 3y_1^2, & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize $Y_1 \sim \text{beta}(3, 1)$ so

$$E(Y_1) = \frac{3}{3+1} = \frac{3}{4}.$$

The pdf of Y_2 is not one of a "named distribution," but we can easily calculate

$$E(Y_2) = \int_0^1 y_2 \, \frac{3}{2}(1 - y_2^2) dy_2 = \frac{3}{8}.$$

Therefore, the covariance of Y_1 and Y_2 is

$$\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) = \frac{3}{10} - \left(\frac{3}{4}\right)\left(\frac{3}{8}\right) = \frac{3}{160}. \; \square$$

Facts: Suppose Y, Y_1, and Y_2 are random variables (discrete or continuous). The covariance operator enjoys the following properties:

1. $\text{Cov}(Y, a) = 0$, for any constant $a \in \mathbb{R}$
2. $\text{Cov}(Y, Y) = V(Y)$
3. $\text{Cov}(Y_1, Y_2) = \text{Cov}(Y_2, Y_1)$
4. $\text{Cov}(a_1 Y_1, a_2 Y_2) = a_1 a_2 \text{Cov}(Y_1, Y_2)$, for constants $a_1, a_2 \in \mathbb{R}$.

Terminology: Suppose Y_1 and Y_2 are random variables (discrete or continuous) with means $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$ and variances $V(Y_1) = \sigma_1^2$ and $V(Y_2) = \sigma_2^2$, respectively. The **correlation** of Y_1 and Y_2 is defined as

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1 \sigma_2}.$$

Interpretation: The correlation, like the covariance, measures the strength and the direction of the linear relationship between two random variables Y_1 and Y_2. However, whereas the covariance can be any real number, the correlation is restricted to $[-1, 1]$; that is,

$$-1 \leq \rho \leq 1.$$

Proof. Define the function

$$h(t) = E\{[(Y_1 - \mu_1)t + (Y_2 - \mu_2)]^2\}.$$

Note first that $h(t) \geq 0$ for all $t \in \mathbb{R}$. Expanding the square and distributing the expectation,

$$\begin{aligned} h(t) &= E[(Y_1 - \mu_1)^2]t^2 + 2E[(Y_1 - \mu_1)(Y_2 - \mu_2)]t + E[(Y_2 - \mu_2)^2] \\ &= \sigma_1^2 t^2 + 2\text{Cov}(Y_1, Y_2)t + \sigma_2^2, \end{aligned}$$

a quadratic function of t. Because nonnegative quadratic functions can have at most one real root, the discriminant of $h(t)$; i.e., $[2\text{Cov}(Y_1, Y_2)]^2 - 4\sigma_1^2\sigma_2^2 \leq 0$. However, note that

$$\begin{aligned} [2\text{Cov}(Y_1, Y_2)]^2 - 4\sigma_1^2\sigma_2^2 \leq 0 &\iff [\text{Cov}(Y_1, Y_2)]^2 \leq \sigma_1^2\sigma_2^2 \\ &\iff -\sigma_1\sigma_2 \leq \text{Cov}(Y_1, Y_2) \leq \sigma_1\sigma_2. \end{aligned}$$

Dividing through by $\sigma_1\sigma_2$ gives

$$-1 \leq \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1\sigma_2} \leq 1. \ \square$$

Note: The correlation ρ is interpreted in the same way as the covariance; i.e.,

- if $\rho > 0$, then Y_1 and Y_2 are positively (linearly) related
- if $\rho < 0$, then Y_1 and Y_2 are negatively (linearly) related
- if $\rho = 0$, then Y_1 and Y_2 are not linearly related.

In addition,

$$Y_1 \perp\!\!\!\perp Y_2 \implies \rho = 0$$

but the relationship does not go the other way. The correlation ρ does not describe nonlinear relationships between Y_1 and Y_2. Finally, because $-1 \leq \rho \leq 1$, we have the following interpretations at the extremes:

- if $\rho = +1$, the bivariate distribution of Y_1 and Y_2 falls entirely on a straight line with positive slope
- if $\rho = -1$, the bivariate distribution of Y_1 and Y_2 falls entirely on a straight line with negative slope.

In both cases, there is a perfect linear relationship between Y_1 and Y_2.

Example 5.18 (continued). We now calculate the correlation ρ for the random variables

$$\begin{aligned} Y_1 &= \text{proportion of the tank available for sale after it is stocked} \\ Y_2 &= \text{proportion sold during the week} \end{aligned}$$

in Example 5.18. We already calculated $\text{Cov}(Y_1, Y_2) = \frac{3}{160}$, so all we need to do is calculate the marginal standard deviations. Recall that

$$f_{Y_1}(y_1) = \begin{cases} 3y_1^2, & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Recall that $Y_1 \sim \text{beta}(3, 1)$ so

$$V(Y_1) = \frac{3(1)}{(3+1)^2(3+1+1)} = \frac{3}{80} \implies \sigma_1 = \sqrt{\frac{3}{80}}.$$

The second moment of Y_2 is

$$E(Y_2^2) = \int_0^1 y_2^2 \, \frac{3}{2}(1 - y_2^2) dy_2 = \frac{1}{5}.$$

Therefore,

$$V(Y_2) = E(Y_2^2) - [E(Y_2)]^2 = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = \frac{19}{320} \implies \sigma_2 = \sqrt{\frac{19}{320}}.$$

Finally, the correlation of Y_1 and Y_2 is

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1 \sigma_2} = \frac{\frac{3}{160}}{\sqrt{\frac{3}{80}}\sqrt{\frac{19}{320}}} \approx 0.397. \; \square$$

Bivariate Normal Distribution

Terminology: The random vector (X, Y) is said to have a **bivariate normal distribution** if the joint pdf of X and Y is given by

$$f_{X,Y}(x, y) = \begin{cases} \dfrac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \, e^{-Q/2}, & (x, y) \in \mathbb{R}^2 \\ 0, & \text{otherwise,} \end{cases}$$

where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

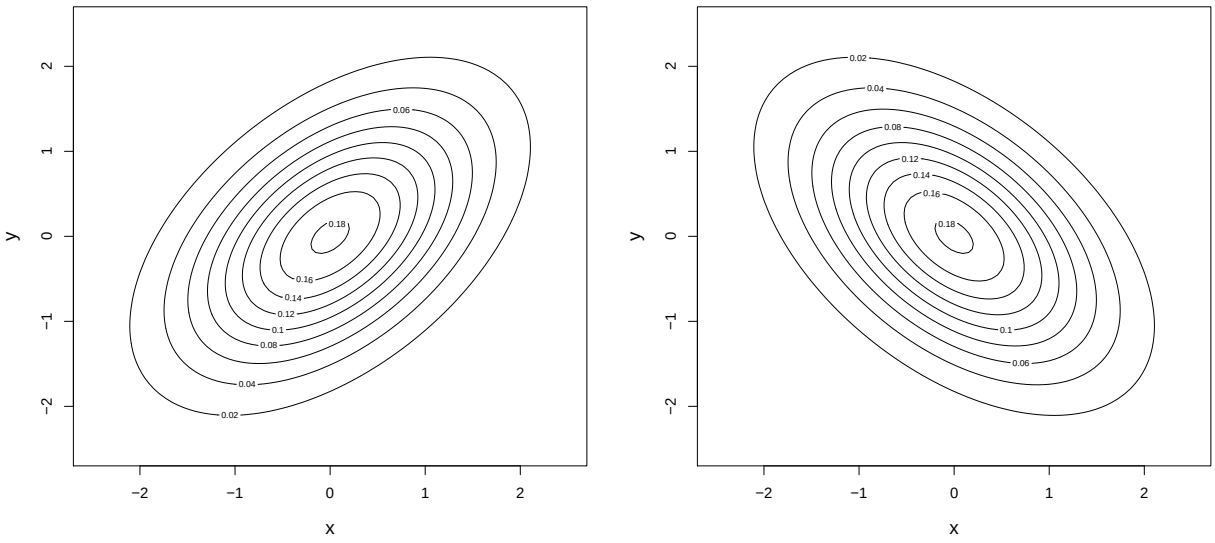

Figure 5.14: Level curves of the bivariate normal pdf with $\mu_X = \mu_Y = 0$ and $\sigma^2_X = \sigma^2_Y = 1$. Left: $\rho = 0.5$. Right: $\rho = -0.5$.

