STAT 509

STATISTICS FOR ENGINEERS

Spring 2018

Lecture Notes

Joshua M. Tebbs

Department of Statistics

University of South Carolina

Contents

1 Introduction

Definition: Statistics is the science of data; how to interpret data, analyze data, and design studies to collect data.

- Statistics is used in all disciplines; not just in engineering.
- "Statisticians get to play in everyone else's back yard." (John Tukey)

Examples:

1. In a reliability (time to event) study, engineers are interested in describing the time until failure for a jet engine fan blade.
2. In a genetics study involving patients with Alzheimer's disease (AD), researchers wish to identify genes that are differentially expressed (when compared to non-AD patients).
3. In an agricultural experiment, researchers want to know which of four fertilizers (which vary in their nitrogen levels) produces the highest corn yield.
4. In a clinical trial, physicians want to determine which of two drugs is more effective for treating HIV in the early stages of the disease.
5. In a public health study involving "at-risk" teenagers, epidemiologists want to know whether smoking is more common in a particular demographic class.
6. A food scientist is interested in determining how different feeding schedules (for pigs) could affect the spread of salmonella during the slaughtering process.
7. A pharmacist is concerned that administering caffeine to premature babies will increase the incidence of necrotizing enterocolitis.
8. A research dietician wants to determine if academic achievement is related to body mass index (BMI) among African American students in the fourth grade.

What we do: Statisticians use their skills in mathematics and computing to formulate statistical models and analyze data for a specific problem at hand. These models are then used to estimate important quantities of interest, to test the validity of proposed conjectures, and to predict future behavior. Being able to identify and model sources of **variability** is an important part of this process.

Definition: A **deterministic model** is one that makes no attempt to explain variability. For example, in chemistry, the ideal gas law states that

$$PV = nRT,$$

where P = pressure of a gas, V = volume, n = the amount of substance of gas (number of moles), R = Boltzmann's constant, and T = temperature. In circuit analysis, Ohm's law states that

$$V = IR,$$

where V = voltage, I = current, and R = resistance.

- In both of these models, the relationship among the variables is **completely determined** without ambiguity.
- In real life, this is rarely true for the obvious reason: there is natural variation that arises in the measurement process.
- For example, a common electrical engineering experiment involves setting up a simple circuit with a known resistance R. For a given current I, different students will then measure the voltage V.
  - With a sample of $n = 20$ students, conducting the experiment in succession, we might very well get 20 different measured voltages!
  - A deterministic model is too simplistic; it does not acknowledge the inherent variability that arises in the measurement process.

Usefulness: Statistical models are not deterministic. They incorporate **variability**. They can also be used to **predict** future outcomes.

Example 1.1. Suppose that I am trying to predict

$$Y = \text{MATH 141 final course percentage}$$

for incoming freshmen enrolled in MATH 141. For each freshmen student, I will record the following variables:

$$\begin{aligned} x_1 &= \text{SAT MATH score} \\ x_2 &= \text{high school GPA.} \end{aligned}$$

Here are SAT/HS GPA data on $n = 50$ freshmen and their final MATH 141 scores. Because there are three variables, each student's data value can be thought of as a point in three-dimensional space.

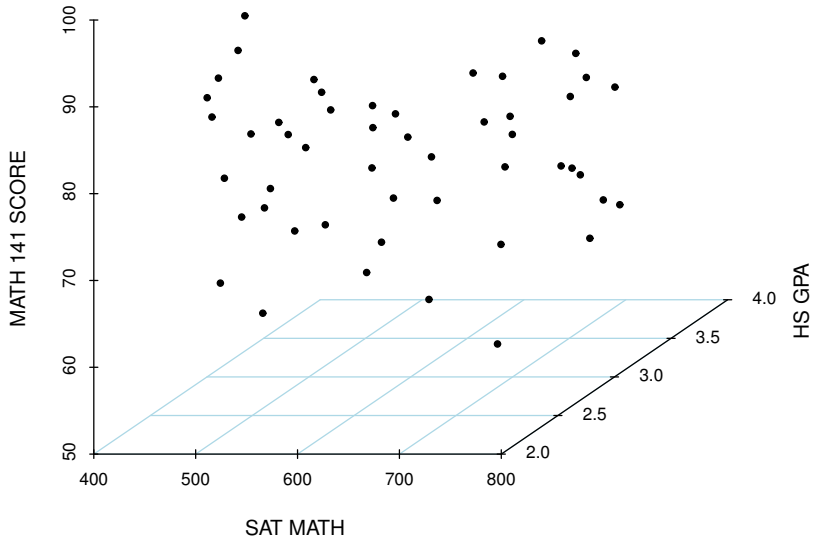

Figure 1.1: Three-dimensional scatterplot (point cloud) of USC freshmen data.

