STAT 515: STATISTICAL METHODS I

Fall 2026

Lecture Notes
Section 001

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

4 Probability and Probability Distributions

4.1 Introduction

Example 4.1. A local television station's meteorologist announces,

"There is a 30 percent chance of rain tomorrow."

How do you interpret this statement?

(a) It will rain tomorrow for 30 percent of the time. That is, for 7.2 hours tomorrow, it will be raining. For the remaining 16.8 hours, it will not be raining.

(b) It will rain tomorrow in 30 percent of the region covered by the local television station. It will not rain in the other 70 percent of the region.

(c) Among all local meteorologists, 30 percent of them think it will rain tomorrow. The remaining 70 percent of the meteorologists think it will not rain tomorrow.

(d) Thirty percent of all inhabitants of the region covered by this local television station will see rain at least once during their day tomorrow; the remaining 70 percent will not see rain during their day.

(e) It will rain on 30 percent of the days in which this same forecast is made.

Discussion: The statement incorporates uncertainty about a future event−the event that it will rain tomorrow. The phrase "30 percent" can be interpreted as a probability. None of us know for sure whether it will rain tomorrow, so we are dealing with a **random event**. We can write this as

$$P(E) = 0.30,$$

where the event

$$E = \{\text{it rains tomorrow}\}.$$

In this example, we are given five different interpretations of $P(E) = 0.30$. I think interpretation (e) makes the most sense, but all five interpretations are perfectly valid depending on how one conceptualizes the problem.

Example 4.2. I have a class with 50 students in it. Assuming there are no twins (or other multiple births), what is the probability there is at least one shared birthday among the students? Remember there are 365 days in a year.

(a) $P(E) = 0.14$
(b) $P(E) = 0.28$
(c) $P(E) = 0.45$
(d) $P(E) = 0.81$
(e) $P(E) = 0.97$

Interpreting probabilities: The authors of your textbook describe three ways to think about probability:

1. Classical approach
2. Relative frequency approach
3. Subjective or personal approach.

Classical approach: We envision a sample space for a phenomenon of interest. A **sample space** is a collection of all possible outcomes that could occur. It is usually denoted by S. An **event** E, like the event that it rains or that there is a shared birthday, is a subset of the sample space. If there are a total of N outcomes in the sample space, N_e outcomes in E, and each outcome in S is equally likely, then the probability of E is simply

$$P(E) = \frac{N_e}{N}.$$

Example 4.3. Pick 3 is a three-digit number game from the South Carolina Education Lottery. We can think of playing this lottery as a random experiment with sample space

$$S = \{000, 001, 002, ..., 998, 999\}.$$

The number of outcomes in S is $N = 1000$. Suppose I purchase three tickets with the following numbers:

$$E = \{364, 450, 545\}.$$

The probability I win is

$$P(E) = \frac{N_e}{N} = \frac{3}{1000} = 0.003.$$

This calculation assumes each of the 1000 outcomes in S is equally likely.

Relative frequency approach: Another way to think about $P(E)$ is that it represents the long-run proportion of times E would occur if we were to repeat the same phenomenon over and over again. The event E will occur on some repetitions and not on others; $P(E)$ is the proportion of times it will occur in the long-run. This is the **relative frequency** interpretation of probability.

Example 4.4. Plastic bottles for liquid laundry detergent are formed by blow molding, a manufacturing process used to create hollow plastic parts by inflating a heated plastic tube inside a mold. Previous data from statistical process control monitoring suggests 10 percent of the bottles do not conform to specifications. Imagine observing bottles one-by-one over a long period of time and, for each bottle, define

$$E = \{\text{bottle does not conform to specifications}\}.$$

For each bottle observed, we observe if E occurs or not. Figure 4.1 graphs what the relative frequency of the event E might look like over time when observing 5000 bottles.

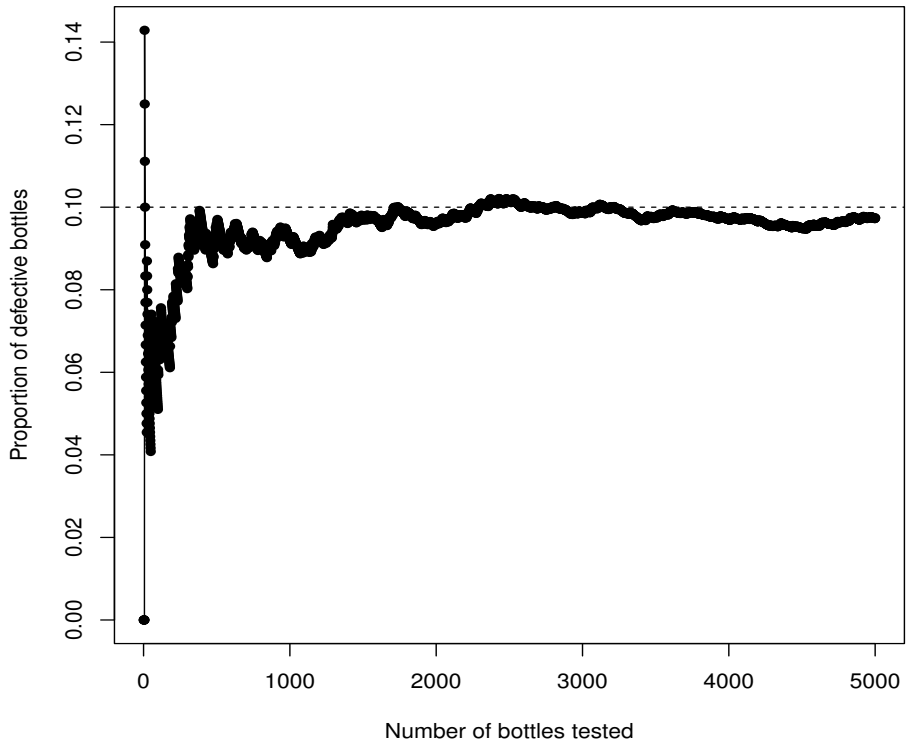

Figure 4.1: Relative frequency plot of the proportion of defective bottles (i.e., bottles that do not conform to specifications). A total of 5000 bottles are observed. A dotted horizontal line at 0.10 is added.

Note: In this simulation, the proportion (relative frequency) of bottles that did not conform to specifications was:

```
> prop[5000]
[1] 0.0974
```

so we would assign

$$P(E) \approx \frac{487}{5000} = 0.0974$$

as an **estimate** of $P(E)$. In general, if we repeat the simulation a total of n times, and n_e counts the number of times E occurs, then

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

This explains why the relative frequency gets closer and closer to 0.10 in Figure 4.1 above. In other words, $P(E)$ describes "long-term behavior" of a random process (not what happens in the short-term).

Subjective or personal approach: This approach relies on an investigator's prior experience and/or subject-matter knowledge.

Example 4.5. Suppose you woke up this morning with a sore throat.

- You find out from your social media accounts that "strep throat is going around." You assign a personal probability of 0.50 to the outcome you have strep throat.
- You then look in the mirror to see what your throat looks like. Your tonsils appear to be red and swollen. You update your personal probability to 0.75.
- Later that day, your doctor tells you, "it looks like strep throat but I can't be sure." You update your personal probability to 0.90.
- An in-house strep test at your doctor's office says you don't have strep. You update your personal probability to 0.10.
- A culture is performed on a swab specimen from your tonsils and is negative for strep bacteria. You update your personal probability to 0.01. Based on this, you conclude you likely don't have strep throat.

Remark: As we continue to get more information, our personal probabilities may change along with it to incorporate that information. This is the central idea behind **Bayesian statistics**. This is a framework for how we update probabilities in the presence of new information (data).

4.2 Probability axioms and laws

Definitions: The **union** of two events A and B is the event containing outcomes in A or B. The **intersection** of two events A and B is the event containing outcomes in A and B. We write

$$\begin{aligned} A \cup B &\longleftarrow \text{union ("or")} \\ A \cap B &\longleftarrow \text{intersection ("and").} \end{aligned}$$

Definition: If the events A and B contain no common outcomes, we say the events are **mutually exclusive**. In this case,

$$P(A \cap B) = P(\emptyset) = 0.$$

In other words, it is not possible for A and B to both occur.

Example 4.6. 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). Imagine observing the next male patient as a random experiment. The sample space is

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

Remarks:

- 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. This illustrates the **multiplication rule of counting**.
- 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; fewer have low SBP.

Q: List the outcomes in $A \cup B$ and $A \cap B$.

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

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

Axioms: Going forward, we need a formal set of rules (or axioms) which will govern how we calculate probabilities in general. The following axioms are due to Kolmogorov. For any sample space S, assigning a probability P must satisfy

(1) $0 \leq P(A) \leq 1$, for any event A

(2) $P(S) = 1$

(3) If $A_1, A_2, ..., A_n$ are pairwise **mutually exclusive** events, then

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

- The term "pairwise mutually exclusive" means that $A_i \cap A_j = \emptyset$, for all $i \neq j$. That is, the intersection of any two events is empty.
- The event

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n$$

means "at least one A_i event occurs."

Discussion: The first axiom guarantees that probabilities must be between 0 and 1. Events with probability 0 can never occur. Events with probability 1 must occur. Both extremes are rare in real life. The third axiom provides the mathematical basis for intuitive calculations like in the following example.

Example 4.7. In the game of craps, two fair dice are rolled initially to start the game. Here is a probability model for the sum of the two faces.

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Q: What is the probability of rolling a "7" or an "11?"
A: The third axiom assures we can add the probabilities, that is,

$$P(\text{rolling a 7 or 11}) = P(\text{rolling a 7}) + P(\text{rolling an 11}) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} \approx 0.222.$$

This is true because

$$\{\text{rolling a 7}\} \quad \text{and} \quad \{\text{rolling an 11}\}$$

are mutually exclusive events.

Complement Rule: The complement of the event A consists of all outcomes not in A. The complement is denoted by $\overline{A}$. The first two axioms can be used to show

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

Additive Rule: If A and B are two events, then

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

Of course, if A and B are mutually exclusive, then $P(A \cap B) = 0$ and

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

This agrees with Axiom 3.

DeMorgan's Laws: If A and B are two events, then

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

Example 4.8. The probability that train 1 is on time is 0.95. The probability that train 2 is on time is 0.93. The probability that both are on time is 0.90. Define the events

$$\begin{aligned} A &= \{\text{train 1 is on time}\} \\ B &= \{\text{train 2 is on time}\}. \end{aligned}$$

We are given $P(A) = 0.95$, $P(B) = 0.93$, and $P(A \cap B) = 0.90$.

Q: What is the probability train 1 is not on time?

$$\begin{aligned} P(\overline{A}) &= 1 - P(A) \\ &= 1 - 0.95 = 0.05. \end{aligned}$$

Q: What is the probability at least one train is on time?

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.95 + 0.93 - 0.90 = 0.98. \end{aligned}$$

Q: What is the probability neither train is on time?
A: Note that by DeMorgan's Law,

$$\{\text{neither train on time}\} = \overline{A} \cap \overline{B} = \overline{A \cup B}.$$

Therefore, by the complement rule,

$$\begin{aligned} P(\overline{A \cup B}) &= 1 - P(A \cup B) \\ &= 1 - 0.98 = 0.02. \end{aligned}$$

4.3 Conditional probability and independence

Remark: In many situations, we want to calculate $P(A)$. However, this probability might be influenced by another event B. That is, the occurrence of B may influence how we assign probability to A. This motivates conditional probability.

Definition: Let A and B be events in a sample space S with $P(B) > 0$. The **conditional probability** of A, given that B has occurred, is

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

Similarly,

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

provided $P(A) > 0$.

Example 4.9. In a manufacturing process, 10% of the parts contain surface flaws and 25% of the parts with surface flaws are defective. However, only 5% of parts without surface flaws are defective. There are two events of interest here:

$$\begin{aligned} D &= \{\text{part is defective}\} \\ F &= \{\text{part has surface flaws}\}. \end{aligned}$$

We are given $P(F) = 0.10$, $P(D|F) = 0.25$, and $P(D|\overline{F}) = 0.05$. Therefore, how we assign probability to D depends on whether F occurs; i.e., it depends on whether the part has surface flaws (F) or not ($\overline{F}$).

Example 4.10. 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.

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 $P(A \cap B) = 0.01$.

Q: Find the probability of catching the usual strain given that the new strain is caught.
A: Using the definition of conditional probability,

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

This means among all residents who have caught the new strain, 20% of them will also catch the usual strain. Notice how $P(A)$ and $P(A|B)$ are different. That is, the occurrence of B has changed how we assign probability to A.

Q: Find the probability of catching the new strain, given that at least one strain is caught.
A: If "at least one strain is caught," this means $A \cup B$ has occurred. Therefore,

$$\begin{aligned} P(B|A \cup B) = \frac{P(B \cap (A \cup B))}{P(A \cup B)} &= \frac{P(B)}{P(A) + P(B) - P(A \cap B)} \\ &= \frac{0.05}{0.10 + 0.05 - 0.01} \approx 0.357. \end{aligned}$$

This means among all residents who have caught at least one strain, 35.7% of them have caught the new strain.

Multiplication Rule: If we take the conditional probability definitions

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

we see that

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

Important: In general, we cannot determine $P(A \cap B)$ from $P(A)$ and $P(B)$ alone. We need more information or we have to make additional assumptions about A and B.

Example 4.11. A corporation is proposing to select 2 of its current regional managers as vice presidents. The corporation has 6 male regional managers and 4 female regional managers. Make the assumption the 10 regional managers are equally qualified and hence all possible groups of 2 managers have the same chance of being selected as the vice presidents.

Q: Find the probability both vice presidents are male.
A: We can think of the sample space as

$$S = \{\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3, \mathrm{M}_4, \mathrm{M}_5, \mathrm{M}_6, \mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \mathrm{W}_4\}.$$

Each outcome in S is equally likely. Define the events

$$\begin{aligned} A &= \{\text{first selection is male}\} \\ B &= \{\text{second selection is male}\}. \end{aligned}$$

We want $P(A \cap B)$. It is easy to see that

$$P(A) = \frac{6}{10} \quad \text{and} \quad P(B|A) = \frac{5}{9}.$$

Therefore,

$$P(A \cap B) = P(A)P(B|A) = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}.$$

Curiosity: $P(A)$ and $P(A|B)$ are both probabilities for the event A. In the first, we look at A by itself. In the second, we incorporate knowledge that the event B has occurred. What does it mean when $P(A) = P(A|B)$? It means the knowledge gained from learning B occurred does not influence how we assign probability to the event A. This is the casual definition of **independence**.