Note that there are 5 parameters that index the bivariate normal distribution:

$$E(X) = \mu_X \qquad E(Y) = \mu_Y$$
$$V(X) = \sigma^2_X \qquad V(Y) = \sigma^2_Y$$

and ρ, the correlation of X and Y.

Facts: Suppose X and Y have a bivariate normal distribution.

1. Marginal distributions are normal; i.e., $X \sim \mathcal{N}(\mu_X, \sigma^2_X)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma^2_Y)$. In other words, both X and Y have (univariate) normal distributions. To establish this, it would suffice to show

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2},$$

   for $-\infty < x < \infty$, and analogously for $f_Y(y)$. Therefore, *bivariate normality implies univariate normality*. This relationship does not go the other way.

2. In general, we learned that

$$X \perp\!\!\!\perp Y \implies \rho = 0.$$

   However, in the bivariate normal distribution, this relationship goes both ways; i.e.,

$$X \perp\!\!\!\perp Y \iff \rho = 0.$$

   This is true because $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ if and only if $\rho = 0$.

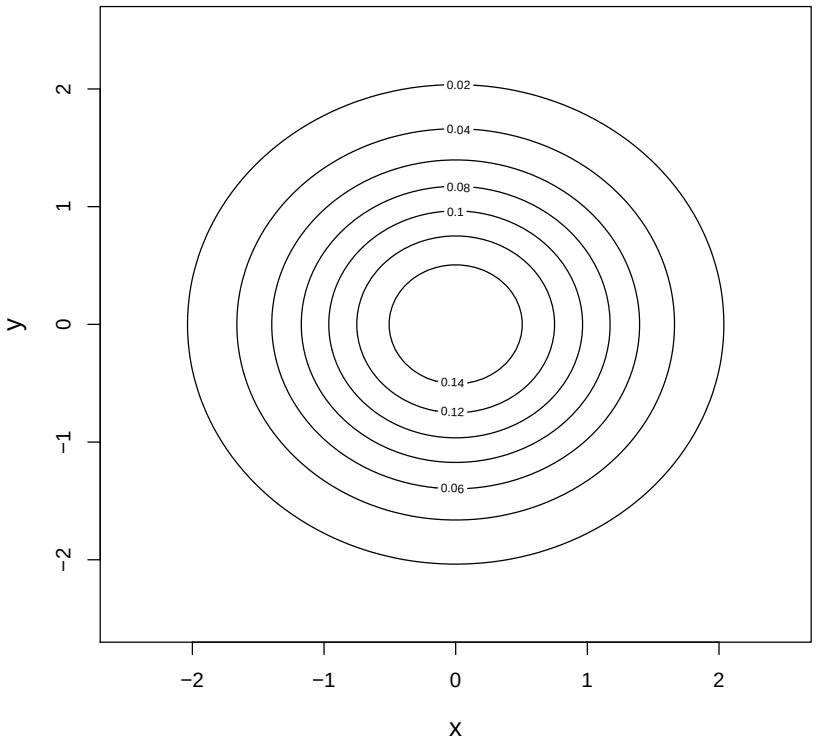

Figure 5.15: Level curves of the bivariate normal pdf with $\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, and correlation $\rho = 0$; i.e., X and Y are independent.

3. Conditional distributions are also normal. For example,

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x,\ \sigma_Y^2(1-\rho^2)),$$

where

$$\begin{aligned} \beta_0 &= \mu_Y - \beta_1 \mu_X \\ \beta_1 &= \rho\left(\frac{\sigma_Y}{\sigma_X}\right). \end{aligned}$$

This result forms the theoretical basis for **simple linear regression**, which is a common statistical technique. Note that the mean of this conditional distribution

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

is a **linear function** of x and the variance of this conditional distribution

$$V(Y|X = x) = \sigma_Y^2(1-\rho^2)$$

is free of x; i.e., the (conditional) variance of Y does not depend on the value of X.

Example 5.19. A fire in an apartment building results in a loss, X, to the owner and a loss, Y, to the tenants. The random variables X and Y have a bivariate normal distribution with

$$E(X) = 40 \qquad E(Y) = 30$$
$$V(X) = 76 \qquad V(Y) = 32$$

and $V(X|Y = 28.5) = 57$. Find $V(Y|X = 25)$.

Solution. The conditional variance formula

$$V(X|Y = y) = \sigma_X^2(1 - \rho^2)$$

applies regardless of what the value of Y is. Therefore, we have

$$57 = 76(1 - \rho^2) \implies 1 - \rho^2 = \frac{57}{76}.$$

Similarly, the conditional variance formula

$$V(Y|X = x) = \sigma_Y^2(1 - \rho^2)$$

applies regardless of what the value of X is. Therefore,

$$V(Y|X = 25) = 32\left(\frac{57}{76}\right) = 24. \ \square$$

Remark: With the information given in Example 5.19, we know that

$$1 - \rho^2 = \frac{57}{76} \implies \rho^2 = \frac{19}{76} = 0.25 \implies \rho = \pm 0.5.$$

Unfortunately, we cannot discern whether $\rho = 0.5$ or $\rho = -0.5$ unless we were provided with either of the conditional means $E(X|Y = y)$ or $E(Y|X = x)$. Intuition should suggest that $\rho > 0$ because X and Y are losses incurred by the owner and the tenants, respectively. Figure 5.16 (next page) displays the level curves of the bivariate normal pdf in Example 5.19 when $\rho = +0.5$.

5.9 Means, variances, and covariances of linear combinations

Terminology: Suppose Y_1 and Y_2 are random variables (discrete or continuous). Suppose $a_1, a_2 \in \mathbb{R}$ are constants. The random variable

$$U = a_1Y_1 + a_2Y_2$$

is called a **linear combination** of Y_1 and Y_2. The mean of U is

$$E(U) = E(a_1Y_1 + a_2Y_2) = a_1E(Y_1) + a_2E(Y_2).$$

The variance of U is

$$V(U) = V(a_1Y_1 + a_2Y_2) = a_1^2V(Y_1) + a_2^2V(Y_2) + 2a_1a_2\text{Cov}(Y_1, Y_2).$$

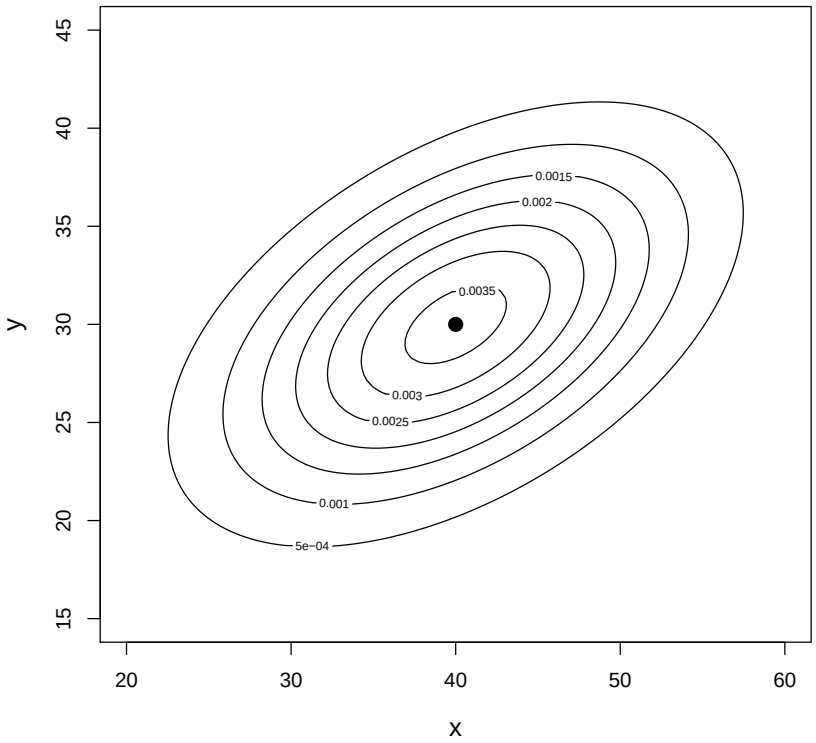

Figure 5.16: Level curves of the bivariate normal pdf in Example 5.19 with $\mu_X = 40$ and $\mu_Y = 30$, $\sigma_X^2 = 76$ and $\sigma_Y^2 = 32$, and correlation $\rho = +0.5$. A solid circle is placed at the vector of means.