Note: In this example, a deterministic model would take the form

$$Y = f(x_1, x_2).$$

This model suggests that for a student with values x_1 and x_2, we could compute Y exactly if the function f was known. Clearly, this is neither realistic nor remotely supported by the data above.

Note: A **statistical model** for Y might look like something like this:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where ϵ is a term that accounts for not only measurement error (e.g., incorrect student information, data entry errors, grading errors, etc.) but also

- all of the other variables not accounted for (e.g., major, difficulty of schedule, study habits, natural ability, etc.) and
- the error induced by assuming a **linear relationship** between Y and x_1 and x_2 when, in fact, the relationship may not be linear.

Discussion:

- Is this sample of students representative of some larger population? After all, we would like our model/predictions to be useful on a larger scale (and not simply for these 50 students).
  - This is the idea behind **statistical inference**. We would like to use sample information to make statements about a larger (relevant) population.
- How should we estimate β_0, β_1, and β_2 in the model above?
  - If we can do this, then we can produce **predictions** of Y on a student-by-student basis (e.g., for future students, etc.).
  - This may be of interest to academic advisers who are trying to predict the success of their incoming students.
  - We can also characterize numerical **uncertainty** with our predictions.
- **Probability** is the "mathematics of uncertainty" and forms the basis for all of statistics. Therefore, we start here.

2 Probability

2.1 Sample spaces and events

Terminology: Probability is a measure of one's belief in the occurrence of a future event. Here are some events to which we may wish to assign a probability:

- tomorrow's temperature exceeding 80 degrees
- manufacturing a defective part
- concluding one fertilizer is superior to another when it isn't
- the NASDAQ losing 5 percent of its value.
- you being diagnosed with prostate/cervical cancer in the next 20 years.

Terminology: The set of all possible outcomes for a given random experiment is called the **sample space**, denoted by S.

- The number of outcomes in S is denoted by n_S.

Example 2.1. In each of the following random experiments, we write out a corresponding sample space.

(a) The Michigan state lottery calls for a three-digit integer to be selected:

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

The size of the set of all possible outcomes is $n_S = 1000$.

(b) A USC undergraduate student is tested for chlamydia (0 = negative, 1 = positive):

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

The size of the set of all possible outcomes is $n_S = 2$.

(c) Four equally qualified applicants (a, b, c, d) are competing for two positions. If the positions are **identical** (so that selection order does not matter), then

$$S = \{ab, ac, ad, bc, bd, cd\}.$$

The size of the set of all possible outcomes is $n_S = 6$. If the positions are **different** (e.g., project leader, assistant project leader, etc.), then

$$S = \{ab, ba, ac, ca, ad, da, bc, cb, bd, db, cd, dc\}.$$

In this case, the size of the set of all possible outcomes is $n_S = 12$.

Terminology: Suppose that S is a sample space for a random experiment. We would like to assign probability to an **event** A. This will quantify how likely the event is. The **probability** that the event A occurs is denoted by $P(A)$.

Terminology: Suppose that a sample space S contains $n_S < \infty$ outcomes, each of which is **equally likely**. If the event A contains n_A outcomes, then

$$P(A) = \frac{n_A}{n_S}.$$

This is called an **equiprobability model**. Its main requirement is that all outcomes in S are equally likely.

- **Important:** If the outcomes in S are not equally likely, then this result is not applicable.

Example 2.1 (continued). In the random experiments from Example 2.1, we use the previous result to assign probabilities to events (if applicable).

(a) The Michigan state lottery calls for a three-digit integer to be selected:

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

There are $n_S = 1000$ outcomes possible. Let the event

$$\begin{aligned} A &= \{000, 005, 010, 015, ..., 990, 995\} \\ &= \{\text{winning number is a multiple of 5}\}. \end{aligned}$$

There are $n_A = 200$ outcomes in A. It is reasonable to assume that each outcome in S is equally likely. Therefore,

$$P(A) = \frac{200}{1000} = 0.20.$$

(b) A USC undergraduate student is tested for chlamydia (0 = negative, 1 = positive):

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

There are $n_S = 2$ outcomes possible. However, is it reasonable to assume that each outcome in S (0 = negative, 1 = positive) is equally likely?