Definition: When the occurrence or non-occurrence of B has no effect on whether or not A occurs, and vice-versa, we say the events A and B are **independent**. Mathematically, we define A and B to be independent if and only if

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

Note that if A and B are independent, then

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

and

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

Example 4.12. An electrical circuit consists of two components (e.g., resistors, capacitors, batteries, etc.). The probability the second component functions satisfactorily during its design life is 0.90. The probability at least one of the components does so is 0.96. The probability both components do so is 0.75.

Q: Do the two components function independently?
A: Define the events

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

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. The two components do not function independently.

4.4 Law of Total Probability and Bayes' Rule

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

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

Therefore,

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

Remark: The Law of Total Probability 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 probability of a related event B. Specifically, $P(A)$ is a linear combination of the conditional probabilities $P(A|B)$ and $P(A|\overline{B})$. Note that the "weights" in the linear combination, $P(B)$ and $P(\overline{B})$, add to 1.

Example 4.13. 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 30 percent accident-prone. 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 $P(B) = 0.3$.

Q: What is the probability a policy holder has an accident?
A: We want $P(A)$, which can be found by using the LOTP:

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

Therefore, 26% of the company's policy holders will have an accident.

Q: If a policy holder has an accident, what is the probability s/he was "accident-prone?"
A: We want $P(B|A)$, which can be calculated as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{0.4(0.3)}{0.26} \approx 0.462.$$

Therefore, among all policy holders who had an accident, 46.2% of them are accident-prone.

Discussion: In the last part, note that if we write

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

we obtain **Bayes' Rule** for two events.

Example 4.14. *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. Define the events

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

We are given the following information:

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

Q: What is the probability a randomly selected subject will test positively?
A: We want $P(A)$. By the LOTP,

$$P(A) = P(A|D)P(D) + P(A|\overline{D})P(\overline{D}) = 0.95(0.08) + 0.01(0.92) \approx 0.085.$$

Therefore, about 8.5% of the population will produce a positive test result.

Q: What is the probability a subject has the disease if his test is positive?
A: We want $P(D|A)$. By Bayes' Rule,

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

Therefore, among all subjects testing positively, about 89.2% of the subjects have the disease.

Remark: In general, Bayes' Rule allows us to update probabilities on the basis of observed information. In Example 4.14, the "observed information" we get to see is the test result.

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

Discussion: The formulas for LOTP and Bayes' Rule are both based on the fact that B and $\overline{B}$ **partition** the sample space S. That is, B and $\overline{B}$ are mutually exclusive and $B \cup \overline{B} = S$. We can restate these rules for a more general partition of S. Suppose $B_1, B_2, ..., B_k$ are pairwise mutually exclusive and $B_1 \cup B_2 \cup \cdots \cup B_k = S$. LOTP and Bayes' Rule become

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k)$$

and

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k)}.$$

Example 4.15. 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.

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 $P(A|B_3) = 0.01$.

Q: What is the probability a policy holder of this certain age dies next year?
A: We want $P(A)$. By the LOTP,

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

Q: A policy holder of this certain age dies next year. What is the probability the deceased was a preferred policy holder?
A: 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.

4.5 Introduction to random variables

Terminology: A **random variable** is a variable whose value is determined by chance.

- By convention, we denote random variables by upper case letters towards the end of the alphabet; e.g., W, X, Y, Z, etc. The authors of the textbook for this course (OL, 2016) favor the use of Y.
- Possible values of a random variable are denoted by the lower case versions; e.g., $X = x$, $Y = y$, etc.

Terminology: Random variables generally break down into one of two types.

- If a random variable Y can have only a finite (or countable) number of values, we say it is **discrete**.
- If it makes more sense that Y has values in an interval of numbers, we say Y is **continuous**.
  - Measurements like part diameters (cm), temperature (deg C), energy expended (kcal), time (days, years, etc.), and distance (miles) are all best regarded as continuous random variables.

Terminology: The set of all values a random variable Y can have is called its **support**.

Example 4.16. Classify the following random variables as discrete or continuous and specify the support of each:

$$\begin{aligned} V &= \text{number of unbroken eggs in a randomly selected carton (dozen)} \\ W &= \text{pH of an aqueous solution} \\ X &= \text{length of time between accidents at a factory} \\ Y &= \text{whether or not you pass this class} \\ Z &= \text{number of aircraft arriving tomorrow at CAE.} \end{aligned}$$

- The random variable V is **discrete**. It can have values in

$$\{0, 1, 2, ..., 12\}.$$

- The random variable W is **continuous**. It can have values in

$$\mathbb{R} = \{-\infty < w < \infty\}.$$

With most solutions, it is more likely that W is not negative (although this is possible) and not larger than, say, 15 (a very reasonable upper bound).

- The random variable X is **continuous**. It can have values in

$$\mathbb{R}^+ = \{x > 0\}.$$

The key point here is that a time cannot be negative. In theory, it is possible that X can be very large.

- The random variable Y is **discrete**. It can have values in

$$\{0, 1\},$$

where I have arbitrarily labeled "1" for passing and "0" for failing. Random variables that can assume exactly two values (e.g., 0, 1) are **binary**.

- The random variable Z is **discrete**. It can have values in

$$\mathbb{N} = \{0, 1, 2, 3, ..., \}.$$

I have allowed for the possibility of a very large number of aircraft arriving.

4.6 Discrete random variables

4.6.1 Probability mass functions

Recall: A random variable Y is **discrete** if it can have a finite or countable number of values.

Terminology: The **probability mass function** (pmf) of a discrete random variable Y tells us two things:

1. the values Y can have
2. a probability $p_Y(y) = P(Y = y)$ for each value of y.

The pmf of Y,

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

describes the **distribution** of Y. It tells us which values of y are possible and how to assign probabilities to these values.

Example 4.17. 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

This pmf is shown in Figure 4.2 (next page). Note how the probabilities sum to 1 (as they should).

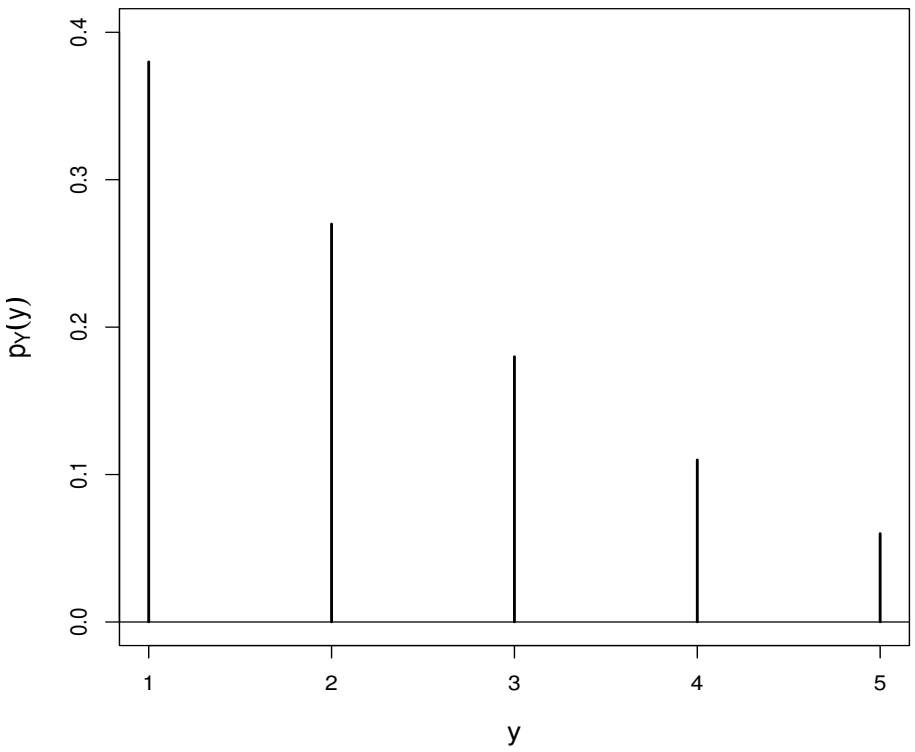

Figure 4.2: Probability mass function of Y in Example 4.17.

Q: What percentage of the population gives a pain rating of 3 or higher? 2 or lower?
A: We have

$$P(Y \geq 3) = P(Y = 3) + P(Y = 4) + P(Y = 5) = 0.18 + 0.11 + 0.06 = 0.35.$$

Therefore, 35% of the population will give a pain rating of 3 or higher. We also have

$$P(Y \leq 2) = P(Y = 1) + P(Y = 2) = 0.38 + 0.27 = 0.65.$$

Therefore, 65% of the population will give a pain rating of 2 or lower.

4.6.2 Mean and variance

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

$$\mu = E(Y) = \sum_{\text{all } y} y p_Y(y).$$

The expected value of a discrete random variable Y is a weighted average of the possible values of Y. Each value y is weighted by its probability $p_Y(y)$.

Note: In statistical applications, $\mu = E(Y)$ is called the **mean** or **population mean**. It is the mean value of Y for all individuals in the population under study.

Example 4.17 (continued). Find the population mean pain rating in Example 4.17.
A: The expected value is

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

In this application, it would be appropriate to call $E(Y) = \mu = 2.2$ the **population mean**. It is the mean pain rating for all patients in the population under study.

Interpretations:

- $E(Y)$ is the "center of gravity" of a discrete random variable's pmf. It's the location on the horizontal axis where $p_Y(y)$ would balance if it were made of solid material.
- $E(Y)$ is a "long-run average." That is, if we were to observe the value of Y for many patients in the population in Example 4.17, the average of these observations should be "close" to $E(Y)$. The more observations there are, the closer this average will be to $E(Y)$. For example, the R code below simulates 500 measurements of Y in Example 4.17 and averages them.

```
> options(digits=3) # control number of digits presented
> n = 500
> y = c(1,2,3,4,5)
> prob = c(0.38,0.27,0.18,0.11,0.06)
> rating = sample(y,n,replace=TRUE,prob=prob)
> mean(rating)
[1] 2.26
```

  This is an illustration of a statistical result known as the **Law of Large Numbers**. This law says the sample mean (e.g., of the 500 ratings we simulated) will "converge" to the population mean $E(Y)$ when the number of observations gets large without bound.

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

$$\sigma^2 = V(Y) = \sum_{\text{all } y} (y - \mu)^2 p_Y(y).$$

The **standard deviation** of Y is the positive square root of the variance:

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

Example 4.17 (continued). Find the population variance $\sigma^2 = V(Y)$ in Example 4.17.
A: The variance is

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

Similarly, it would be appropriate to call $V(Y) = \sigma^2 = 1.5$ the **population variance**. It is the variance associated with the pain ratings for all patients in the population. The population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.5} \approx 1.22.$$

Interpreting the variance and standard deviation:

1. Whereas $E(Y)$ measures the "center" of a distribution, the variance $V(Y)$ measures the "spread," that is, how spread out the values of Y are about the mean.
2. The larger $V(Y)$ is, the more spread (variability) in the distribution of Y.
3. The variance $V(Y) \geq 0$. The only time $V(Y) = 0$ is when Y has a **degenerate distribution**; i.e., all the probability mass is at one point.
4. $V(Y)$ is measured in (units)^2 and σ is measured in original units.

4.7 Binomial distribution

Bernoulli trials: Many random experiments can be envisioned as consisting of a sequence of "trials," where

1. each trial results in a "success" or a "failure" (only 2 outcomes are possible)
2. the trials are independent (the result on one trial is not affected by the results from other trials)
3. the probability of success π is the same on every trial.

Examples:

- When circuit boards used in the manufacture of laptops are tested, one percent of the boards are found to be defective.
  - circuit board = "trial"
  - defective board is observed = "success"
  - $\pi = 0.01$
- Ninety-eight percent of all air traffic radar signals are correctly interpreted the first time they are transmitted.
  - radar signal = "trial"
  - signal is correctly interpreted = "success"
  - $\pi = 0.98$

- Albino rats used to study the hormonal regulation of a metabolic pathway are injected with a drug that inhibits body synthesis of protein. Twenty percent of all rats will die before the study is complete.
  - rat = "trial"
  - dies before study is over = "success"
  - $\pi = 0.2$
- During her WNBA tenure, Caitlyn Clark's free throw percentage is 88.7%.
  - free throw = "trial"
  - made free throw = "success"
  - $\pi = 0.887$

Definition: A **binomial distribution** arises when we observe a fixed number of Bernoulli trials:

$$\begin{aligned} n &= \text{number of trials} \\ Y &= \text{number of successes (out of } n\text{)}. \end{aligned}$$

If the Bernoulli trial assumptions hold, then the probability mass function (pmf) of Y is given by the formula

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

We write $Y \sim b(n, \pi)$, where π is the probability of success on any one trial.

Example 4.18. In an agricultural study, 40 percent of all plots respond to a certain treatment. In this context, we interpret

- plot = "trial"
- plot responds to treatment = "success"
- $\pi = 0.4$

Four plots are observed. If the Bernoulli trial assumptions hold (independent plots, same response probability for each plot), then

$$Y = \text{number of plots responding to treatment} \sim b(n = 4, \pi = 0.4).$$

That is, the random variable Y has a binomial distribution with $n = 4$ trials and probability of success $\pi = 0.4$.

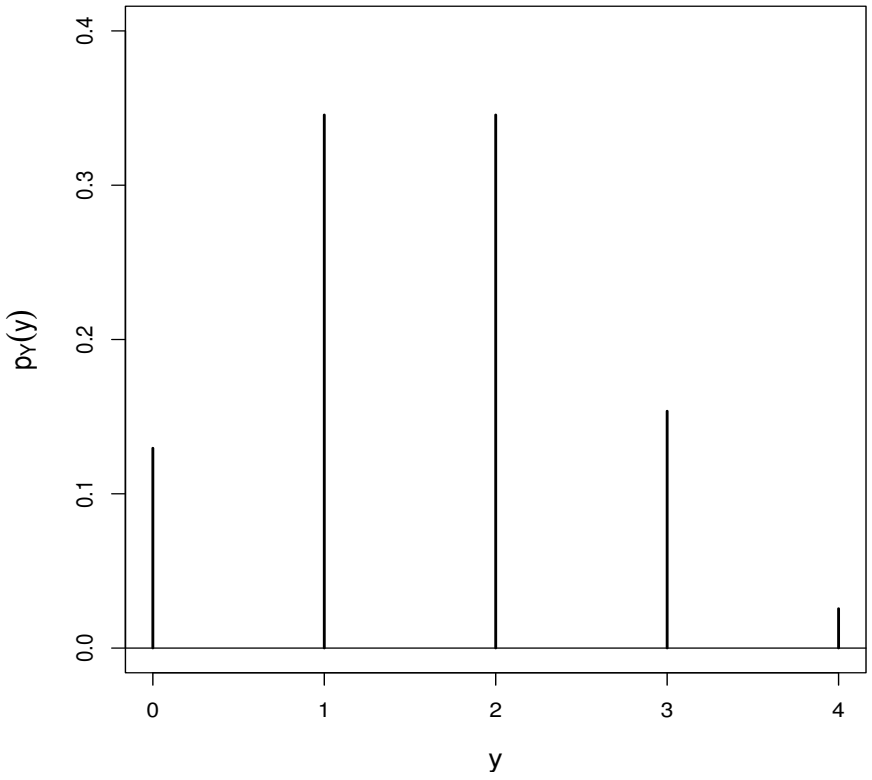

Figure 4.3: Probability mass function of $Y \sim b(4, 0.4)$ in Example 4.18.