Note: The following are common special cases of the variance result on the last page:

- $a_1 = a_2 = 1$:
$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2)$$
- $a_1 = 1$, $a_2 = -1$:
$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2)$$
- $Y_1 \perp\!\!\!\perp Y_2$, $a_1 = 1$, $a_2 = \pm 1$:
$$V(Y_1 \pm Y_2) = V(Y_1) + V(Y_2).$$

Remark: That
$$E(a_1Y_1 + a_2Y_2) = a_1E(Y_1) + a_2E(Y_2)$$
follows from the linearity properties of expectation. Showing
$$V(a_1Y_1 + a_2Y_2) = a_1^2V(Y_1) + a_2^2V(Y_2) + 2a_1a_2\text{Cov}(Y_1, Y_2)$$

is done as follows. Define $U = a_1Y_1 + a_2Y_2$. The variance of U is $V(U) = E(U^2) - [E(U)]^2$. Note that

$$U^2 = (a_1Y_1 + a_2Y_2)^2 = a_1^2Y_1^2 + 2a_1a_2Y_1Y_2 + a_2^2Y_2^2.$$

Therefore,

$$E(U^2) = a_1^2E(Y_1^2) + 2a_1a_2E(Y_1Y_2) + a_2^2E(Y_2^2).$$

Now,

$$\begin{aligned}[E(U)]^2 &= [a_1E(Y_1) + a_2E(Y_2)]^2 \\ &= a_1^2[E(Y_1)]^2 + 2a_1a_2E(Y_1)E(Y_2) + a_2^2[E(Y_2)]^2.\end{aligned}$$

Therefore,

$$\begin{aligned}V(U) = E(U^2) - [E(U)]^2 &= \underbrace{a_1^2E(Y_1^2) - a_1^2[E(Y_1)]^2} + \underbrace{a_2^2E(Y_2^2) - a_2^2[E(Y_2)]^2} \\ &\quad + \underbrace{2a_1a_2E(Y_1Y_2) - 2a_1a_2E(Y_1)E(Y_2)} \\ &= a_1^2V(Y_1) + a_2^2V(Y_2) + 2a_1a_2\text{Cov}(Y_1, Y_2). \ \square\end{aligned}$$

Example 5.20. In Example 5.18, we considered the random variables

$$\begin{aligned}Y_1 &= \text{proportion of the tank available for sale after it is stocked} \\ Y_2 &= \text{proportion sold during the week}\end{aligned}$$

whose joint pdf was given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The quantity $U = Y_1 - Y_2$ represents the proportion of the tank occupied by gasoline after the week (i.e., the amount of gasoline not sold). Find $E(Y_1 - Y_2)$ and $V(Y_1 - Y_2)$.

Solution. We have already calculated

$$\begin{aligned}E(Y_1) &= \frac{3}{4} \qquad E(Y_2) = \frac{3}{8} \\ V(Y_1) &= \frac{3}{80} \qquad V(Y_2) = \frac{19}{320}\end{aligned}$$

and $\text{Cov}(Y_1, Y_2) = \frac{3}{160}$. Therefore,

$$E(Y_1 - Y_2) = E(Y_1) - E(Y_2) = \frac{3}{4} - \frac{3}{8} = \frac{3}{8}$$

and

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2) = \frac{3}{80} + \frac{19}{320} - 2\left(\frac{3}{160}\right) = \frac{19}{320}. \ \square$$

Generalization: We now describe the mean and variance for a linear combination of n random variables $Y_1, Y_2, ..., Y_n$. These are natural extensions of the $n = 2$ case. Suppose $Y_1, Y_2, ..., Y_n$ are random variables (discrete or continuous). Suppose $a_1, a_2, ..., a_n \in \mathbb{R}$ are constants. The random variable

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

is a linear combination of $Y_1, Y_2, ..., Y_n$. The mean of U is

$$E(U) = E(a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n) = a_1 E(Y_1) + a_2 E(Y_2) + \cdots + a_n E(Y_n),$$

that is,

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i).$$

The variance of U is

$$\begin{aligned} V(U) = V\left(\sum_{i=1}^{n} a_i Y_i\right) &= V(a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n) \\ &= \sum_{i=1}^{n} a_i^2 V(Y_i) + \sum\sum_{i \neq j} a_i a_j \text{Cov}(Y_i, Y_j). \end{aligned}$$

Note: Because $\text{Cov}(Y_i, Y_j) = \text{Cov}(Y_j, Y_i)$, the variance formula is sometimes written as

$$V\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum\sum_{i<j} a_i a_j \text{Cov}(Y_i, Y_j).$$

The proofs of these results are analogous to the $n = 2$ case.

Result: Suppose $Y_1, Y_2, ..., Y_n$ and $X_1, X_2, ..., X_m$ are two sets of random variables. In some problems, it will be necessary to calculate the covariance of two linear combinations, say,

$$\begin{aligned} U_1 &= \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n \\ U_2 &= \sum_{j=1}^{m} b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_m X_m. \end{aligned}$$

This is done as follows:

$$\text{Cov}(U_1, U_2) = \text{Cov}\left(\sum_{i=1}^{n} a_i Y_i, \sum_{j=1}^{m} b_j X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(Y_i, X_j).$$

Note the linear combinations do not have to be of the same length. In many problems, the Y_i's and the X_j's are the same set of random variables.

Example 5.21. Suppose that Y_1, Y_2, and Y_3 are random variables with

$$\begin{array}{ccc} E(Y_1) = 1 & E(Y_2) = 2 & E(Y_3) = 3 \\ V(Y_1) = 1 & V(Y_2) = 4 & V(Y_3) = 9 \\ \text{Cov}(Y_1, Y_2) = 0 & \text{Cov}(Y_1, Y_3) = 1 & \text{Cov}(Y_2, Y_3) = -1. \end{array}$$

(a) Find the mean and variance of $U = Y_1 - 2Y_2 + 4Y_3$.
(b) Find $\text{Cov}(Y_1 + 2Y_2 - Y_3, Y_2 - 5Y_1)$.