- The prevalence of chlamydia among college age students is much less than 50 percent (in SC, this prevalence is probably somewhere between 1-5 percent).
- Therefore, it would be illogical to assign probabilities using an equiprobability model.

(c) Four equally qualified applicants (a, b, c, d) are competing for two positions. If the positions are **identical** (so that selection order does not matter), then

$$S = \{ab, ac, ad, bc, bd, cd\}.$$

There are $n_S = 6$ outcomes possible. If A is the event that applicant d is selected for one of the two positions, then

$$\begin{aligned} A &= \{ad, bd, cd\} \\ &= \{\text{applicant } d \text{ is chosen}\}. \end{aligned}$$

There are $n_A = 3$ outcomes in A. If each applicant has the same chance of being selected (an assumption), then each of the $n_S = 6$ outcomes in S is equally likely. Therefore,

$$P(A) = \frac{3}{6} = 0.50.$$

Again, this calculation is valid only if the outcomes in S are equally likely.

Interpretation: What does $P(A)$ measure? There are two main interpretations:

- $P(A)$ measures the likelihood that A will occur on any given experiment.
- If the experiment is performed many times, then $P(A)$ can be interpreted as the percentage of times that A will occur "over the long run." This is called the **relative frequency** interpretation.

Example 2.2. Suppose a baseball's team winning percentage is 0.571. We can interpret this as the probability that the team will win a particular game. We can also interpret this as the "long-run" percentage of games won (over the course of a season, say). I used R to simulate this team's winning percentages over the course of a 162-game season.

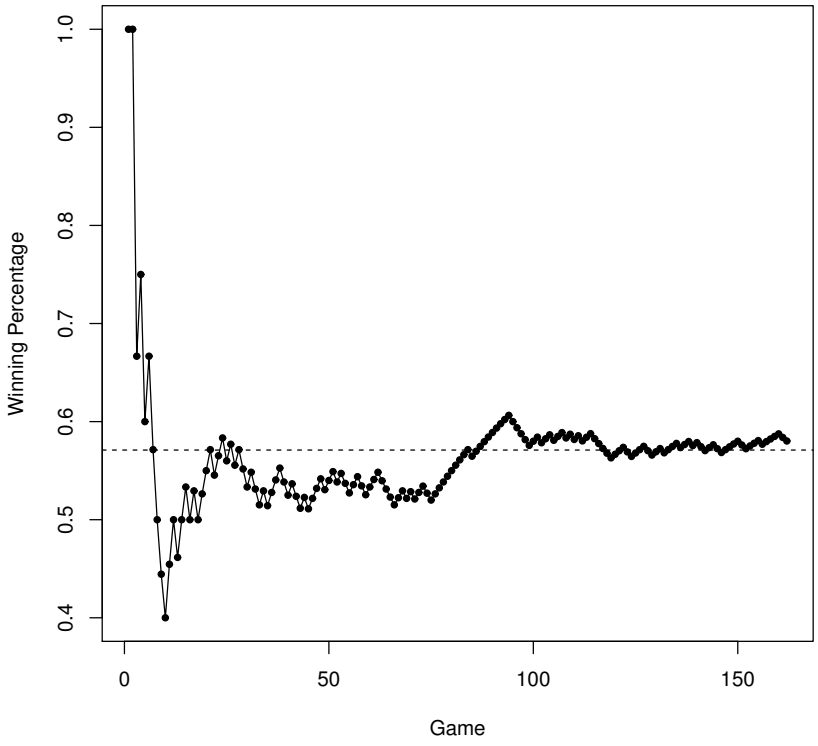

Figure 2.1: Plot of baseball team's winning percentage over 162 games. A horizontal line at 0.571 has been added.

Curiosity: Why did I pick 0.571? This was the winning percentage of the Oakland A's during the 2002 season after 119 games (68-51). If you have seen Moneyball, you know that the A's then went on a 20-game winning streak ("The Streak"). To get an idea of how amazing this run was, let's simulate 20 game outcomes assuming that 0.571 is the correct winning percentage for each game:

```
> games = rbinom(20,1,0.571)
> games
 [1] 1 0 0 1 1 1 1 1 0 0 0 1 1 1 0 1 1 1 0 1
> sum(games)
[1] 13
```

In this simulation, the A's won 13 games out of 20.

Now, let's simulate the process of playing 20 games **1000 times**. Let's keep track of the number of times (out of 1000) that the team would win 20 games in a row:

```
> games = rbinom(1000,20,0.571)
> length(games[games>19])
[1] 0
```

In 1000 simulated 20-game stretches, the team **never** won 20 games in a row.