The $b(4, 0.4)$ pmf is shown in Figure 4.3 (above). Here are all the pmf calculations:

$$\begin{aligned}
P(Y=0) &= \binom{4}{0}(0.4)^0(0.6)^4 = 0.1296 \\
P(Y=1) &= \binom{4}{1}(0.4)^1(0.6)^3 = 0.3456 \\
P(Y=2) &= \binom{4}{2}(0.4)^2(0.6)^2 = 0.3456 \\
P(Y=3) &= \binom{4}{3}(0.4)^3(0.6)^1 = 0.1536 \\
P(Y=4) &= \binom{4}{4}(0.4)^4(0.6)^0 = 0.0256.
\end{aligned}$$

Putting these in tabular form (like before), we have

y	0	1	2	3	4
$p_Y(y)$	0.1296	0.3456	0.3456	0.1536	0.0256

MEAN/VARIANCE: If $Y \sim b(n, \pi)$, then

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

Example 4.18 (continued). The mean number of plots which respond to treatment is

$$\mu = 4(0.4) = 1.6 \text{ plots.}$$

The variance is

$$\sigma^2 = 4(0.4)(0.6) = 0.96 \text{ (plots)}^2.$$

The standard deviation is

$$\sigma = \sqrt{0.96} \approx 0.98 \text{ plots.}$$

Example 4.19. A computer's power supply unit (PSU) is hardware that converts high-voltage alternating current (from a wall outlet) into lower-voltage direct current power which is needed by a computer's components (e.g., CPU, etc.). A manufacturer claims "no more than 5 percent" of its power supply units need servicing during their warranty period. We can interpret

- PSU = "trial"
- PSU needs service during warranty period = "success"
- $\pi = 0.05$ (in fact, the manufacturer claims π is no larger than 0.05).

Technicians access a sample of 30 units and simulate their usage during the warranty period. If the Bernoulli trial assumptions hold (independent units, same probability of needing service for each unit), then

$$Y = \text{number of PSUs needing service during warranty period} \sim b(n = 30, \pi = 0.05).$$

Q: Among the 30 units tested, what is the probability the technicians find 5 or more PSUs requiring service during the warranty period?
A: We want $P(Y \geq 5)$. We could calculate

$$P(Y = 5) + P(Y = 6) + P(Y = 7) + \cdots + P(Y = 29) + P(Y = 30),$$

but this would require using the binomial pmf formula 26 times and adding up the results.

Easier: It is easier to write

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y \leq 4) \\ &= 1 - [P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4)] \\ &= 1 - P(Y = 0) - P(Y = 1) - P(Y = 2) - P(Y = 3) - P(Y = 4), \end{aligned}$$

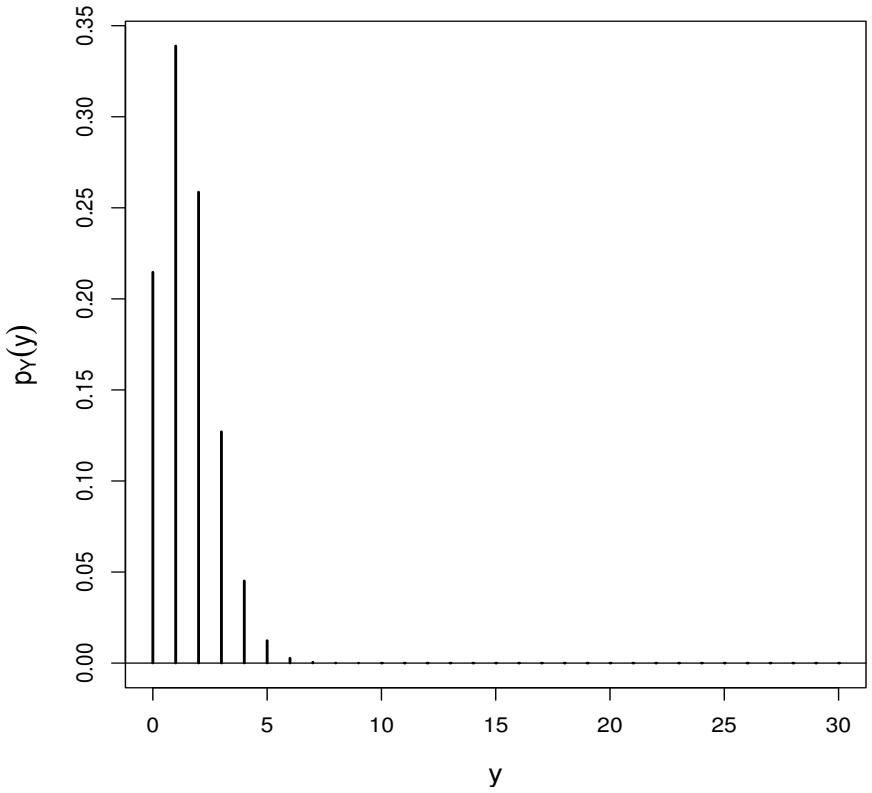

Figure 4.4: Probability mass function of $Y \sim b(30, 0.05)$ in Example 4.19.

which requires using the binomial pmf formula only 5 times. Note that

$$\begin{aligned}
P(Y=0) &= \binom{30}{0}(0.05)^0(0.95)^{30} \approx 0.2146 \\
P(Y=1) &= \binom{30}{1}(0.05)^1(0.95)^{29} \approx 0.3389 \\
P(Y=2) &= \binom{30}{2}(0.05)^2(0.95)^{28} \approx 0.2586 \\
P(Y=3) &= \binom{30}{3}(0.05)^3(0.95)^{27} \approx 0.1270 \\
P(Y=4) &= \binom{30}{4}(0.05)^4(0.95)^{26} \approx 0.0451.
\end{aligned}$$

Therefore,

$$P(Y \geq 5) \approx 1 - 0.2146 - 0.3389 - 0.2586 - 0.1270 - 0.0451 = 0.0158.$$

Under this model, it is unlikely the technicians would find 5 or more PSUs requiring service during the warranty period. This would occur only in 1.58% of the samples of size 30 tested.

BINOMIAL R CODE: Suppose $Y \sim b(n, \pi)$.

$P(Y = y)$	$P(Y \leq y)$
`dbinom(y,n,`π`)`	`pbinom(y,n,`π`)`

For example, $P(Y = 2)$ would be calculated using `dbinom(2,30,0.05)`, and $P(Y \leq 4)$ would be calculated using `pbinom(4,30,0.05)`.

```
> options(digits=4)
> dbinom(2,30,0.05) # P(Y=2)
[1] 0.2586
> pbinom(4,30,0.05) # P(Y<=4)
[1] 0.9844
```

4.8 Poisson distribution

Relevance: The Poisson distribution is the most common discrete distribution when modeling **counts** such as

- the number of customers entering a post office per hour
- the number of insurance claims received per day
- the number of machine breakdowns per month
- the number of severe weather events per year
- the number of raw material defects per square foot
- the number of airborne aerosol particles per cubic inch.

Definition: A **Poisson distribution** arises when we are counting the number of "occurrences" over a unit interval of time (e.g., hour, day, month, year etc.) or a unit region of space (e.g., square mile, cubic inch, etc.). These occurrences must obey the following postulates:

P1. The probability of 2 or more occurrences in a sufficiently small interval of time (or region of space) is zero.

P2. The number of occurrences in non-overlapping intervals of time (or regions of space) are independent.

P3. The mean number of events in one unit interval of time (or region of space), denoted by μ, is the same as the mean number of events in another unit interval of time (or region of space).

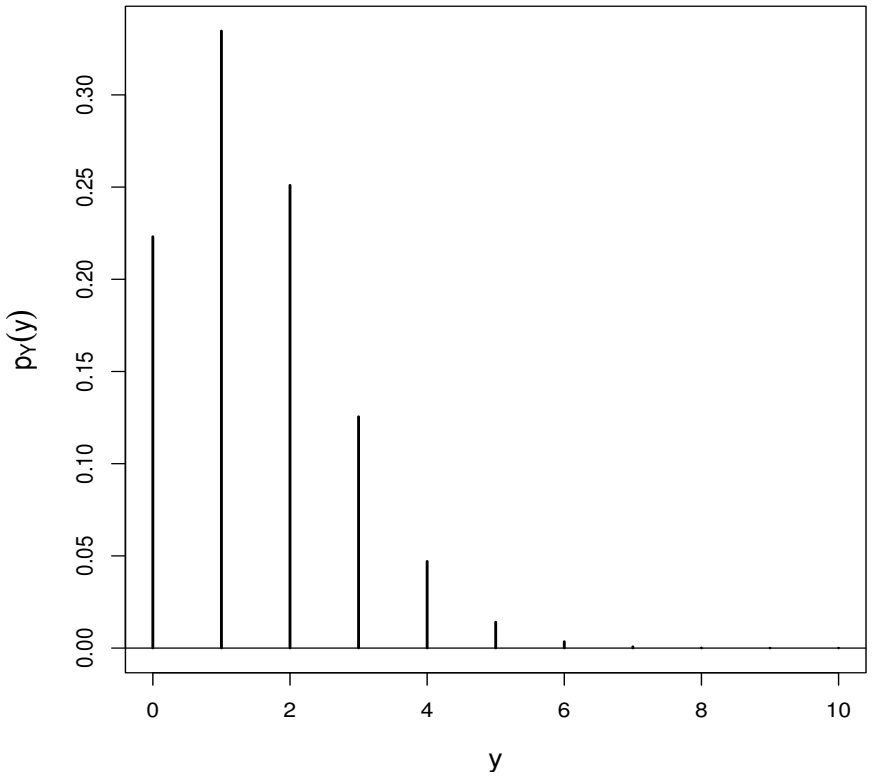

Figure 4.5: Probability mass function of $Y \sim$ Poisson(1.5) in Example 4.20.

Define

$$Y = \text{number of occurrences in a unit interval of time (or region of space).}$$

If the Poisson postulate assumptions hold, then the probability mass function (pmf) of Y is given by the formula

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

We write $Y \sim$ Poisson(μ), where μ is the mean number of occurrences per unit interval of time or region of space.

Example 4.20. 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 $\mu = 1.5$. The pmf of Y is shown in Figure 4.5 (above).

Q: What is the probability there are exactly 2 severe weather events in a given year?

$$P(Y = 2) = \frac{(1.5)^2 e^{-1.5}}{2!} \approx 0.251.$$

Q: What is the probability there are more than 3 severe weather events in a given year?
A: We want $P(Y > 3)$.

$$\begin{aligned} P(Y > 3) &= 1 - P(Y \leq 3) \\ &= 1 - [P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3)] \\ &= 1 - P(Y=0) - P(Y=1) - P(Y=2) - P(Y=3) \\ &= 1 - \frac{(1.5)^0 e^{-1.5}}{0!} - \frac{(1.5)^1 e^{-1.5}}{1!} - \frac{(1.5)^2 e^{-1.5}}{2!} - \frac{(1.5)^3 e^{-1.5}}{3!} \approx 0.066. \end{aligned}$$

MEAN/VARIANCE: If $Y \sim \text{Poisson}(\mu)$, then

$$\begin{aligned} E(Y) &= \mu \\ V(Y) &= \mu. \end{aligned}$$

This is a unique feature of the Poisson distribution−its mean and variance are equal.

POISSON R CODE: Suppose $Y \sim \text{Poisson}(\mu)$.

$P(Y = y)$	$P(Y \leq y)$
`dpois(y,`λ`)`	`ppois(y,`μ`)`

```
> options(digits=3)
> dpois(2,1.5) # P(Y=2)
[1] 0.251
> 1-ppois(3,1.5) # P(Y>3)
[1] 0.0656
```

Comparison: It turns out the binomial and Poisson distributions are sometimes "close" to each other in appearance. In a more mathematical course (like STAT 511), you would see the Poisson distribution arises from looking at the binomial distribution when the number of trials n is large and the success probability π is small (i.e., close to 0). What this means for us is the

$$b(n, \pi) \quad \text{and} \quad \text{Poisson}(\mu = n\pi)$$

distributions can be used interchangeably under these conditions.

Example 4.21. In observing patients administered a new drug in a clinical trial, the number of persons experiencing a particular side effect might be very small. Suppose π, the probability a person experiences a side effect to the drug, is 0.001 and $n = 1000$ patients in the trial received the drug. Find the probability that none of the patients experience the side effect (e.g., damage to a heart valve, etc.) when $\pi = 0.001$.

Let Y denote the number of patients (out of 1000) that will experience the side effect. Under the **binomial** model, $b(n = 1000, \pi = 0.001)$, we have

$$P(Y = 0) = \binom{1000}{0} (0.001)^0 (0.999)^{1000} \approx 0.3676954.$$

Under the **Poisson** model, Poisson($\mu = 1$), we have

$$P(Y = 0) = \frac{1^0 e^{-1}}{0!} = e^{-1} \approx 0.3678794.$$

These answers are very close which illustrates how similar these distributions are when n is large and π is small.

```
> dbinom(0,1000,0.001) # binomial
[1] 0.3676954
> dpois(0,1) # Poisson
[1] 0.3678794
```

4.9 Continuous random variables

Recall: A random variable Y is **continuous** if, at least in theory, it can have any value in an interval of numbers. For example,

- Y = pH of an aqueous solution $\longrightarrow \mathbb{R} = \{-\infty < y < \infty\}$
- Y = BMI of a patient $\longrightarrow \mathbb{R}^+ = \{y > 0\}$
- Y = proportion of students who graduate college in 4 years $\longrightarrow [0, 1] = \{0 \leq y \leq 1\}$
- Y = current (mA) measured in a thin copper wire $\longrightarrow [4.9, 5.1] = \{4.9 \leq y \leq 5.1\}$

Important: Assigning probabilities to events with continuous random variables is different than in the discrete case. We do not assign probability to specific values like $y = 2$ as we might in discrete models. Instead, we assign probabilities to events which are **intervals** like $1 < y < 3$, acknowledging that Y can have any value between 1 and 3.