Solutions. (a) Note that U is a linear combination with $a_1 = 1$, $a_2 = -2$, and $a_3 = 4$. The mean is

$$E(Y_1 - 2Y_2 + 4Y_3) = E(Y_1) - 2E(Y_2) + 4E(Y_3) = 1 - 2(2) + 4(3) = 9.$$

The variance is

$$\begin{aligned} V(Y_1 - 2Y_2 + 4Y_3) &= 1^2V(Y_1) + (-2)^2V(Y_2) + 4^2V(Y_3) \\ &\quad + 2(1)(-2)\text{Cov}(Y_1, Y_2) + 2(1)(4)\text{Cov}(Y_1, Y_3) + 2(-2)(4)\text{Cov}(Y_2, Y_3) \\ &= 1(1) + 4(4) + 16(9) - 4(0) + 8(1) - 16(-1) = 185. \end{aligned}$$

(b) Using the covariance result on the last page, we have

$$\begin{aligned} \text{Cov}(Y_1 + 2Y_2 - Y_3, Y_2 - 5Y_1) &= \text{Cov}(Y_1, Y_2) + \text{Cov}(Y_1, -5Y_1) + \text{Cov}(2Y_2, Y_2) + \text{Cov}(2Y_2, -5Y_1) \\ &\quad + \text{Cov}(-Y_3, Y_2) + \text{Cov}(-Y_3, -5Y_1) \\ &= \text{Cov}(Y_1, Y_2) - 5\text{Cov}(Y_1, Y_1) + 2\text{Cov}(Y_2, Y_2) - 10\text{Cov}(Y_2, Y_1) \\ &\quad - \text{Cov}(Y_3, Y_2) + 5\text{Cov}(Y_3, Y_1) \\ &= 0 - 5V(Y_1) + 2V(Y_2) - 10(0) - (-1) + 5(1) = 9. \ \square \end{aligned}$$

Example 5.22. Suppose $Y_1, Y_2, ..., Y_n$ are mutually independent random variables, each with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Define

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

the arithmetic average of $Y_1, Y_2, ..., Y_n$. In a statistical setting, $\overline{Y}$ is called the **sample mean** of $Y_1, Y_2, ..., Y_n$. Recognizing that $\sum_{i=1}^{n} Y_i$ is simply a linear combination of $Y_1, Y_2, ..., Y_n$ with

$$a_1 = a_2 = \cdots = a_n = 1,$$

find $E(\overline{Y})$ and $V(\overline{Y})$.

Solution. The mean of $\overline{Y}$ is

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu.$$

Because $Y_1, Y_2, ..., Y_n$ are mutually independent, $\text{Cov}(Y_i, Y_j) = 0$ whenever $i \neq j$. Therefore, the variance of $\overline{Y}$ is

$$V(\overline{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V(Y_i) + 2\sum\sum_{i<j}\text{Cov}(Y_i, Y_j)\right]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \square$$

Multinomial Distribution

Experiment: Perform n mutually independent trials. Each trial results in one of k distinct category outcomes:

			Probability	Count
	$\longrightarrow$	Category 1	p_1	Y_1
	$\longrightarrow$	Category 2	p_2	Y_2
Trial outcome	$\longrightarrow$	Category 3	p_3	Y_3
		$\vdots$	$\vdots$	$\vdots$
	$\longrightarrow$	Category k	p_k	Y_k

The probabilities $p_1, p_2, ..., p_k$ do not change from trial to trial and $\sum_{j=1}^{k} p_j = 1$. Define

$$Y_j = \text{ number of outcomes in Category } j \text{ (out of } n \text{ trials)},$$

for $j = 1, 2, ..., k$. We call $\mathbf{Y} = (Y_1, Y_2, ..., Y_k)$ a **multinomial random vector**. The joint probability mass function (pmf) of $Y_1, Y_2, ..., Y_k$ is

$$p_{\mathbf{Y}}(y_1, y_2, ..., y_k) = \frac{n!}{y_1!y_2!\cdots y_k!}\, p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}$$

with the support $R = \{(y_1, y_2, ..., y_k) : y_j = 0, 1, 2, ..., n; \sum_{j=1}^{k} y_j = n\}$. We write $\mathbf{Y} \sim \text{mult}(n, \mathbf{p}; \sum_{j=1}^{k} p_j = 1)$. The parameter $\mathbf{p} = (p_1, p_2, ..., p_k)$ is k-dimensional. However, because $\sum_{j=1}^{k} p_i = 1$, only $k - 1$ of these parameters are "free to vary."

Remark: The multinomial distribution is obviously an extension of the binomial distribution to allow for more than 2 categories. Recall that in the binomial distribution, there were only 2 categories: "success" and "failure;" i.e.,

			Probability	Count
Trial outcome	$\longrightarrow$	Category 1 ("success")	p	Y
	$\longrightarrow$	Category 2 ("failure")	$1-p$	$n-Y$

We write $Y \sim b(n, p)$. Because $Y + (n - Y) = n$ and $p + (1 - p) = 1$, it suffices to consider only the "success category."

Example 5.23. The State Hygienic Laboratory at the University of Iowa tests thousands of Iowa residents each year for chlamydia (CT) and gonorrhea (NG). On a given day, suppose the lab receives $n = 100$ specimens to be tested. Define

$$\begin{array}{lll} \text{Category 1:} & \text{CT}-/\text{NG}- & (p_1 = 0.90) \\ \text{Category 2:} & \text{CT}+/\text{NG}- & (p_2 = 0.07) \\ \text{Category 3:} & \text{CT}-/\text{NG}+ & (p_3 = 0.02) \\ \text{Category 4:} & \text{CT}+/\text{NG}+ & (p_4 = 0.01) \end{array}$$

and let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ denote the category counts observed after testing. Envisioning each specimen as a "trial," regarding the specimens as mutually independent, and assuming the category probabilities in $\mathbf{p} = (p_1, p_2, p_3, p_4)$ are the same for each specimen, then

$$\mathbf{Y} \sim \text{mult}\left(n = 100, \mathbf{p}; \sum_{j=1}^{4} p_j = 1\right).$$

The pmf of $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where nonzero, is given by

$$p_{\mathbf{Y}}(y_1, y_2, y_3, y_4) = \frac{100!}{y_1!y_2!y_3!y_4!}\,(0.90)^{y_1}(0.07)^{y_2}(0.02)^{y_3}(0.01)^{y_4}.$$

For example,

$$\begin{aligned} p_{\mathbf{Y}}(88, 10, 1, 1) &= P(Y_1 = 88, Y_2 = 10, Y_3 = 1, Y_4 = 1) \\ &= \frac{100!}{88!\ 10!\ 1!\ 1!}\,(0.90)^{88}(0.07)^{10}(0.02)^{1}(0.01)^{1} \approx 0.017. \ \square \end{aligned}$$

Result: If $\mathbf{Y} \sim \text{mult}(n, \mathbf{p}; \sum_{j=1}^{k} p_j = 1)$, then $Y_j \sim b(n, p_j)$, $j = 1, 2, ..., k$. That is, the category counts $Y_1, Y_2, ..., Y_k$ have marginal binomial distributions, so

$$\begin{aligned} E(Y_j) &= np_j \\ V(Y_j) &= np_j(1 - p_j), \end{aligned}$$

for $j = 1, 2, ..., k$. The covariance between any two category counts Y_j and $Y_{j'}$, $j \neq j'$, is

$$\text{Cov}(Y_j, Y_{j'}) = -np_jp_{j'}.$$

Proof. Let Y_j denote the count for Category j (i.e., "success") and collapse all other categories into "not Category j" (i.e., "failure"). Therefore, Y_j counts the number of "successes" in n Bernoulli trials. Note that we can write

$$Y_j = \sum_{i=1}^{n} Y_{ij},$$

where the random variable

$$Y_{ij} = \begin{cases} 1, & \text{individual } i \text{ in category } j \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, the mean of Y_j is

$$E(Y_j) = E\left(\sum_{i=1}^{n} Y_{ij}\right) = \sum_{i=1}^{n} E(Y_{ij}).$$

Because Y_{ij} is a discrete random variable (with 2 outcomes "1" and "0"), its mean is

$$E(Y_{ij}) = 1 \times \underbrace{P(\text{individual } i \text{ in category } j)}_{=\, p_j} \;+\; 0 \times \underbrace{P(\text{individual } i \text{ not in category } j)}_{=\, 1-p_j} = p_j.$$