Let's simulate **10,000** 20-game stretches:

```
> games = rbinom(10000,20,0.571)
> length(games[games>19])
[1] 0
```

In 10,000 simulated 20-game stretches, the team **never** won 20 games in a row.

Let's simulate **1,000,000** 20-game stretches:

```
> games = rbinom(1000000,20,0.571)
> length(games[games>19])
[1] 13
```

In 1,000,000 simulated 20-game stretches, the team won 20 games in a row 13 times. Using the relative frequency interpretation of probability, we could say that

$$P(\text{winning 20 games in a row}) \approx 0.000013.$$

2.2 Unions and intersections

Terminology: The **null event**, denoted by $\emptyset$, is an event that contains no outcomes (therefore, the null event cannot occur). The null event has probability $P(\emptyset) = 0$.

Terminology: The **union** of two events A and B contains all outcomes ω in either event or in both. We denote the union of two events A and B by

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}.$$

Terminology: The **intersection** of two events A and B contains all outcomes ω in both events. We denote the intersection of two events A and B by

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

Terminology: If the events A and B contain no common outcomes, we say the events are **mutually exclusive** (or **disjoint**). In this case,

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

Example 2.3. Hemophilia is a sex-linked hereditary blood defect of males characterized by delayed clotting of the blood. When a woman is a carrier of classical hemophilia, there is a 50 percent chance that a male child will inherit this disease. If a carrier gives birth to two males (not twins), what is the probability that **either** will have the disease? **both** will have the disease?

SOLUTION. We can envision the process of having two male children as an experiment with sample space

$$S = \{++, +-, -+, --\},$$

where "+" means the male offspring has the disease; "−" means the male does not have the disease. **To compute the probabilities, we will assume that each outcome in S is equally likely.** Define the events:

$$\begin{aligned} A &= \{\text{first child has disease}\} = \{++, +-\} \\ B &= \{\text{second child has disease}\} = \{++, -+\}. \end{aligned}$$

The union and intersection of A and B are, respectively,

$$\begin{aligned} A \cup B &= \{\text{either child has disease}\} = \{++, +-, -+\} \\ A \cap B &= \{\text{both children have disease}\} = \{++\}. \end{aligned}$$

The probability that either male child will have the disease is

$$P(A \cup B) = \frac{n_{A \cup B}}{n_S} = \frac{3}{4} = 0.75.$$

The probability that both male children will have the disease is

$$P(A \cap B) = \frac{n_{A \cap B}}{n_S} = \frac{1}{4} = 0.25.$$

2.3 Axioms of probability

Kolmogorov's Axioms: For any sample space S, 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 $i \neq j$.
- The event

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

means "at least one A_i occurs."

2.4 Conditional probability and independence

Note: In some situations, we may have **prior knowledge** about the likelihood of other events related to the event of interest. We can then incorporate this information into a probability calculation.

Terminology: 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,

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

Example 2.4. In a company, 36 percent of the employees have a degree from a SEC university, 22 percent of those employees with a degree from the SEC are engineers, and 30 percent of the employees are engineers. An employee is selected at random.

(a) Compute the probability that the employee is an engineer **and** is from the SEC.

(b) Compute the conditional probability that the employee is from the SEC, **given** that s/he is an engineer.

SOLUTION: Define the events

$$\begin{aligned} A &= \{\text{employee is an engineer}\} \\ B &= \{\text{employee is from the SEC}\}. \end{aligned}$$

From the information in the problem, we are given:

$$\begin{aligned} P(A) &= 0.30 \\ P(B) &= 0.36 \\ P(A|B) &= 0.22. \end{aligned}$$

In part (a), we want $P(A \cap B)$. Note that

$$0.22 = P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{0.36}.$$

Therefore,

$$P(A \cap B) = 0.22(0.36) = 0.0792.$$

In part (b), we want $P(B|A)$. From the definition of conditional probability:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.0792}{0.30} = 0.264.$$

Important: Note that, in this example, the conditional probability $P(B|A)$ and the unconditional probability $P(B)$ are not equal.

- In other words, knowledge that A "has occurred" has changed the likelihood that B occurs.
- In other situations, it might be that the occurrence (or non-occurrence) of a companion event has no effect on the probability of the event of interest. This leads us to the definition of independence.

Terminology: When the occurrence or non-occurrence of B has no effect on whether or not A occurs, and vice-versa, we say that 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,

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

and

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

These results only apply if A and B are independent.