Terminology: The **probability density function** (pdf) of a continuous random variable Y is a function $f_Y(y)$ which has the following characteristics:

1. $f_Y(y) \geq 0 \longrightarrow f_Y(y)$ is nonnegative.
2. the area under any pdf is equal to 1.

Result: If Y is a continuous random variable with pdf $f_Y(y)$, then

$$P(a < Y < b) = \int_a^b f_Y(y)dy.$$

This gives the **area under the curve** over the interval from a to b. That is, for continuous distributions, probabilities are areas under the pdf.

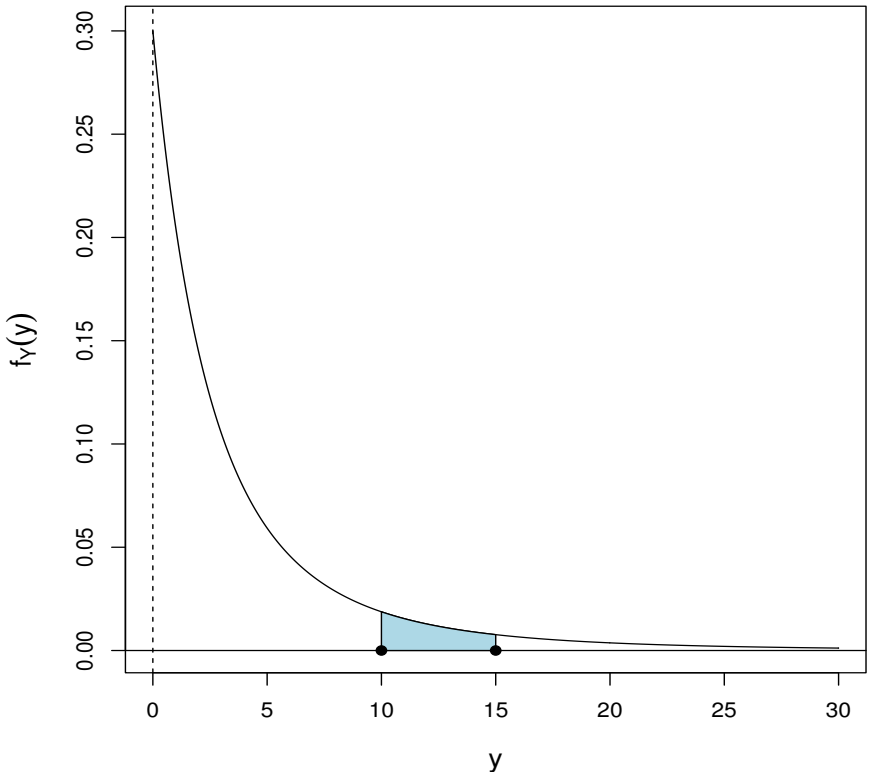

Figure 4.6: Probability density function of Y in Example 4.22. The area under the curve is $P(10 < Y < 15)$.

Example 4.22. The amount of loss/damage (in millions of dollars) due to catastrophic weather is a continuous random variable Y with pdf

$$f_Y(y) = \begin{cases} \dfrac{3000}{(10+y)^4}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.6 (above). Calculus shows

$$P(10 < Y < 15) = \int_{10}^{15} \frac{3000}{(10+y)^4} dx \approx 0.061.$$

Therefore, the probability the amount of loss/damage will be between 10 and 15 million dollars is approximately 0.061.

Example 4.23. Let the continuous random variable Y denote the current measured in a thin copper wire (in milliamperes, mA). Assume the range (support) of Y is $[4.9, 5.1]$ mA and the pdf of Y is

$$f_Y(y) = \begin{cases} 5, & 4.9 \leq y \leq 5.1 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is shown in Figure 4.7 (next page).

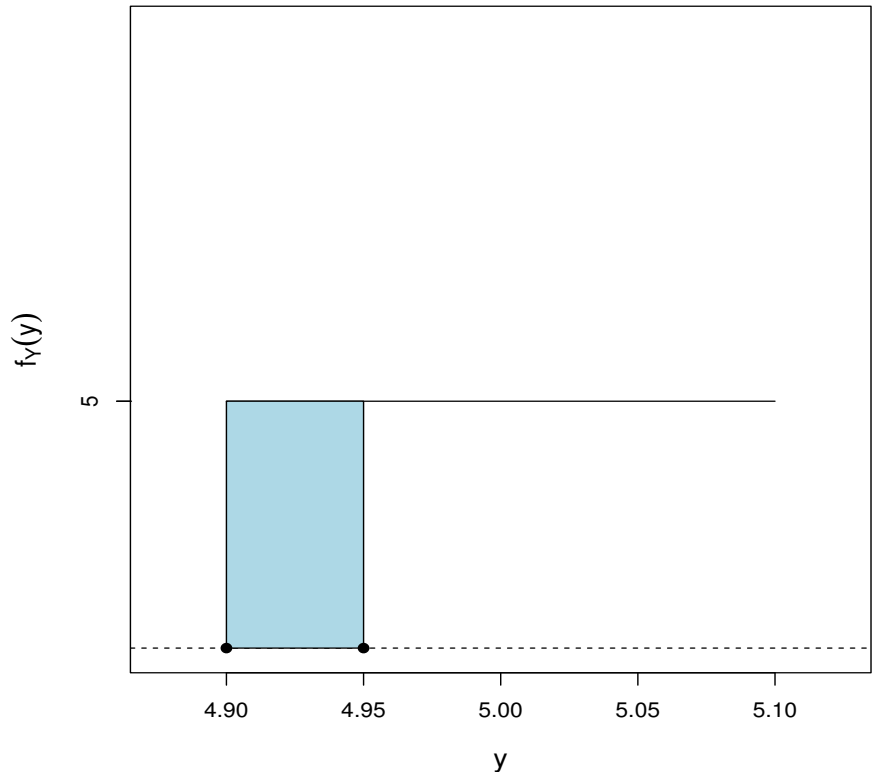

Figure 4.7: Probability density function of Y in Example 4.23. The area under the curve is $P(Y < 4.95)$.

Q: What is the probability the current measurement is less than 4.95 mA?
A: We want

$$P(Y < 4.95) = \int_{4.9}^{4.95} 5 \, dy = 0.25.$$

That is, 25% of all current measurements will be less than 4.95 mA. Note this area can be determined using simple geometry. The area of the rectangle above is simply its base (0.05) times its height (5); i.e., $0.05 \times 5 = 0.25$.

4.10 Normal distribution

Importance: The normal distribution is, by far, the most important probability distribution in statistics. Mathematically, a random variable Y is said to have a normal distribution if its probability density function is given by

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{-(y-\mu)^2/2\sigma^2}, \quad \text{for } -\infty < y < \infty.$$

We write $Y \sim \mathcal{N}(\mu, \sigma^2)$.

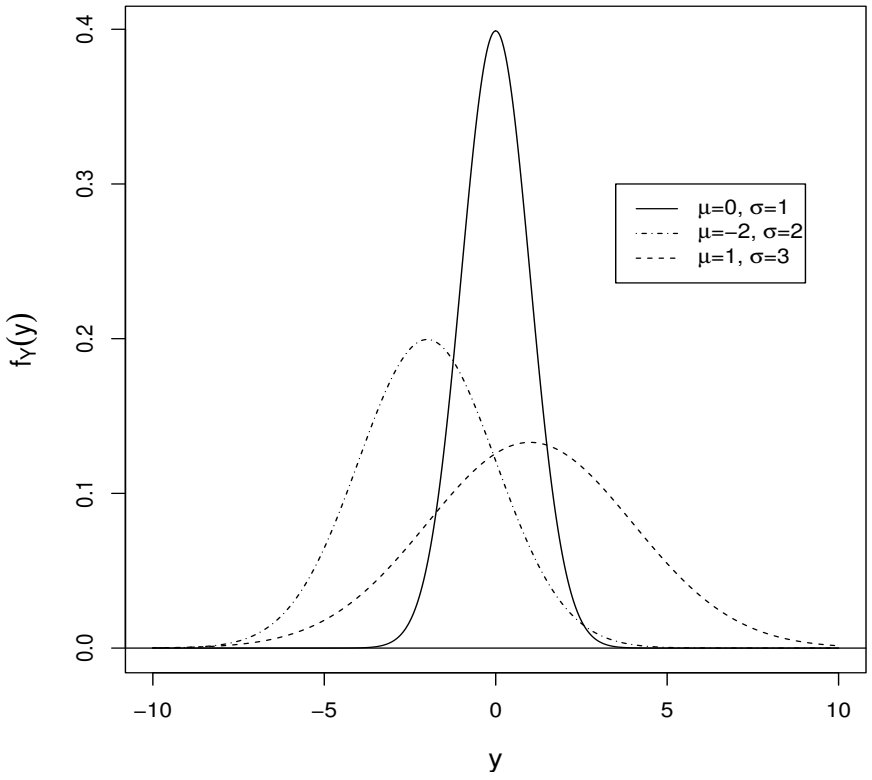

Figure 4.8: Normal pdfs for different values of μ and σ^2. Note the standard deviation σ is used instead of the variance σ^2.

Notes: The parameters μ and σ^2 are the mean and variance of Y, respectively, that is,

$$\begin{aligned} E(Y) &= \mu \\ V(Y) &= \sigma^2. \end{aligned}$$

The mean μ identifies where the "center" of the distribution is. The variance σ^2 measures the "spread" of the distribution.

- The $\mathcal{N}(\mu, \sigma^2)$ pdf is **symmetric** about the mean μ.
- The $\mathcal{N}(\mu, \sigma^2)$ pdf has points of inflection at $\mu - \sigma$ and $\mu + \sigma$.
- The $\mathcal{N}(\mu, \sigma^2)$ pdf follows the **68-95-99.7% Rule**:

$$\begin{aligned} P(\mu - \sigma < Y < \mu + \sigma) &\approx 0.68 \\ P(\mu - 2\sigma < Y < \mu + 2\sigma) &\approx 0.95 \\ P(\mu - 3\sigma < Y < \mu + 3\sigma) &\approx 0.997. \end{aligned}$$

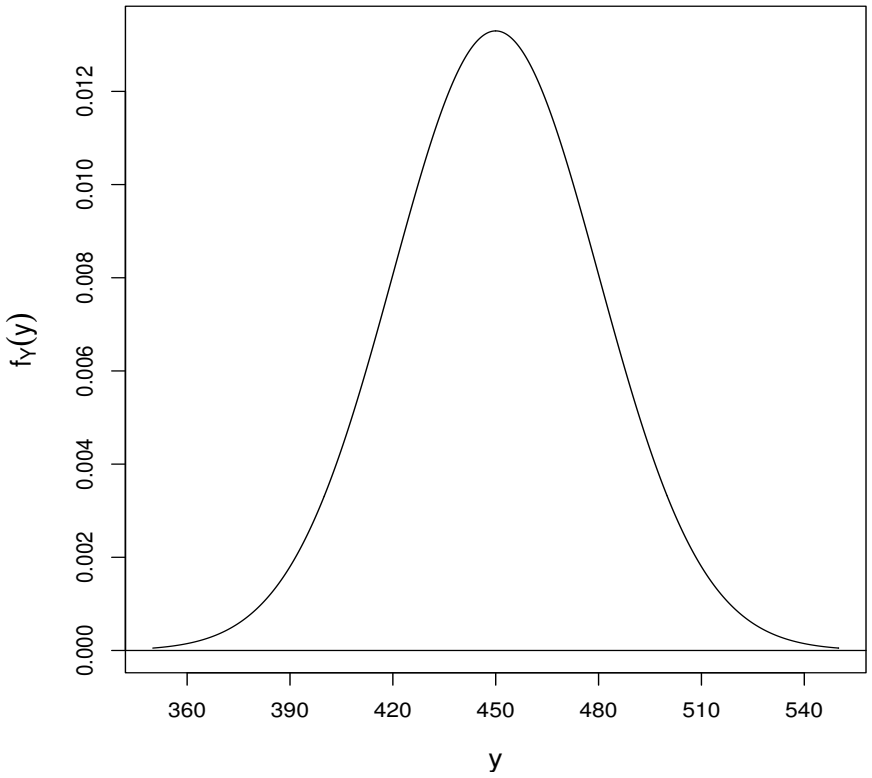

Figure 4.9: Probability density function of $Y \sim \mathcal{N}(450, 30^2)$ in Example 4.24.

That is,

- approximately 68% of the observations will be within 1 standard deviation of the mean
- approximately 95% of the observations will be within 2 standard deviations of the mean
- approximately 99.7% of the observations will be within 3 standard deviations of the mean.

Therefore, it is unlikely for a $\mathcal{N}(\mu, \sigma^2)$ random variable to have a value y further than 3 standard deviations away from its mean in either direction; this probability is approximately 0.003 or 0.3%.

Example 4.24. The daily demand for water use in Atlanta, GA, is a continuous random variable Y, measured in millions of gallons. Suppose the distribution of Y is normal with mean $\mu = 450$ and standard deviation $\sigma = 30$. The $\mathcal{N}(450, 30^2)$ pdf is shown in Figure 4.9 (above).

Q: Form intervals 1, 2, and 3 standard deviations from the mean. Interpret.

A: For reference, refer to the $\mathcal{N}(450, 30^2)$ pdf on the last page:

- Approximately 68% of the days will have a water demand between 420 and 480 million gallons.
- Approximately 95% of the days will have a water demand between 390 and 510 million gallons.
- Approximately 99.7% of the days (or nearly all) will have a water demand between 360 and 540 million gallons.

In addition, under this $\mathcal{N}(450, 30^2)$ model, it would very unlikely to have a day where the demand is below 360 million gallons or above 540 million gallons.

Terminology: A normal random variable with mean 0 and variance 1 is called a **standard normal** random variable. The pdf of $Z \sim \mathcal{N}(0, 1)$ is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \text{for } -\infty < z < \infty.$$

Result: If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Z = \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

This result says any normal random variable Y can be "converted" to a standard normal random variable Z by applying this linear transformation. This conversion is known as **standardization**. For example, suppose exam scores for a population of students have mean $\mu = 70$ and standard deviation $\sigma = 10$. A student's score of 85 has standardized value

$$z = \frac{85 - 70}{10} = 1.5.$$

This means the student's score is 1.5 standard deviations above the mean. Similarly, a student's score of 55 produces

$$z = \frac{55 - 70}{10} = -1.5,$$

meaning the score is 1.5 standard deviations below the mean. From the 68-95-99.7% Rule, we know almost all standardized values z will be between -3 and 3.