Therefore,

$$E(Y_j) = \sum_{i=1}^{n} p_j = np_j.$$

The variance of Y_j is

$$V(Y_j) = V\left(\sum_{i=1}^{n} Y_{ij}\right) = \sum_{i=1}^{n} V(Y_{ij}) + \underbrace{\sum_{i=1}^{n}\sum_{\substack{i'=1 \\ i\neq i'}}^{n} \text{Cov}(Y_{ij}, Y_{i'j})}_{=\, 0} = \sum_{i=1}^{n} V(Y_{ij}),$$

because all of the covariances are 0 (i.e., Y_{ij} and $Y_{i'j}$ correspond to different trials so they are independent). We can calculate $V(Y_{ij})$ using the variance computing formula; i.e.,

$$V(Y_{ij}) = E(Y_{ij}^2) - [E(Y_{ij})]^2 = E(Y_{ij}^2) - p_j^2.$$

The second moment of Y_{ij} is

$$E(Y_{ij}^2) = 1^2 \times \underbrace{P(\text{individual } i \text{ in category } j)}_{=\, p_j} \;+\; 0^2 \times \underbrace{P(\text{individual } i \text{ not in category } j)}_{=\, 1-p_j} = p_j.$$

Therefore, $V(Y_{ij}) = E(Y_{ij}^2) - p_j^2 = p_j - p_j^2 = p_j(1-p_j)$ and

$$V(Y_j) = \sum_{i=1}^{n} p_j(1-p_j) = np_j(1-p_j).$$

Finally,

$$\text{Cov}(Y_j, Y_{j'}) = \text{Cov}\left(\sum_{i=1}^{n} Y_{ij}, \sum_{i=1}^{n} Y_{ij'}\right) = \sum_{i=1}^{n} \text{Cov}(Y_{ij}, Y_{ij'}).$$

Using the covariance computing formula, we have

$$\text{Cov}(Y_{ij}, Y_{ij'}) = \underbrace{E(Y_{ij}Y_{ij'})}_{=\, 0} - E(Y_{ij})E(Y_{ij'}) = -p_j p_{j'}.$$

Therefore,

$$\text{Cov}(Y_j, Y_{j'}) = \sum_{i=1}^{n} -p_j p_{j'} = -np_j p_{j'}. \;\square$$

Example 5.23 (continued). In Example 5.23, the lab receives $n = 100$ specimens which will be tested and then classified into one of four categories. The random variable

$$Y_1 = \text{the number of CT}-/\text{NG}- \text{ specimens (Category 1)}$$

follows a binomial distribution with $n = 100$ and $p_1 = 0.90$. In addition, with

$$Y_4 = \text{the number of CT}+/\text{NG}+ \text{ specimens (Category 4)},$$

the covariance of Y_1 and Y_4 is

$$\text{Cov}(Y_1, Y_4) = -np_1p_4 = -100(0.90)(0.01) = -0.9.$$

It makes sense that Y_1 and Y_4 have a negative (linear) relationship. As the number of disease-free specimens increases, the number of specimens with both diseases tends to decrease.

Q: How would we calculate the correlation of Y_1 and Y_4?
A: Use the definition of correlation:

$$\rho_{Y_1,Y_4} = \frac{\text{Cov}(Y_1, Y_4)}{\sqrt{V(Y_1)V(Y_4)}}$$

We have already calculated $\text{Cov}(Y_1, Y_4) = -0.9$. The variance of Y_1 is

$$V(Y_1) = np_1(1 - p_1) = 100(0.90)(0.10) = 9;$$

i.e., the variance of a binomial random variable. Similarly,

$$V(Y_4) = np_4(1 - p_4) = 100(0.01)(0.99) = 0.99.$$

Therefore,

$$\rho_{Y_1,Y_4} = \frac{-0.9}{\sqrt{9 \times 0.99}} \approx -0.302. \ \square$$

5.10 Conditional expectations

Recall: Suppose Y_1 and Y_2 are random variables. When Y_1 and Y_2 are discrete, the conditional pmf of Y_1, given $Y_2 = y_2$, is

$$p_{Y_1|Y_2}(y_1|y_2) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)},$$

whenever $p_{Y_2}(y_2) > 0$. When Y_1 and Y_2 are continuous, the conditional pdf of Y_1, given $Y_2 = y_2$, is

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)},$$

whenever $f_{Y_2}(y_2) > 0$. Recall that conditional distributions describe how one variable behaves (here, Y_1) when the other variable is fixed (here, Y_2). Of course, $p_{Y_2|Y_1}(y_2|y_1)$ and $f_{Y_2|Y_1}(y_2|y_1)$ are defined analogously.

Terminology: Suppose Y_1 and Y_2 are continuous random variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. We define **conditional expectations** as follows:

$$\begin{aligned} E[g(Y_1)|Y_2 = y_2] &= \int_{\mathbb{R}} g(y_1) f_{Y_1|Y_2}(y_1|y_2) dy_1 \\ E[h(Y_2)|Y_1 = y_1] &= \int_{\mathbb{R}} h(y_2) f_{Y_2|Y_1}(y_2|y_1) dy_2. \end{aligned}$$

Notes:

1. If Y_1 and Y_2 are discrete, then pdfs are replaced by pmfs and integrals by sums.
2. The same existence issues still remain; for example, for $E[g(Y_1)|Y_2 = y_2]$ to exist, we need $\int_{\mathbb{R}} |g(y_1)| f_{Y_1|Y_2}(y_1|y_2) dy_1 < \infty$.

Special case: If $g(Y_1) = Y_1$ and $h(Y_2) = Y_2$, then

$$\begin{aligned} E(Y_1|Y_2 = y_2) &= \int_{\mathbb{R}} y_1 f_{Y_1|Y_2}(y_1|y_2) dy_1 \\ E(Y_2|Y_1 = y_1) &= \int_{\mathbb{R}} y_2 f_{Y_2|Y_1}(y_2|y_1) dy_2. \end{aligned}$$

These are called **conditional means**. For example, $E(Y_1|Y_2 = y_2)$ is the mean of Y_1 when $Y_2 = y_2$. Similarly, $E(Y_2|Y_1 = y_1)$ is the mean of Y_2 when $Y_1 = y_1$. It is insightful to compare these formulas to the **marginal means**; i.e.,

$$\begin{aligned} E(Y_1) &= \int_{\mathbb{R}} y_1 f_{Y_1}(y_1) dy_1 \\ E(Y_2) &= \int_{\mathbb{R}} y_2 f_{Y_2}(y_2) dy_2, \end{aligned}$$

where $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ are the marginal pdfs. If Y_1 and Y_2 are discrete, then pdfs are replaced by pmfs and integrals by sums.

Important: Conditional expectations are *always* functions of the variable on which you are conditioning. Furthermore, the use of notation for the conditioning variable is important in describing whether a conditional expectation is a fixed quantity or a random variable.

$$\begin{aligned} E(Y_1|Y_2 = y_2) &\longleftarrow \text{function of } y_2\text{; fixed quantity} \\ E(Y_1|Y_2) &\longleftarrow \text{function of } Y_2\text{; random variable} \end{aligned}$$

$$\begin{aligned} E(Y_2|Y_1 = y_1) &\longleftarrow \text{function of } y_1\text{; fixed quantity} \\ E(Y_2|Y_1) &\longleftarrow \text{function of } Y_1\text{; random variable} \end{aligned}$$

Note: Computing formulas for **conditional variances** are analogous to the unconditional versions:

$$\begin{aligned} V(Y_1|Y_2 = y_2) &= E(Y_1^2|Y_2 = y_2) - [E(Y_1|Y_2 = y_2)]^2 \\ V(Y_2|Y_1 = y_1) &= E(Y_2^2|Y_1 = y_1) - [E(Y_2|Y_1 = y_1)]^2. \end{aligned}$$

Example 5.24. In Example 5.1 (notes), we examined the joint distribution of

$$\begin{aligned} Y_1 &= \text{the number of tornados recorded each year in Lee County} \\ Y_2 &= \text{the number of tornados recorded each year in Van Buren County.} \end{aligned}$$

The joint pmf of Y_1 and Y_2, $p_{Y_1,Y_2}(y_1, y_2)$, was described in the table

$p_{Y_1,Y_2}(y_1,y_2)$	$y_2=0$	$y_2=1$	$y_2=2$	$p_{Y_1}(y_1)$
$y_1=0$	0.64	0.08	0.04	0.76
$y_1=1$	0.12	0.06	0.02	0.20
$y_1=2$	0.02	0.01	0.01	0.04
$p_{Y_2}(y_2)$	0.78	0.15	0.07	

The marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$ are in the margins of this table. We derived the conditional pmf of Y_1 given $Y_2 = 0$, shown in the following table:

y_1	0	1	2
$p_{Y_1\|Y_2}(y_1\|y_2=0)$	0.820	0.154	0.026

(a) Calculate the conditional mean $E(Y_1|Y_2 = 0)$.
(b) Calculate the conditional variance $V(Y_1|Y_2 = 0)$.

Solutions. (a) The conditional mean $E(Y_1|Y_2 = 0)$ is a weighted average of the three possible values of Y_1 where the conditional probabilities $p_{Y_1|Y_2}(y_1|y_2 = 0)$ play the role of the weights; i.e.,

$$E(Y_1|Y_2=0) = \sum_{y_1=0}^{2} y_1 p_{Y_1|Y_2}(y_1|y_2=0) = 0(0.820) + 1(0.154) + 2(0.026) = 0.206.$$

(b) To find $V(Y_1|Y_2 = 0)$, we can use the variance computing formula (for conditional variances); i.e.,

$$V(Y_1|Y_2=0) = E(Y_1^2|Y_2=0) - [E(Y_1|Y_2=0)]^2.$$

The conditional second moment is

$$E(Y_1^2|Y_2=0) = \sum_{y_1=0}^{2} y_1^2 p_{Y_1|Y_2}(y_1|y_2=0) = 0^2(0.820) + 1^2(0.154) + 2^2(0.026) = 0.258.$$

Therefore,

$$V(Y_1|Y_2=0) = 0.258 - (0.206)^2 \approx 0.216. \ \square$$

Example 5.25. In Example 5.6, we considered continuous random variables Y_1 and Y_2 with joint pdf

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} e^{-y_2}, & 0 < y_1 < y_2 < \infty \\ 0, & \text{otherwise.} \end{cases}$$

We derived the conditional pdfs to be

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \dfrac{1}{y_2}, & 0 < y_1 < y_2 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} e^{-(y_2-y_1)}, & y_2 > y_1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the conditional means; i.e., calculate $E(Y_1|Y_2 = y_2)$ and $E(Y_2|Y_1 = y_1)$.

Solutions. The conditional mean of Y_1 is

$$E(Y_1|Y_2 = y_2) = \int_{\mathbb{R}} y_1 f_{Y_1|Y_2}(y_1|y_2)dy_1 = \int_0^{y_2} y_1 \, \frac{1}{y_2} dy_1 = \frac{1}{y_2}\left(\frac{y_1^2}{2}\bigg|_0^{y_2}\right) = \frac{y_2}{2}.$$

This makes sense. Note that $Y_1|Y_2 = y_2 \sim \mathcal{U}(0, y_2)$; i.e., conditional on $Y_2 = y_2$, the random variable Y_1 is uniformly distributed from 0 to y_2. The conditional mean $E(Y_1|Y_2 = y_2) = y_2/2$ is the midpoint. The conditional mean of Y_2 is

$$\begin{aligned} E(Y_2|Y_1 = y_1) = \int_{\mathbb{R}} y_2 f_{Y_2|Y_1}(y_2|y_1)dy_2 &= \int_{y_1}^{\infty} y_2 \, e^{-(y_2-y_1)} dy_2 \\ &\overset{u=y_2-y_1}{=} \int_0^{\infty} (u + y_1)e^{-u} du = E(U + y_1), \end{aligned}$$

where $U \sim$ exponential(1); in the last integral, note that e^{-u} is the pdf of $U \sim$ exponential(1) and we are integrating over $(0, \infty)$. Therefore,

$$E(Y_2|Y_1 = y_1) = E(U + y_1) = E(U) + y_1 = 1 + y_1.$$

This also makes sense. The conditional pdf of Y_2 is that of a "shifted" exponential(1) pdf, where the fixed value of y_1 denotes the shift. The conditional mean $E(Y_2|Y_1 = y_1)$ is 1 plus the shift. □

Exercise: Show the conditional variances are $V(Y_1|Y_2 = y_2) = y_2^2/12$ and $V(Y_2|Y_1 = y_1) = 1$.

Remark: In Example 5.25, we derived the conditional means

$$\begin{aligned} E(Y_1|Y_2 = y_2) &= \frac{y_2}{2} \\ E(Y_2|Y_1 = y_1) &= 1 + y_1. \end{aligned}$$

These are fixed quantities; i.e., when we condition on a fixed value of y_2 (y_1), a conditional mean is a function of this fixed value. The random versions of these are

$$\begin{aligned} E(Y_1|Y_2) &= \frac{Y_2}{2} \\ E(Y_2|Y_1) &= 1 + Y_1. \end{aligned}$$

These are functions of random variables; therefore, these are random variables themselves. As such, they have their own means, their own variances, in fact, they have their own distributions!

Result: Suppose Y_1 and Y_2 are random variables (discrete or continuous). Then

$$E(Y_1) = E[E(Y_1|Y_2)],$$

provided that all expectations exist. Similarly,

$$E(Y_2) = E[E(Y_2|Y_1)].$$

This is called **the law of iterated expectation**.

Remark: With so many expectations floating around, it is important to keep track of which distributions are being used. For example, for

$$E(Y_1) = E[E(Y_1|Y_2)],$$

there are three expectations:

$$\begin{aligned} \underline{E}(Y_1) &\longrightarrow \text{refers to the marginal distribution of } Y_1 \\ \underline{E}(Y_1|Y_2) &\longrightarrow \text{refers to the conditional distribution of } Y_1 \text{ given } Y_2 \\ \underline{E}[E(Y_1|Y_2)] &\longrightarrow \text{calculated using the marginal distribution of } Y_2. \end{aligned}$$

Remember that, in general, $E(Y_1|Y_2)$ is a function of Y_2.

Proof. We prove $E(Y_1) = E[E(Y_1|Y_2)]$ in the continuous case. Showing $E(Y_2) = E[E(Y_2|Y_1)]$ is done in the same way. Suppose Y_1 and Y_2 are continuous with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ and marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. Note that

$$\begin{aligned} E(Y_1) &= \int\int_{\mathbb{R}^2} y_1 f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} y_1 f_{Y_1|Y_2}(y_1|y_2) f_{Y_2}(y_2) dy_1 dy_2 \\ &= \int_{\mathbb{R}} \underbrace{\left[\int_{\mathbb{R}} y_1 f_{Y_1|Y_2}(y_1|y_2) dy_1 \right]}_{E(Y_1|Y_2=y_2)} f_{Y_2}(y_2) dy_2 \\ &= \int_{\mathbb{R}} E(Y_1|Y_2 = y_2) f_{Y_2}(y_2) dy_2 \;=\; E[E(Y_1|Y_2)]. \end{aligned}$$

The discrete case is proven analogously by replacing pdfs with pmfs and replacing integrals with sums. □

Example 5.26. An automobile repair shop makes an initial estimate Y_1 (in \$1000s) needed to fix a car after an accident. The probability density function (pdf) of Y_1 is

$$f_{Y_1}(y_1) = \begin{cases} \frac{1}{3} e^{-(y_1 - 0.5)/3}, & y_1 > 0.5 \\ 0, & \text{otherwise.} \end{cases}$$

Conditional on the estimate $Y_1 = y_1$, the final payment by the car owner Y_2 (in \$1000s) has a uniform distribution from $y_1 - 0.2$ and $y_1 + 0.3$. Find $E(Y_2)$, the mean final payment by the car owner.