Implication: We can find any area under any normal pdf by using the fact above. To see why, note that if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then

$$P(a < Y < b) = P\left(\frac{a - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right).$$

Table 1 (OL, 2016, pp 1086-1087) catalogues standardized values z and **areas to the left** of z. These areas (probabilities) are determined using

`pnorm(z)` or `pnorm(z,0,1)`

in R.

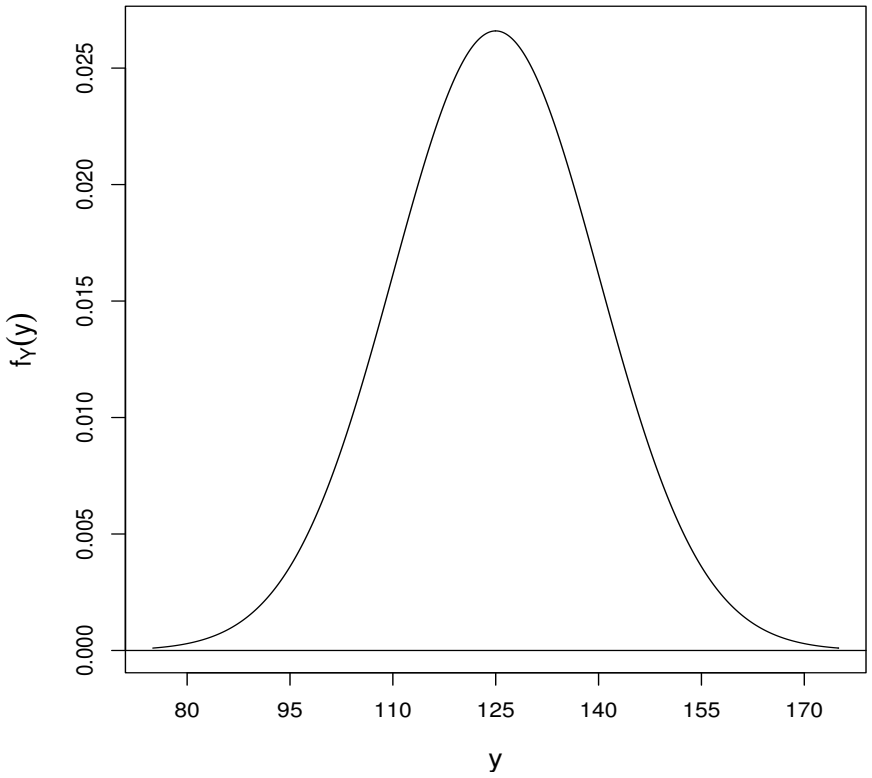

Figure 4.10: $\mathcal{N}(125, 15^2)$ pdf in Example 4.25.

Example 4.25. The World Health Organization uses a normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 15$ to describe the distribution of systolic blood pressure (Y, SBP, measured in mm Hg) of American males aged 18 and over. The $\mathcal{N}(125, 15^2)$ pdf is shown in Figure 4.10 (above).

Q: Find the probability a randomly selected male has SBP **less than** 100 mm Hg.
A: We want

$$P(Y < 100) = P\left(\frac{Y - 125}{15} < \frac{100 - 125}{15}\right) = P\left(Z < \frac{100 - 125}{15}\right) \approx P(Z < -1.67).$$

From R, we have

```
> pnorm(-1.67) # area to the left
[1] 0.0475
```

Interpretation: Approximately 4.75% of the population of all American males (aged 18 and over) have SBP less than 100 mm Hg.

Q: Find the probability a randomly selected male has SBP **greater than** 140 mm Hg. This is the threshold for being classified as "high SBP."

A: We want

$$P(Y > 140) = P\left(\frac{Y-125}{15} > \frac{140-125}{15}\right) = P\left(Z > \frac{140-125}{15}\right) \approx P(Z > 1).$$

From R, we have

```
> 1-pnorm(1) # area to the right
[1] 0.1587
```

Therefore, approximately 15.87% of the population of all American males (aged 18 and over) has "high SBP."

Q: Find the probability a randomly selected male has SBP **between** 100 and 140 mm Hg.
A: We want

$$P(100 < Y < 140) = P\left(\frac{100-125}{15} < \frac{Y-125}{15} < \frac{140-125}{15}\right) \approx P(-1.67 < Z < 1).$$

From R, we have

```
> pnorm(1)-pnorm(-1.67) # area between
[1] 0.7939
```

Therefore, approximately 79.39% of the population of all American males (aged 18 and over) has SBP between 100 and 140 mm Hg.

Finding percentiles: For $0 < p < 1$, the $100p$**th percentile** of a distribution is the value y_p such that

- $100p\%$ of the area is to the left of y_p, that is, $P(Y < y_p) = p$
- $100(1-p)\%$ of the area is to the right of y_p, that is, $P(Y > y_p) = 1 - p$.

Figure 4.11 (next page) makes this definition clear. The most common percentile is the **50th percentile**. This is called the **median** of the distribution. That is,

- the area to the left of the median $y_{0.50}$ is 0.5. The area to the right of the median $y_{0.50}$ is 0.5.
- In any $\mathcal{N}(\mu, \sigma^2)$ distribution, the median and the mean (μ) are the same. This is because the normal distribution is **symmetric**.

Other important percentiles are 25th and 75th percentiles:

- $y_{0.25} \longrightarrow$ **1st quartile** $\longrightarrow P(Y < y_{0.25}) = 0.25$
- $y_{0.75} \longrightarrow$ **3rd quartile** $\longrightarrow P(Y < y_{0.75}) = 0.75$.

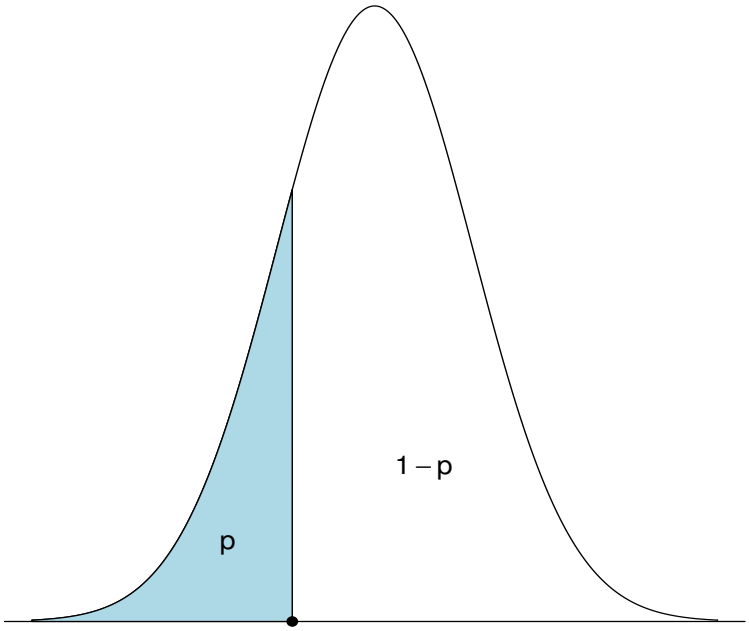

Figure 4.11: Normal pdf. The 100pth percentile y_p is shown using a dark circle.

In general, note that by standardizing, we have

$$p = P(Y < y_p) = P\left(\frac{Y-\mu}{\sigma} < \frac{y_p-\mu}{\sigma}\right) = P\left(Z < \frac{y_p-\mu}{\sigma}\right).$$

This means that

$$z_p = \frac{y_p-\mu}{\sigma} \quad \Longrightarrow \quad y_p = \mu + z_p\sigma,$$

where z_p is the 100pth percentile of a standard normal distribution $\mathcal{N}(0,1)$. Standard normal percentiles z_p are found using

`qnorm(p)` or `qnorm(p,0,1)`

in R. Table 1 (OL, 2016, pp 1086-1087) can also be used, although approximation may be needed if an exact area (e.g., 0.2) is not shown in the table.

Example 4.26. For a population of runners, the time it takes to complete the New York Marathon (Y, in minutes) is normally distributed with mean $\mu = 250$ and standard deviation $\sigma = 40$. The $\mathcal{N}(250, 40^2)$ pdf is shown in Figure 4.12 (next page).

Q: The fastest 10% of the runners will complete the marathon in how long?

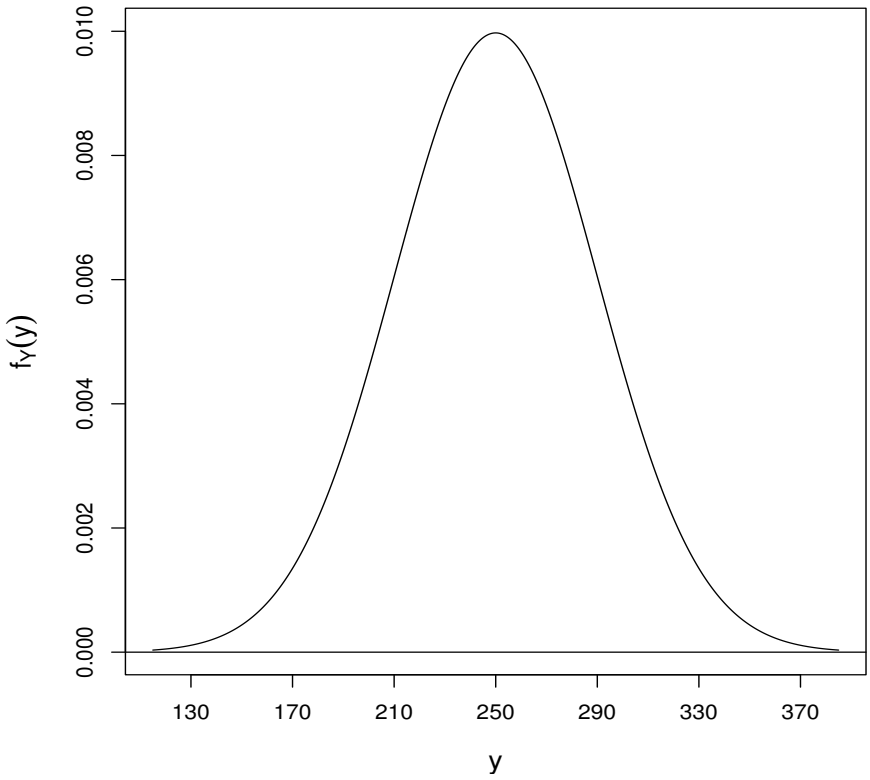

Figure 4.12: $\mathcal{N}(250, 40^2)$ pdf in Example 4.27.

A: We want to find the 10th percentile $y_{0.10}$ of the $\mathcal{N}(250, 40^2)$ distribution. We can begin by finding $z_{0.10}$, the 10th percentile of the $\mathcal{N}(0, 1)$ distribution. From Table 1 (OL, 2016, pp 1086-1087) or R, we have

```
> options(digits=3)
> qnorm(0.10) # z_{0.10}, 10th percentile of N(0,1)
[1] -1.28
```

so that

$$y_{0.10} = \mu + z_{0.10}\sigma = 250 - 1.28(40) = 198.8 \text{ minutes}.$$

Remark: We have learned how to calculate probabilities and percentiles for any normal distribution by first standardizing.

- Doing this allows us to use the standard normal distribution which is catalogued in Table 1 (OL, 2016, pp 1086-1087).
- Ultimately, this is not needed as R's functions `pnorm` and `qnorm` allow you to calculate probabilities and percentiles for any $\mathcal{N}(\mu, \sigma^2)$ distribution!

NORMAL R CODE: Suppose $Y \sim \mathcal{N}(\mu, \sigma^2)$.

$P(Y < y)$	y_p
`pnorm(y,`μ`,`σ`)`	`qnorm(p,`μ`,`σ`)`

Note that the `pnorm` and `qnorm` functions in R are written in terms of the standard deviation σ and not the variance σ^2.

4.11 Sampling distributions and the Central Limit Theorem

Scenario: Consider the last two examples and the populations that are represented:

- Example 4.25. Population: all American males aged 18 and over
- Example 4.26. Population: all marathon runners.

The notion of a **sampling distribution** arises when we observe a sample from a population; e.g., sample 100 American males, sample 25 marathon runners, and we calculate the value of a statistic from our sample. Common examples of statistics are

- the sample mean $\overline{y}$
- the sample variance s^2 (or sample standard deviation s).

Notation: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample of measurements from a population of individuals.

Example 4.27. I have 1000 pennies manufactured between 1985 and 2025. This is the population (see Figure 4.13; next page, upper left). I sample $n = 5$ pennies and record the ages of the pennies:

```
> sample(pennies,5,replace=F)
[1] 17  2 40 26 39
```

The sample mean is $\overline{y} = 24.8$ years and the sample variance is $s^2 = 253.7$ (years)^2. Now, suppose I return these pennies to the population and observe a new random sample:

```
> sample(pennies,5,replace=F)
[1] 25 27 14 16 40
```

For this sample, the sample mean is $\overline{y} = 24.4$ years and the sample variance is $s^2 = 107.3$ (years)^2.

Key point: Statistics' values change from sample to sample! The **sampling distribution** of a statistic is a distribution that describes how the statistic varies from sample to sample.