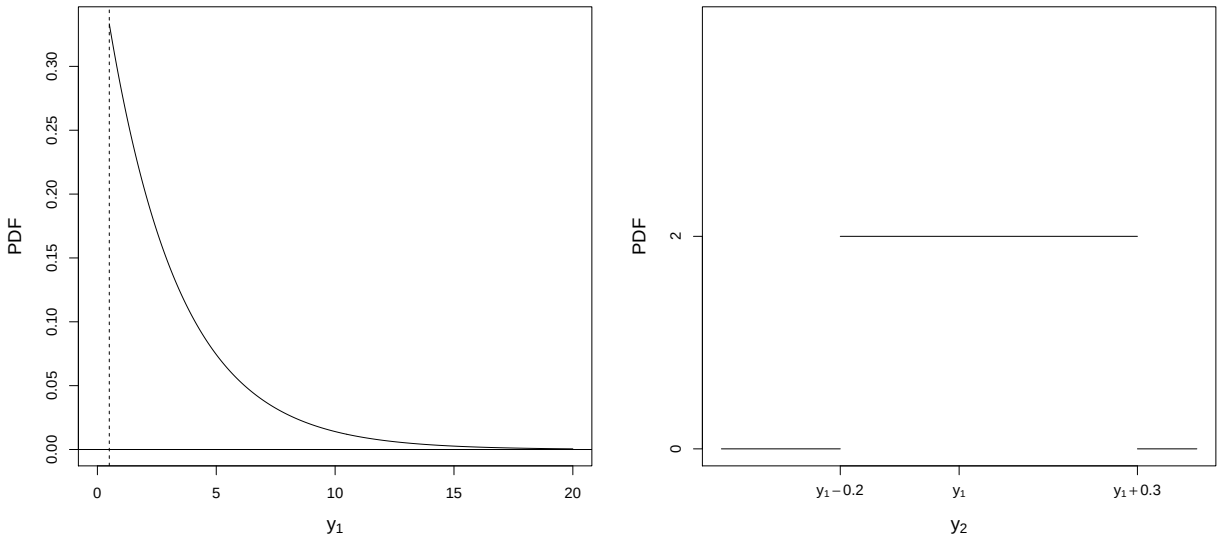

Figure 5.17: Probability density functions in Example 5.26. Left: The marginal pdf $f_{Y_1}(y_1)$. Right: The conditional pdf $f_{Y_2|Y_1}(y_2|y_1)$.

Solution. The marginal pdf $f_{Y_1}(y_1)$ is shown in Figure 5.17 above (left). We also know the conditional pdf

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} 2, & y_1 - 0.2 < y_2 < y_1 + 0.3 \\ 0, & \text{otherwise,} \end{cases}$$

which is shown in Figure 5.17 above (right). We want to find $E(Y_2)$.

Hard way: We could first derive $f_{Y_2}(y_2)$, the marginal pdf of Y_2, and then calculate $E(Y_2)$ as usual via

$$E(Y_2) = \int_{\mathbb{R}} y_2 f_{Y_2}(y_2) dy_2.$$

This is straightforward conceptually, but performing this calculation is hard. Note that the joint pdf of Y_1 and Y_2, where nonzero, is

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_2|Y_1}(y_2|y_1) f_{Y_1}(y_1) \;=\; \frac{2}{3} e^{-(y_1 - 0.5)/3}.$$

That is,

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{2}{3} e^{-(y_1-0.5)/3}, & \begin{array}{c} y_1 > 0.5 \\ y_1 - 0.2 < y_2 < y_1 + 0.3 \end{array} \\ 0, & \text{otherwise.} \end{cases}$$

The bivariate support of Y_1 and Y_2,

$$R = \{(y_1, y_2) : y_1 > 0.5,\ y_1 - 0.2 < y_2 < y_1 + 0.3\},$$

is shown in Figure 5.18 (see next page).

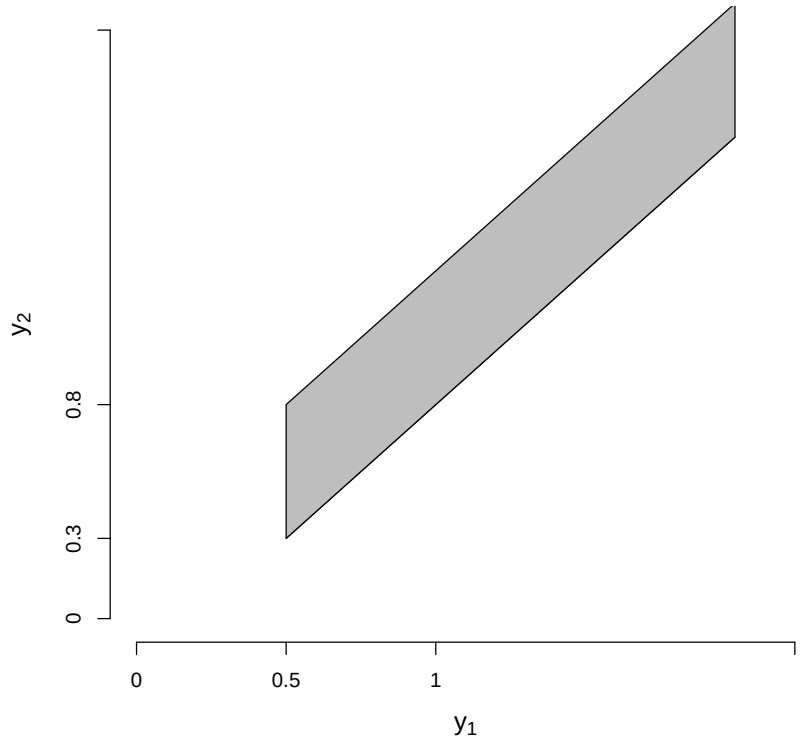

Figure 5.18: The bivariate support $R = \{(y_1, y_2) : y_1 > 0.5,\ y_1 - 0.2 < y_2 < y_1 + 0.3\}$ in Example 5.26. The lower line is $y_2 = y_1 - 0.2$. The upper line is $y_2 = y_1 + 0.3$.

Note: The marginal pdf

$$f_{Y_2}(y_2) = \int_{\mathbb{R}} f_{Y_1,Y_2}(y_1, y_2)dy_1$$

assumes different values, depending on the value of y_2.

Case 1: $0.3 < y_2 < 0.8$.

$$f_{Y_2}(y_2) = \int_{y_1=0.5}^{y_2+0.2} \frac{2}{3}e^{-(y_1-0.5)/3}dy_1 = 2\left[1 - e^{-(y_2-0.3)/3}\right].$$

Case 2: $y_2 > 0.8$.

$$f_{Y_2}(y_2) = \int_{y_1=y_2-0.3}^{y_2+0.2} \frac{2}{3}e^{-(y_1-0.5)/3}dy_1 = 2\left[e^{-(y_2-0.8)/3} - e^{-(y_2-0.3)/3}\right].$$

Summarizing,

$$f_{Y_2}(y_2) = \begin{cases} 2\left[1 - e^{-(y_2-0.3)/3}\right], & 0.3 < y_2 < 0.8 \\ 2\left[e^{-(y_2-0.8)/3} - e^{-(y_2-0.3)/3}\right], & y_2 > 0.8 \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, to find $E(Y_2)$, we would be left to calculate

$$E(Y_2) = \int_{y_2=0.3}^{0.8} 2y_2 \left[1 - e^{-(y_2-0.3)/3}\right] dy_2 + \int_{y_2=0.8}^{\infty} 2y_2 \left[e^{-(y_2-0.8)/3} - e^{-(y_2-0.3)/3}\right] dy_2.$$

Although it is possible to do both of these integrals analytically (it would not be friendly), I used R to calculate both integrals numerically and got $E(Y_2) = 3.55$.

Easy way: Use the law of iterated expectation. Write $E(Y_2) = E[E(Y_2|Y_1)]$. We know that $Y_2|Y_1 = y_1 \sim \mathcal{U}(y_1 - 0.2, y_1 + 0.3)$. Therefore,

$$E(Y_2|Y_1 = y_1) = y_1 + 0.05,$$

the midpoint of $y_1 - 0.2$ and $y_1 + 0.3$. Therefore,

$$E(Y_2) = E[E(Y_2|Y_1)] = E(Y_1 + 0.05) = E(Y_1) + 0.05.$$

The marginal mean of Y_1 is easy to calculate; i.e.,

$$E(Y_1) = \int_{\mathbb{R}} y_1 f_{Y_1}(y_1) dy_1 = \int_{0.5}^{\infty} y_1 \, \frac{1}{3} e^{-(y_1-0.5)/3} dy_1.$$

Let $u = y_1 - 0.5 \Longrightarrow du = dy_1$. Therefore,

$$E(Y_1) = \int_{0.5}^{\infty} y_1 \, \frac{1}{3} e^{-(y_1-0.5)/3} dy_1 = \int_0^{\infty} (u + 0.5) \frac{1}{3} e^{-u/3} du = E(U + 0.5) = E(U) + 0.5,$$

where $U \sim$ exponential(3). Therefore,

$$E(Y_1) = 3 + 0.5 = 3.5 \quad \Longrightarrow \quad E(Y_2) = 3.5 + 0.05 = 3.55.$$

The mean final payment by the car owner is \$3,550. $\square$

Morale: This example illustrates an important lesson when finding expected values. In some problems, it is difficult to calculate $E(Y_1)$ or $E(Y_2)$ directly. Using the law of iterated expectation can make calculations much easier.

Adam's Rule: Suppose Y_1 and Y_2 are random variables (discrete or continuous). Then

$$V(Y_1) = E[V(Y_1|Y_2)] + V[E(Y_1|Y_2)],$$

provided that all expectations and variances exist. Similarly,

$$V(Y_2) = E[V(Y_2|Y_1)] + V[E(Y_2|Y_1)].$$

This is also called **the law of iterated variances**.

Proof. We prove $V(Y_1) = E[V(Y_1|Y_2)] + V[E(Y_1|Y_2)]$. First, note that

$$\begin{aligned} E[V(Y_1|Y_2)] &= E\{E(Y_1^2|Y_2) - [E(Y_1|Y_2)]^2\} \\ &= E[E(Y_1^2|Y_2)] - E\{[E(Y_1|Y_2)]^2\} = E(Y_1^2) - E\{[E(Y_1|Y_2)]^2\}. \end{aligned}$$

Second, note that

$$\begin{aligned} V[E(Y_1|Y_2)] &= E\{[E(Y_1|Y_2)]^2\} - \{E[E(Y_1|Y_2)]\}^2 \\ &= E\{[E(Y_1|Y_2)]^2\} - [E(Y_1)]^2. \end{aligned}$$

Combining these two equations yields

$$\begin{aligned} E[V(Y_1|Y_2)] + V[E(Y_1|Y_2)] &= E(Y_1^2) - E\{[E(Y_1|Y_2)]^2\} + E\{[E(Y_1|Y_2)]^2\} - [E(Y_1)]^2 \\ &= E(Y_1^2) - [E(Y_1)]^2 = V(Y_1). \ \square \end{aligned}$$

Remark: Adam's Rule is helpful. In the same way we realized great simplification when using the law of iterated expectation; e.g.,

$$E(Y_1) = E[E(Y_1|Y_2)],$$

calculating marginal variances can also be done more easily by incorporating information on conditional distributions (through their means and variances).

Example 5.26 (continued). In Example 5.26, find $V(Y_2)$, the variance of the final payment by the car owner.

Solution. Doing things the hard way would require us to first calculate

$$E(Y_2^2) = \int_{y_2=0.3}^{0.8} 2y_2^2 \left[1 - e^{-(y_2-0.3)/3}\right] dy_2 + \int_{y_2=0.8}^{\infty} 2y_2^2 \left[e^{-(y_2-0.8)/3} - e^{-(y_2-0.3)/3}\right] dy_2$$

and then use the variance computing formula $V(Y_2) = E(Y_2^2) - [E(Y_2)]^2 = E(Y_2^2) - (3.55)^2$. The easy way is to write

$$V(Y_2) = E[V(Y_2|Y_1)] + V[E(Y_2|Y_1)].$$

Because $Y_2|Y_1 = y_1 \sim \mathcal{U}(y_1 - 0.2, y_1 + 0.3)$, we have

$$E(Y_2|Y_1 = y_1) = y_1 + 0.05$$

as before and

$$V(Y_2|Y_1 = y_1) = \frac{[(y_1 + 0.3) - (y_1 - 0.2)]^2}{12} = \frac{1}{48}.$$

Therefore,

$$V(Y_2) = E\left(\frac{1}{48}\right) + V(Y_1 + 0.5) = \frac{1}{48} + V(Y_1) = \frac{1}{48} + 9 \approx 9.021.$$

Note that the marginal pdf of Y_1,

$$f_{Y_1}(y_1) = \begin{cases} \frac{1}{3}e^{-(y_1-0.5)/3}, & y_1 > 0.5 \\ 0, & \text{otherwise}, \end{cases}$$

is a shifted exponential(3) distribution (with shift 0.5). Because the location shift does not affect the variance of the distribution, $V(Y_1) = 3^2 = 9$. $\square$

Example 5.27. A clinical trial is performed to assess the drug enzalutamide in treating women with advanced breast cancer. Suppose n patients are recruited and we record

$$Y = \text{number of patients who respond to the drug.}$$

Instead of assuming the probability of response P is the same for each patient (which would admit a binomial distribution for Y), suppose P is potentially different for each patient. In this case, P may be best regarded as random itself and having its own distribution. Suppose we assume

$$\begin{aligned} Y|P &\sim \text{binomial}(n, P) \\ P &\sim \text{beta}(\alpha, \beta). \end{aligned}$$

This is an example of a **hierarchical model**. The mean of Y can be computed using the law of iterated expectation:

$$E(Y) = E[E(Y|P)] = E(nP) = nE(P) = n\left(\frac{\alpha}{\alpha+\beta}\right).$$

The variance of Y can be computed using the law of iterated variances (i.e., Adam's Rule):

$$\begin{aligned} V(Y) = E[V(Y|P)] + V[E(Y|P)] &= E[nP(1-P)] + V(nP) \\ &= nE[P(1-P)] + n^2V(P). \end{aligned}$$

Because $P \sim \text{beta}(\alpha, \beta)$, we know

$$V(P) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

and

$$\begin{aligned} E[P(1-P)] &= \int_0^1 p(1-p)\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{(\alpha+1)-1}(1-p)^{(\beta+1)-1}dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)}. \end{aligned}$$

Therefore,

$$\begin{aligned} V(Y) &= nE[P(1-P)] + n^2V(P) \\ &= \frac{n\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)} + \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= n\left(\frac{\alpha}{\alpha+\beta}\right)\left[1-\left(\frac{\alpha}{\alpha+\beta}\right)\right] + \frac{n(n-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

Interestingly, the variance of Y, the number of patients who respond, takes the form

$$V(Y) = nE(P)[1-E(P)] + \underbrace{\frac{n(n-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}_{\text{excess variation}}.$$

This extra term acknowledges the excess variation in Y arising from treating P as random and not constant (as it is in the usual binomial distribution). □