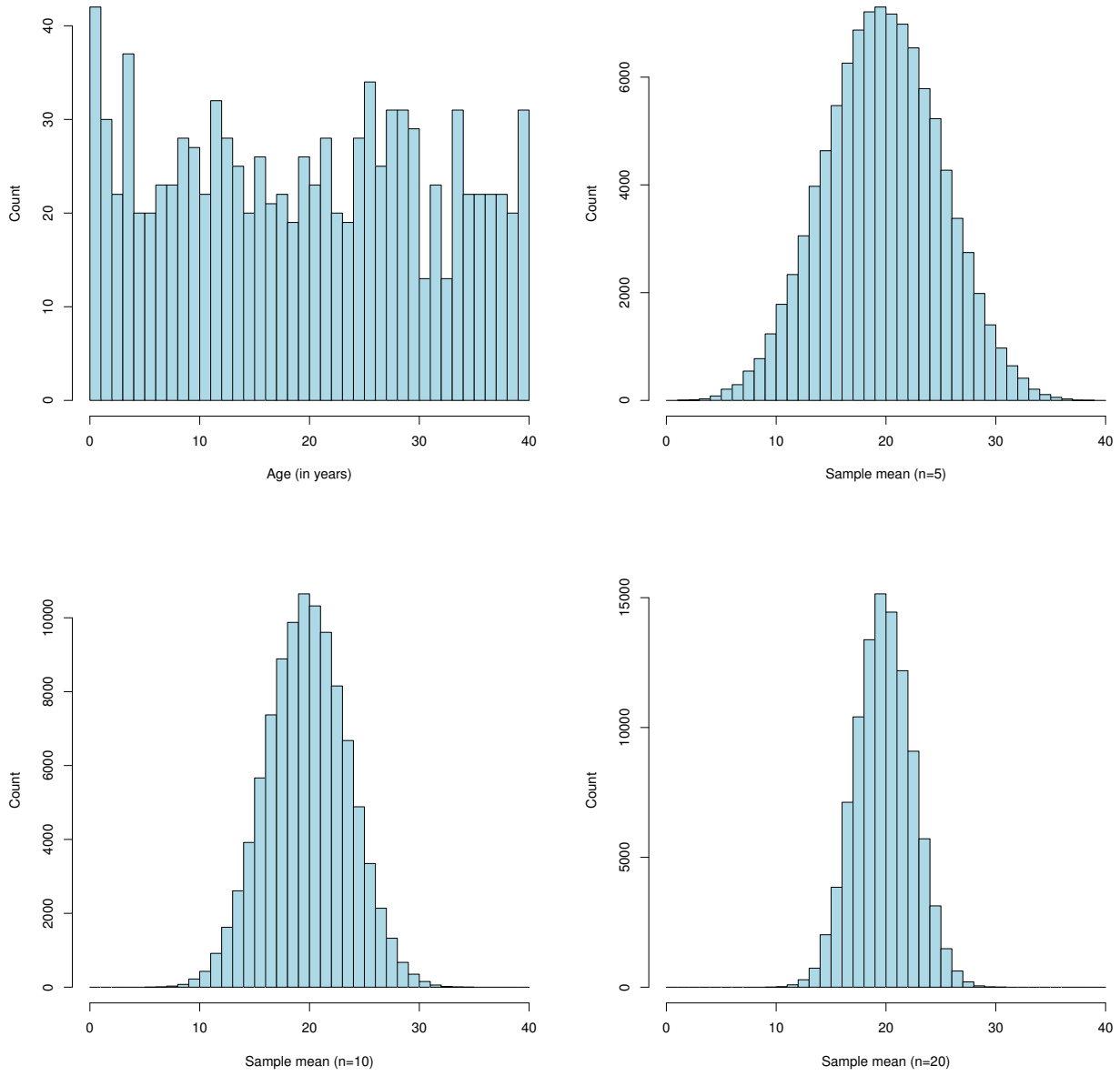

Figure 4.13: Top left: Population distribution of ages of 1000 pennies. Top right: Sampling distribution of $\overline{y}$ when $n = 5$. Bottom left: Sampling distribution of $\overline{y}$ when $n = 10$. Bottom right: Sampling distribution of $\overline{y}$ when $n = 20$. **Note:** Sampling distributions were created by simulating $B = 100,000$ random samples from the population, calculating the value of $\overline{y}$ for each sample, and then displaying the distribution of the $100,000$ sample means in a histogram.

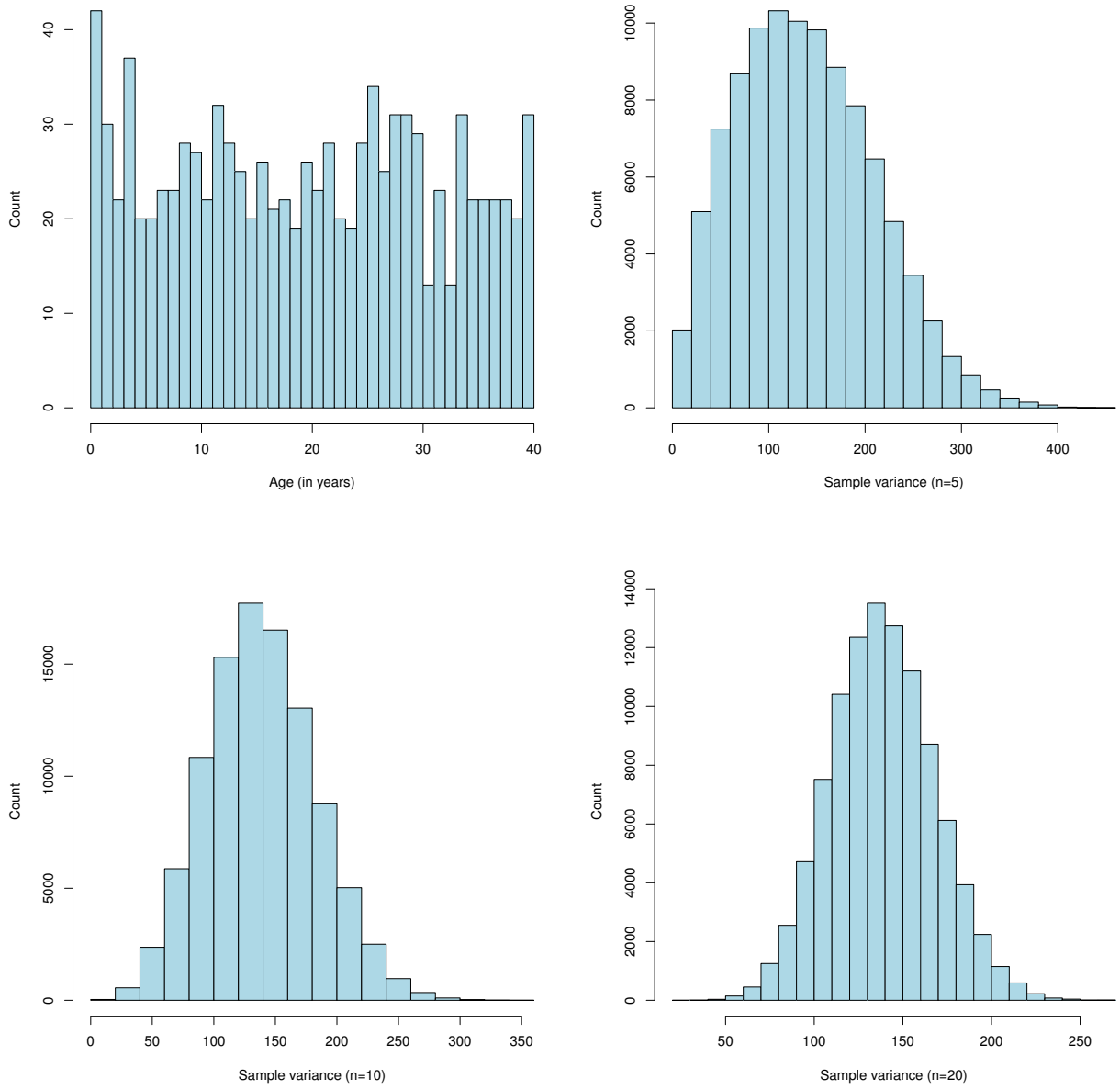

Figure 4.14: Top left: Population distribution of ages of 1000 pennies. Top right: Sampling distribution of s^2 when $n = 5$. Bottom left: Sampling distribution of s^2 when $n = 10$. Bottom right: Sampling distribution of s^2 when $n = 20$. **Note:** Sampling distributions were created by simulating $B = 100,000$ random samples from the population, calculating the value of s^2 for each sample, and then displaying the distribution of the $100,000$ sample variances in a histogram.

Result 1: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample from a population with mean μ and variance σ^2. Let $\overline{Y}$ denote the sample mean. We have the following results:

$$\mu_{\overline{Y}} = \mu \quad \text{and} \quad \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

The first result says the sampling distribution of the sample mean $\overline{Y}$ has mean equal to the population mean μ. The second result says what the standard deviation of the sampling distribution of $\overline{Y}$ is equal to the population standard deviation σ divided by the square root of the sample size n. We call

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

the **standard error** of $\overline{Y}$. It is the standard deviation of the sampling distribution of $\overline{Y}$.

Exercise: I used R to calculate the mean and the standard deviation for the population distribution of the ages of pennies:

$$\begin{aligned} \mu &= 19.84 \\ \sigma &= 11.77. \end{aligned}$$

Find the mean and standard deviation (standard error) of the sample mean $\overline{Y}$ when $n = 5$, $n = 10$, and $n = 20$. Refer to Figure 4.13.

$\underline{n = 5}$:

$$\mu_{\overline{Y}} = \qquad \sigma_{\overline{Y}} =$$

$\underline{n = 10}$:

$$\mu_{\overline{Y}} = \qquad \sigma_{\overline{Y}} =$$

$\underline{n = 20}$:

$$\mu_{\overline{Y}} = \qquad \sigma_{\overline{Y}} =$$

Result 2: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution. The sample mean $\overline{Y}$ has the following sampling distribution:

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- Our first observation here is

  normal population distribution $\implies$ sampling distribution of $\overline{Y}$ is also normal.

- This result also reminds us that

  $$\mu_{\overline{Y}} = \mu,$$

  that is, the sample mean $\overline{Y}$ is an **unbiased estimator** of the population mean μ.

- This result shows the standard error of $\overline{Y}$ is

$$\sigma_{\overline{Y}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

  This reveals the variability in the possible values of $\overline{Y}$ depends on

  - the population standard deviation σ (for individuals in the population)
  - the sample size n.

Implication: Larger samples reduce the variability of $\overline{Y}$. This leads to more precise estimates when using $\overline{Y}$ as an estimate of μ. This is also true when the population distribution is non-normal (Result 3; coming up).

Example 4.28. In cardiology, an ejection fraction (EF) measures your heart's ability to pump oxygen-rich blood out to your body. This is measured as a percentage and quantifies the amount of blood pumped out of the left ventricle when your heart contracts. Suppose for a population of healthy male subjects, the ejection fraction Y is normally distributed with mean $\mu = 56$ and standard deviation $\sigma = 8$.

Q: What is the population distribution?
A: $Y \sim \mathcal{N}(56, 8^2)$. This is the distribution of EF for all male subjects in the population. See Figure 4.15 (next page).

Q: A random sample of $n = 16$ males is selected from the population and the EF is measured on each subject producing $Y_1, Y_2, ..., Y_{16}$. What is the sampling distribution of $\overline{Y}$, the sample mean EF?
A: Use Result 2:

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \overline{Y} \sim \mathcal{N}\left(56, \frac{64}{16}\right) \implies \overline{Y} \sim \mathcal{N}(56, 4).$$

The sample mean $\overline{Y}$ is normally distributed with mean 56 and variance 4.

Q: In the last part, what is the standard error of $\overline{Y}$?
A: The standard error of $\overline{Y}$ is the standard deviation of its sampling distribution, here,

$$\sigma_{\overline{Y}} = \sqrt{4} = 2.$$

Note that this is also

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{16}} = 2$$

using the formula above.

Q: Calculate $P(Y > 60)$ and $P(\overline{Y} > 60)$ and explain what these mean.

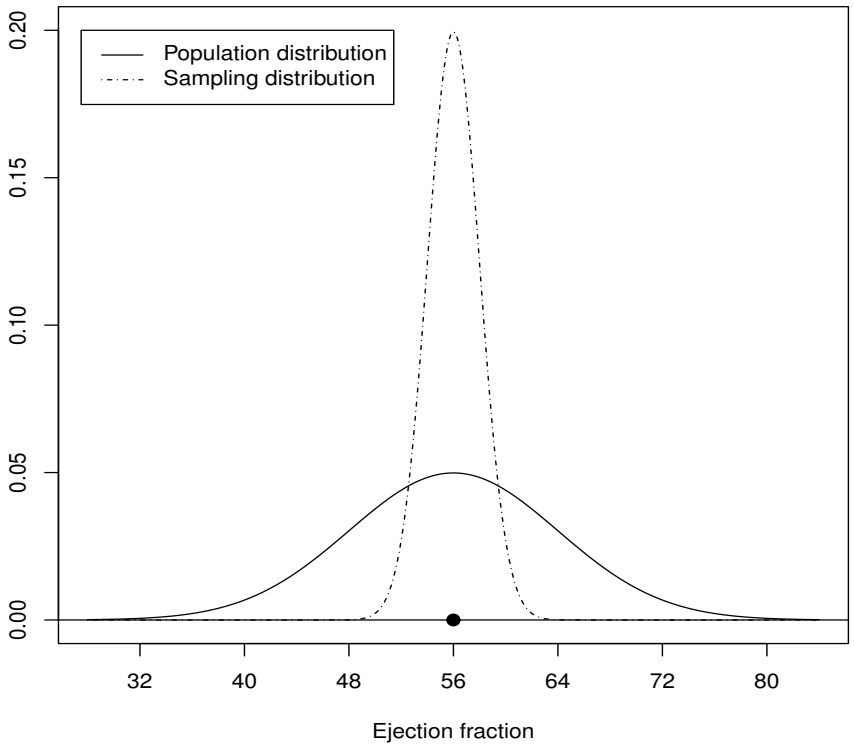

Figure 4.15: $\mathcal{N}(56, 64)$ pdf for Y in Example 4.28. This serves as a model for the population of all healthy male subjects. The sampling distribution of $\overline{Y}$ with $n = 16$ is also shown.

A: We want

$$P(Y > 60) = P\left(\frac{Y - 56}{8} > \frac{60 - 56}{8}\right) = P(Z > 0.5) \approx 0.3085.$$

This means about 30.9% of the population of male subjects will have EF larger than 60.

```
> 1-pnorm(0.5) # area to the right
[1] 0.3085
```

Also,

$$P(\overline{Y} > 60) = P\left(\frac{\overline{Y} - 56}{8/\sqrt{16}} > \frac{60 - 56}{8/\sqrt{16}}\right) = P(Z > 2) \approx 0.0228.$$

If we observed a random sample of $n = 16$ male subjects from this population, the probability the sample mean ejection fraction $\overline{Y}$ would be larger than 60 is about 0.0228.

```
> 1-pnorm(2) # area to the right
[1] 0.0228
```

Recall: Result 2 informed us that when $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, the sample mean

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Q: What if the population distribution is something else other than a normal distribution? What is the sampling distribution of $\overline{Y}$ in this case?
A: We answer this question now, stating one of the most fascinating results in statistics.

Result 3: Suppose $Y_1, Y_2, ..., Y_n$ is a random sample from a population distribution with mean μ and variance σ^2. When the sample size n is large, the sample mean

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right).$$

The symbol $\mathcal{AN}$ is read "approximately normal." This result is called the **Central Limit Theorem (CLT)**.

Remark: The CLT implies the sample mean $\overline{Y}$ will behave like a normal random variable (approximately) regardless of what the population distribution looks like. The population distribution could be skewed, bimodal, discrete, binary, or whatever. The only mathematical requirement is that the population variance $\sigma^2 < \infty$, which holds for nearly all probability distributions.

Example 4.29. In a textile production process, the number of defects (per square meter) in a certain fabric is assumed to follow a Poisson distribution with mean $\lambda = 1.5$. A quality control plan involves sampling 50 square meter pieces per day and inspecting them for defects.

Q: What is the population distribution?
A: Define

$$Y = \text{number of defects per square meter of fabric.}$$

The population distribution is Poisson($\lambda = 1.5$). This is the distribution of the number of defects Y for each square meter piece of fabric in the population.

Q: What is the sampling distribution of $\overline{Y}$, the average number of defects for the 50 pieces observed on a given day?
A: The population mean is $\mu = 1.5$ and the population variance is $\sigma^2 = 1.5$. Recall that in the Poisson distribution, the mean and variance are equal. Now, use Result 3 (CLT):

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \implies \overline{Y} \sim \mathcal{AN}\left(1.5, \frac{1.5}{50}\right) \implies \overline{Y} \sim \mathcal{AN}(1.5, 0.03).$$

Q: Assuming the population distribution is correct, how likely would it be to observe a sample mean $\overline{Y}$ larger than 2 as part of the daily quality control plan?

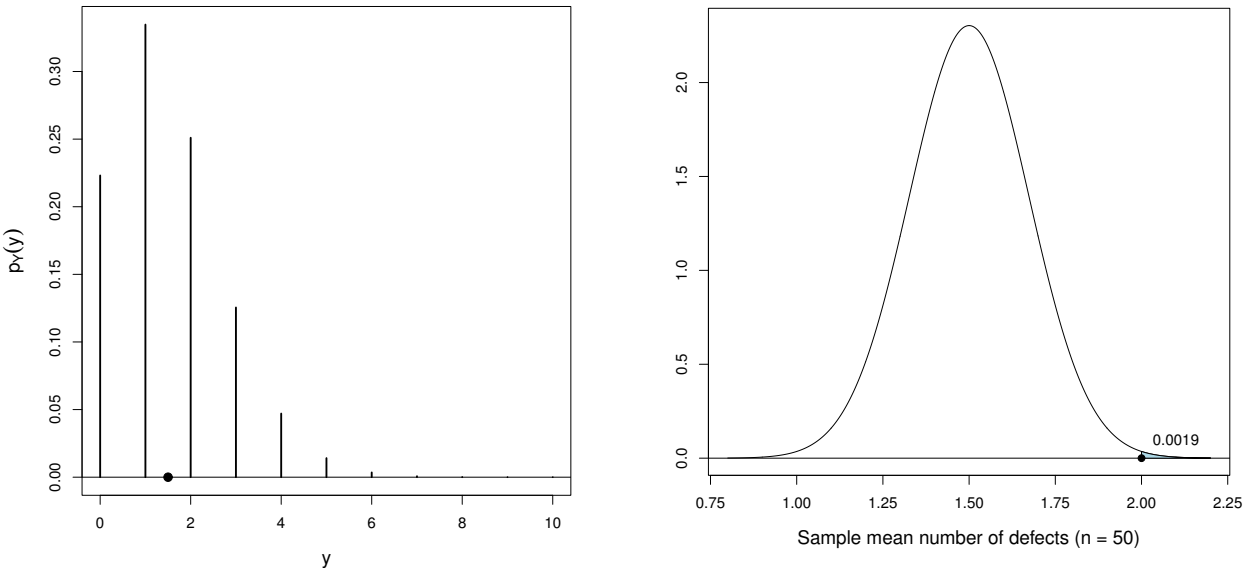

Figure 4.16: Left: Population distribution of $Y \sim \text{Poisson}(\lambda = 1.5)$. The population mean $\mu = 1.5$ is shown using a solid circle. Right: Approximate sampling distribution of $\overline{Y}$, the sample mean of $n = 50$ observations from the population.

A: We can calculate

$$P(\overline{Y} > 2) = P\left(\frac{\overline{Y} - 1.5}{\sqrt{1.5/50}} > \frac{2 - 1.5}{\sqrt{1.5/50}}\right) \approx P(Z > 2.89) \approx 0.0019.$$

```
> 1-pnorm(2.89) # area to the right
[1] 0.0019
```

Question for thought: Because $P(\overline{Y} > 2) \approx 0.0019$ is so small, what might be true if we actually observed a sample mean $\overline{Y}$ larger than 2 on any given day? This would not be expected if the population mean $\mu = 1.5$ was correct.

Back to the CLT: Because the CLT only approximates the sampling distribution of $\overline{Y}$, that is,

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right),$$

it is natural to wonder how good the approximation actually is. This depends primarily on two factors:

1. the sample size n. The larger the sample size, the better the approximation.
2. the amount of skewness in the population distribution. The closer the population distribution is to being **symmetric**, the better the approximation.

There is no "one sample size n that fits all" to ensure the CLT will offer a good approximation (although some textbooks claim $n \geq 30$ is the magic threshold). For population distributions which are symmetric or approximately symmetric, the sample size n doesn't have to be that large. For severely skewed distributions or distributions with other nonstandard shapes, the sample size might have to be larger.

Misinterpretation: Some students will interpret the CLT as "with a large enough sample, the data should look approximately normal." This is not correct. When you are sampling from a population, the population distribution remains fixed−it doesn't change. The fact that you have a larger sample simply means that you have more observations from the population. The CLT is a statement about the sampling distribution of the sample mean $\overline{Y}$; not the shape of the population from which you are sampling. The population doesn't "become more normal" when you have larger samples.

Summary: Results 1-3 are sampling distributional results for the sample mean $\overline{Y}$. These are summarized here:

1. Mean and standard error:

$$\mu_{\overline{Y}} = \mu \quad \text{and} \quad \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

2. If $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, then

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

3. **Central Limit Theorem:** If $Y_1, Y_2, ..., Y_n$ is a random sample from a population with mean μ and variance σ^2, then

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right),$$

when the sample size n is large.

Note: We can restate all of these results in terms of the sample sum $\sum_{i=1}^{n} Y_i$. Note that

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \implies \sum_{i=1}^{n} Y_i = n\overline{Y}.$$

That is, the sample sum $\sum_{i=1}^{n} Y_i$ and the sample mean $\overline{Y}$ are **linear functions** of each other. Mathematics shows we have the following distributional results for the sample sum:

1. Mean and standard error:

$$\mu_{\sum Y_i} = n\mu \quad \text{and} \quad \sigma_{\sum Y_i} = \sqrt{n}\sigma$$

2. If $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population distribution, then

$$\sum_{i=1}^{n} Y_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right).$$

3. **Central Limit Theorem:** If $Y_1, Y_2, ..., Y_n$ is a random sample from a population with mean μ and variance σ^2, then

$$\sum_{i=1}^{n} Y_i \sim \mathcal{AN}\left(n\mu, n\sigma^2\right),$$

when the sample size n is large.

Example 4.30. A clinical trial is being planned to test the effectiveness of semaglutide for weight loss in pre-diabetic patients. The time Y (in days) to enroll a patient from this population into the trial has a population distribution with mean $\mu = 20$ days and variance $\sigma^2 = 400$ (days)^2. The goal is recruit $n = 40$ patients.

Q: What is the approximate sampling distribution of $\sum_{i=1}^{40} Y_i$, the total time it will take to recruit 40 patients?
A: We first calculate

$$\mu_{\sum Y_i} = n\mu = 40(20) = 800$$

and

$$\sigma^2_{\sum Y_i} = n\sigma^2 = 40(400) = 16000.$$

The Central Limit Theorem (for sample sums) says

$$\sum_{i=1}^{40} Y_i \sim \mathcal{AN}\left(800, 16000\right).$$

Q: What is the probability it will take less than 600 hours to recruit the 40 patients?
A: We want

$$P\left(\sum_{i=1}^{40} Y_i < 600\right) = P\left(\frac{\sum_{i=1}^{40} Y_i - 800}{\sqrt{16000}} < \frac{600 - 800}{\sqrt{16000}}\right) \approx P(Z < -1.58) \approx 0.0571.$$

```
> pnorm(-1.58) # area to the left
[1] 0.0571
```

4.12 Normal approximation to the binomial

Interesting: We can use a (continuous) normal distribution to approximate a (discrete) binomial distribution. This may sound unrealistic, but it actually not surprising given what we have just learned about the CLT approximation for sample sums.

Recall: A **binomial distribution** arises when we observe a fixed number of Bernoulli trials:

$$\begin{aligned} n &= \text{number of trials} \\ Y &= \text{number of successes (out of } n\text{)}. \end{aligned}$$

We write $Y \sim b(n, \pi)$, where π is the probability of success on any one trial.

Observation: Note that the number of successes Y is really a sample sum. To see why, define

$$Y_i = \begin{cases} 1, & \text{ith trial is a "success"} \\ 0, & \text{ith trial is a "failure."} \end{cases}$$

Then the number of successes Y can be written as

$$Y = Y_1 + Y_2 + \cdots + Y_n = \sum_{i=1}^{n} Y_i,$$

a sample sum of zeros and ones. Applying the CLT to the sample sum, we have

$$Y \sim \mathcal{AN}(n\pi, n\pi(1-\pi)),$$

provided that n is large. In other words, the

$$b(n, \pi) \quad \text{and} \quad \mathcal{N}(n\pi, n\pi(1-\pi))$$

distributions should be "close" to each other when the number of trials (sample size) n is large. In Chapter 10, we will learn the **sample proportion**

$$\widehat{\pi} = \frac{Y}{n} \sim \mathcal{AN}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

under the same condition. This result is commonly used to perform statistical inference for π when it is unknown.

Example 4.31. Suppose a random sample of $n = 1000$ voters is polled to determine sentiment toward the consolidation of city and county government. Suppose that, in truth, 50% of the population of voters is in favor of consolidation so that $\pi = 0.5$.

Q: Calculate the probability of observing 460 or fewer favoring consolidation in the sample of 1,000 voters. Do this under the binomial model $b(1000, 0.5)$ and by using a normal approximation.
A: Let Y denote the number of voters favoring consolidation so that $Y \sim b(1000, 0.5)$. The probability we want is

$$P(Y \leq 460) = P(Y = 0) + P(Y = 1) + P(Y = 2) + \cdots + P(Y = 460).$$

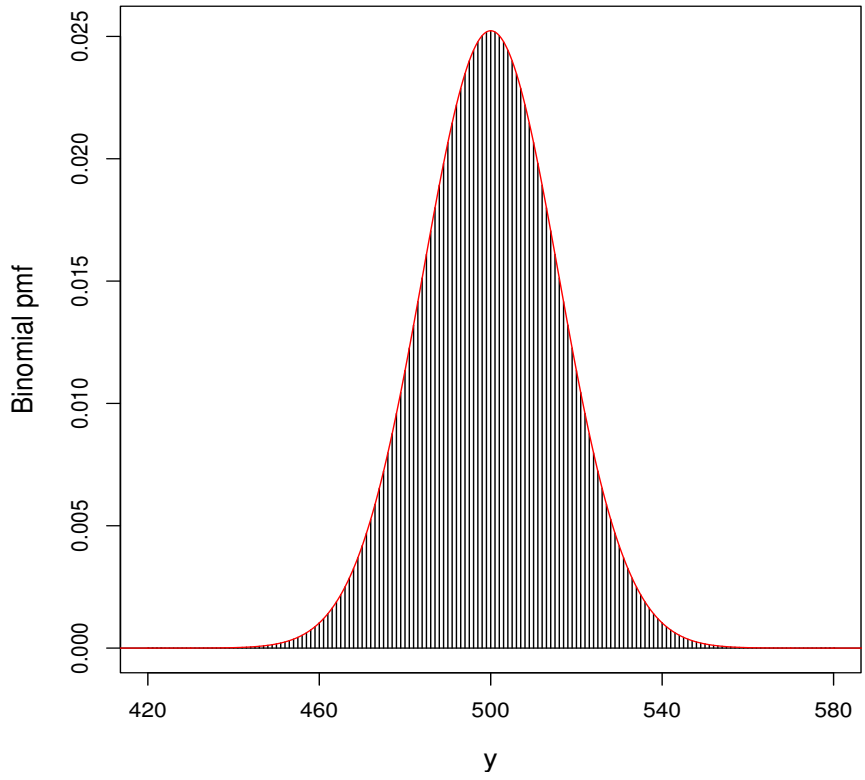

Figure 4.17: $b(1000, 0.5)$ pmf with a normal pdf approximation.

Each probability $P(Y = y)$ would be calculated using the $b(1000, 0.5)$ pmf and we would add up the results. This is what the `pbinom` function in R does automatically, that is,

```
> pbinom(460,1000,0.5) # binomial (exact)
[1] 0.006222
```

On the other hand, a normal approximation for Y has mean

$$\mu = n\pi = 1000(0.5) = 500$$

and variance

$$\sigma^2 = n\pi(1-\pi) = 1000(0.5)(1-0.5) = 250.$$

That is, $Y \sim \mathcal{AN}(500, 250)$. Using this distribution, we have

$$P(Y \leq 460) = P\left(\frac{Y-500}{\sqrt{250}} \leq \frac{460-500}{\sqrt{250}}\right) \approx P(Z \leq -2.53).$$

```
> pnorm(-2.53) # normal (approximation)
[1] 0.005703
```

The two answers (exact and approximation) are very close as expected.

Discussion: It's important to remember that the normal approximation to the binomial is only an approximation (it's actually just the CLT). The approximation is best when

- the number of trials (sample size) n is large
- π is closer to 0.5.

The authors recommend only using this approximation when $n\pi$ and $n(1-\pi)$ are at least 20. Otherwise, the Central Limit Theorem, which is responsible for the approximation, may be inadequate. In other words, probability calculations under the binomial and normal models may be quite different. If you really want to use a normal approximation when $n\pi$ and $n(1-\pi)$ are not large, a possible remedy is to use a **continuity correction**. This correction involves adding or subtracting 1/2 in the probability you're calculating. For example, writing

$$P(Y \leq 15) = P(Y < 15.5) \quad \text{or} \quad P(Y \geq 15) = P(Y > 14.5).$$

See Example 4.26 (OL, 2016, pp 203).

4.13 Normal quantile-quantile plots

Importance: We will often assume measurements of a continuous random variable Y follow a normal distribution. That is, a normal distribution is the true **population distribution** producing the data we see. How do we know if this assumption is reasonable?

- Because we are making an assumption about the distribution of all individuals in the population, we never get to know for sure if the distribution we have chosen is correct.
- We *can* assess if the distribution we have chosen is "reasonable" based on the observed data in the sample.
- This is part of the "model diagnostics" phase of any data analysis. We are assessing (or diagnosing) the plausibility of the assumptions made as part of the analysis.

Example 4.32. Aluminum-lithium alloys are primarily used in the aerospace industry due to their high strength-to-weight ratio and stiffness. The data below are the compressive strengths (in psi) of $n = 80$ specimens of a new alloy undergoing evaluation as a possible material for aircraft structural elements.

105	221	183	186	121	181	180	143	97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110	163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123	134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169	199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135	196	201	200	176	150	170	118	149

Population: all alloy specimens (of this type) produced using the current process
Sample: the 80 specimens.

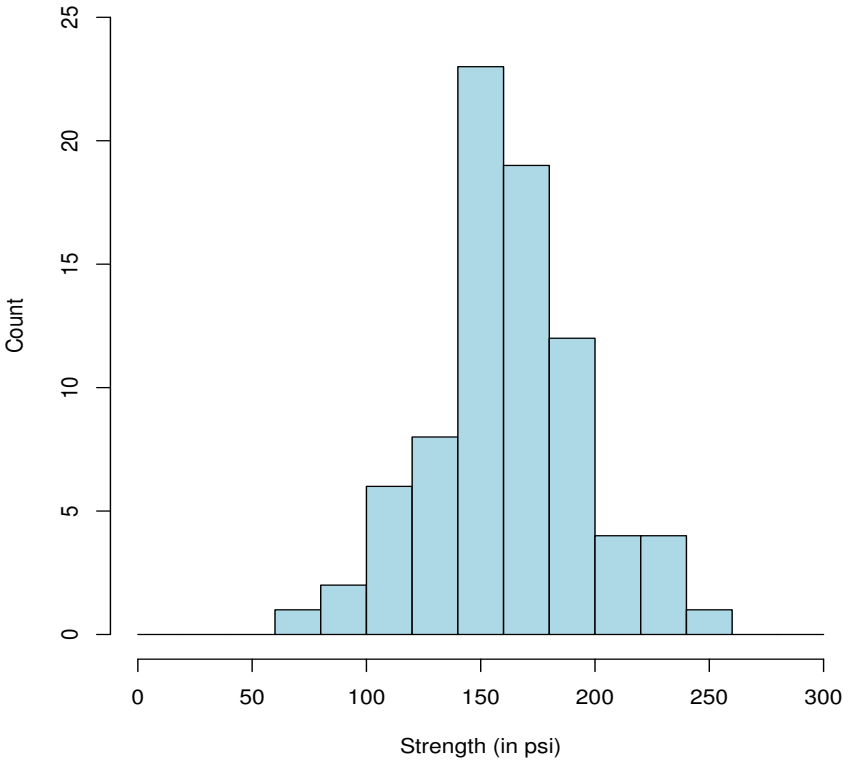

Figure 4.18: Histogram of the sample of $n = 80$ alloy specimens.

Remark: The first thing I do in any data analysis is look at the data graphically. A histogram of the $n = 80$ specimen strengths is shown in Figure 4.18 (above). With just a sample, it can be hard to make good prognostications about "what's going on" in the population of all aluminum-lithium alloy specimens. However, at least the shape we see in the histogram does align with symmetric shape of a normal distribution. This is reassuring but by no means determinative. There are many types of distributions which are symmetric or approximately symmetric.

Terminology: A **quantile-quantile plot** (**qq plot**) is a graphical display that can help assess how well a distribution fits a data set. Here is how the plot is constructed:

- On the vertical axis, we plot the observed data ordered from low to high.
- On the horizontal axis, we plot the same number of (ordered) percentiles from the distribution assumed for the observed data (i.e., a normal distribution).

Our intuition should suggest the following:

- If the observed data align with the normal distribution's percentiles, then the qq plot should look like a **straight line**. This suggests the normal distribution fits the data well and is therefore a reasonable choice for the population.

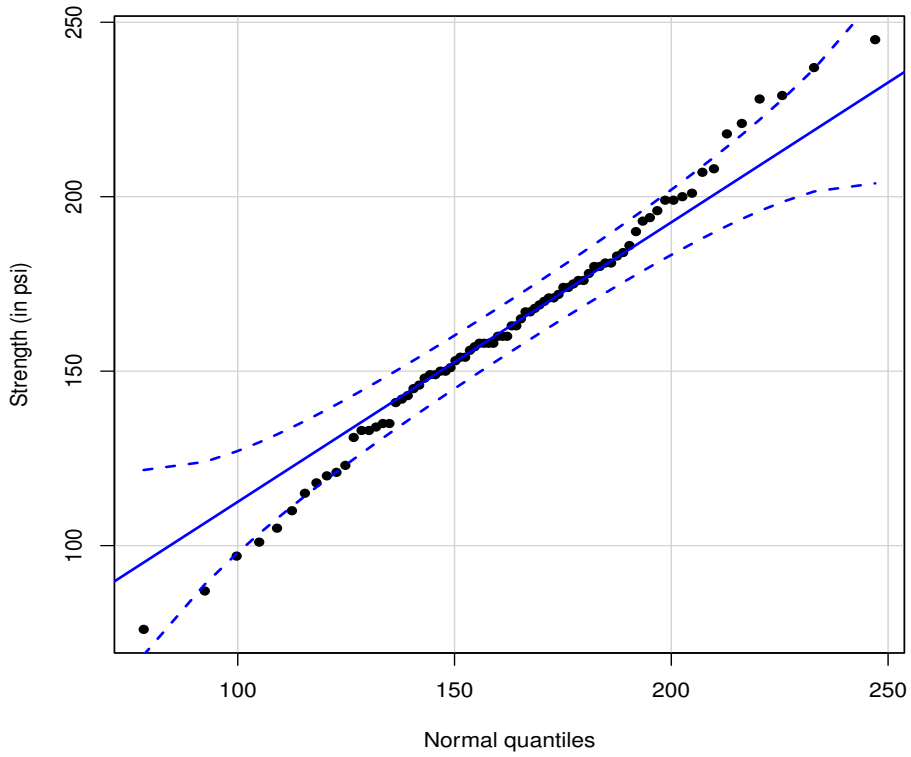

Figure 4.19: Normal qq plot of the sample of $n = 80$ alloy specimens.

- If the observed data do not align with the normal distribution's percentiles, then the qq plot should have **curvature** in it. This suggests the distribution may not be a good choice for the population.

Important: When you interpret qq plots, you are looking for **general agreement**. The observed data will never line up perfectly with the normal distribution's percentiles due to natural variability−even when the normal distribution is the correct model for the population! In other words, don't be "too picky" when interpreting these plots, especially with small sample sizes.

- If there is disagreement, it will usually happen in the tails of the distribution (left or right).
- The bands in Figure 4.19 are shown to give the viewer a reference for how much variability about the line "is allowed." However, even if a few points fall outside the bands, especially if in the tails, this should not cause dramatic concern.

Example 4.33. Arsenic (As) is a chemical element found naturally in ground water. Excessive levels may result from contamination caused by hazardous waste or by industries that make or use arsenic. Environmental engineers sampled $n = 102$ water wells in Texas and

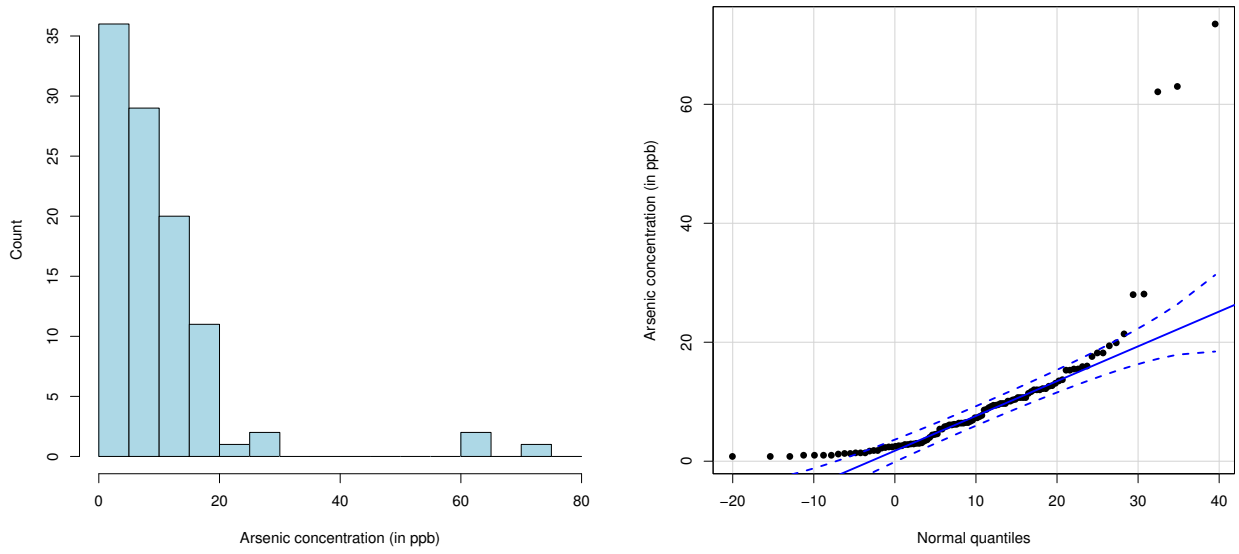

Figure 4.20: Left: Histogram of the $n = 102$ arsenic concentrations in Example 6.5. Right: Normal qq plot.

measured the arsenic concentration Y (in parts per billion, ppb) for each well. The observed data are shown below:

17.6	10.4	13.5	4.0	19.9	16.0	12.0	12.2	11.4	12.7	3.0	10.3	21.4	19.4	9.0
6.5	10.1	8.7	9.7	6.4	9.7	63.0	15.5	10.7	18.2	7.5	6.1	6.7	6.9	0.8
73.5	12.0	28.0	12.6	9.4	6.2	15.3	7.3	10.7	15.9	5.8	1.0	8.6	1.3	13.7
2.8	2.4	1.4	2.9	13.1	15.3	9.2	11.7	4.5	1.0	1.2	0.8	1.0	2.4	4.4
2.2	2.9	3.6	2.5	1.8	5.9	2.8	1.7	4.6	5.4	3.0	3.1	1.3	2.6	1.4
2.3	1.5	4.0	1.8	2.6	3.4	1.4	10.7	18.2	7.7	6.5	12.2	10.1	6.4	10.7
6.1	0.8	12.0	28.1	9.4	6.2	7.3	9.7	62.1	15.5	6.4	9.5			

Population: all water wells in Texas
Sample: the 102 water wells sampled.

Q: Is it reasonable to assume these data arise from a normal population distribution?
A: Figure 4.20 (above) shows the histogram and the normal qq plot for these data.

- The histogram shows a strong skewed right shape with multiple outliers, which we know doesn't align with the normal distribution.
- The appearance of the histogram appears to align more with a population distribution that is skewed right.
- Not surprisingly, we see **strong disagreement** in the observed data and the normal quantiles in the qq plot. Normality is not a good assumption for these data.

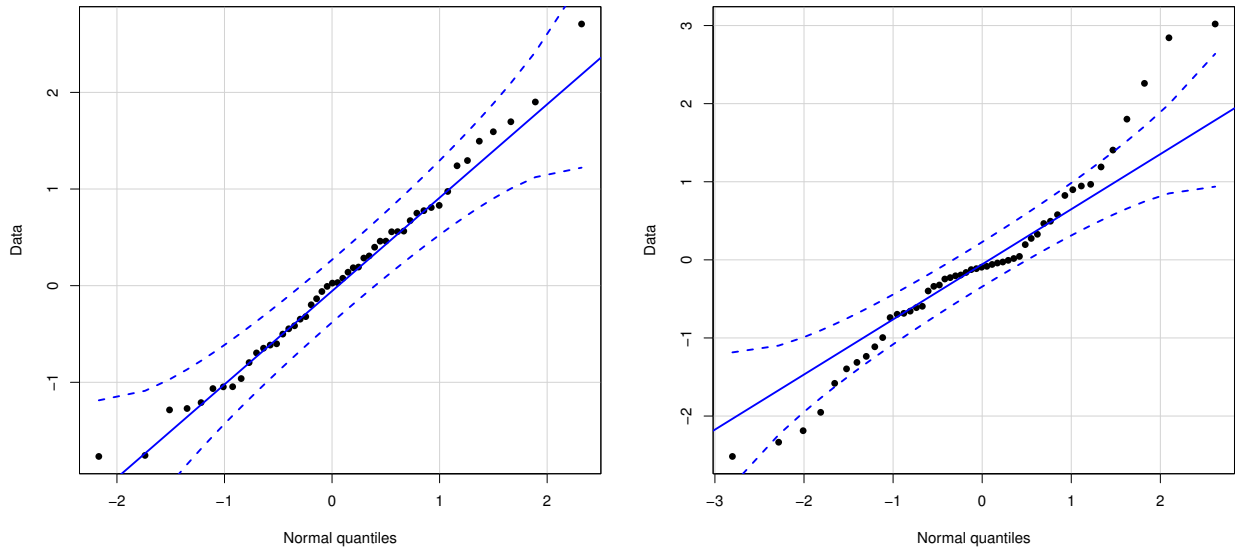

Figure 4.21: Two qq plots with sample size $n = 50$. The $\mathcal{N}(0, 1)$ distribution is assumed for the population.

Exercise: Figure 4.21 (above) shows two qq plots, each with sample size $n = 50$. The population distribution is assumed to be normal with mean $\mu = 0$ and $\sigma^2 = 1$.

Q: I simulated two data sets. I then constructed qq plots (above) for each data set under the assumption the $\mathcal{N}(0, 1)$ population distribution is correct. Which plot do you think used data simulated from the correct distribution?
A: This was a trick question! Both qq plots show data sets simulated from the correct distribution.

- In reality, I simulated about 20 data sets from the correct $\mathcal{N}(0, 1)$ distribution and constructed qq plots for each one.
- I then selected the qq plot that looked "the best" (left) and the one that looked "the worst" (right).
- The lesson here is that qq plots, while helpful in model assessment, should not be meticulously overanalyzed. This is especially true with small sample sizes.

Remark: The reason this exercise is important is so you can see that even when the assumed model is correct, the observed data may not reflect that. Some students may look at the qq plot on the right and say, *"there is too much disagreement here so a normal distribution is not appropriate."* Unbeknownst to them, this would be incorrect. Of course, in real life, no one tells us what the correct distribution is, so all we can do is make judgements based on what we see in the sample. Fortunately, many statistical analyses which "assume normality" enjoy a certain type of robustness to this assumption. This will be discussed repeatedly going forward.