Example 2.5. In an engineering system, two components are placed in a **series**; that is, the system is functional as long as both components are. Each component is functional with probability 0.95. Define the events

$$\begin{aligned} A_1 &= \{\text{component 1 is functional}\} \\ A_2 &= \{\text{component 2 is functional}\} \end{aligned}$$

so that $P(A_1) = P(A_2) = 0.95$. Because we need both components to be functional, the probability that the system is functional is given by $P(A_1 \cap A_2)$.

- If the components operate independently, then A_1 and A_2 are independent events and the system reliability is

$$P(A_1 \cap A_2) = P(A_1)P(A_2) = 0.95(0.95) = 0.9025.$$

- If the components do not operate independently; e.g., failure of one component "wears on the other," we can not compute $P(A_1 \cap A_2)$ without additional knowledge.

Extension: The notion of independence extends to any finite collection of events $A_1, A_2, ..., A_n$. **Mutual independence** means that the probability of the intersection of any sub-collection of $A_1, A_2, ..., A_n$ equals the product of the probabilities of the events in the sub-collection. For example, if A_1, A_2, A_3, and A_4 are mutually independent, then

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1 \cap A_2 \cap A_3 \cap A_4) &= P(A_1)P(A_2)P(A_3)P(A_4). \end{aligned}$$

2.5 Probability rules

Terminology: Suppose S is a sample space and that A is an event. The **complement** of A, denoted by $\overline{A}$, is the collection of all outcomes in S not in A. That is,

$$\overline{A} = \{\omega \in S : \omega \notin A\}.$$

1. Complement rule: Suppose that A is an event.

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

2. Additive law: Suppose that A and B are two events.

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

3. Multiplicative law: Suppose that A and B are two events.

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

4. Law of Total Probability (LOTP): Suppose that A and B are two events.

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

5. Bayes' Rule: Suppose that A and B are two events.

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

Example 2.6. 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

$$\begin{aligned} A_1 &= \{\text{train 1 is on time}\} \\ A_2 &= \{\text{train 2 is on time}\}. \end{aligned}$$

We are given that $P(A_1) = 0.95$, $P(A_2) = 0.93$, and $P(A_1 \cap A_2) = 0.90$.

(a) What is the probability that train 1 is **not on time**?

$$\begin{aligned} P(\overline{A}_1) &= 1 - P(A_1) \\ &= 1 - 0.95 = 0.05. \end{aligned}$$

(b) What is the probability that **at least one** train is on time?

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.95 + 0.93 - 0.90 = 0.98. \end{aligned}$$

(c) What is the probability that train 1 is on time **given** that train 2 is on time?

$$\begin{aligned} P(A_1|A_2) &= \frac{P(A_1 \cap A_2)}{P(A_2)} \\ &= \frac{0.90}{0.93} \approx 0.968. \end{aligned}$$

(d) What is the probability that train 2 is on time **given** that train 1 is not on time?

$$\begin{aligned} P(A_2|\overline{A}_1) &= \frac{P(\overline{A}_1 \cap A_2)}{P(\overline{A}_1)} \\ &= \frac{P(A_2) - P(A_1 \cap A_2)}{1 - P(A_1)} \\ &= \frac{0.93 - 0.90}{1 - 0.95} = 0.60. \end{aligned}$$

(e) Are A_1 and A_2 **independent** events?

ANSWER: They are not independent because

$$P(A_1 \cap A_2) \neq P(A_1)P(A_2).$$

Equivalently, note that $P(A_1|A_2) \neq P(A_1)$. In other words, knowledge that A_2 has occurred changes the likelihood that A_1 occurs.

Example 2.7. An insurance company classifies people as "accident-prone" and "non-accident-prone." For a fixed year, the probability that an accident-prone person has an accident is 0.4, and the probability that a non-accident-prone person has an accident is 0.2. The population is estimated to be 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 that

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

(a) What is the probability that a new policy-holder will have an accident?

SOLUTION: By the Law of Total Probability,

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

(b) Suppose that the policy-holder does have an accident. What is the probability that s/he was "accident-prone?"

SOLUTION: We want $P(B|A)$. By Bayes' Rule,

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})} \\ &= \frac{0.4(0.3)}{0.4(0.3) + 0.2(0.7)} \approx 0.46. \end{aligned}$$

2.6 Random variables

Terminology: A **random variable** Y is a variable whose value is determined by chance. The **distribution** of a random variable consists of two parts:

1. an elicitation of the set of all possible values of Y (called the **support**)
2. a function that describes how to assign probabilities to events involving Y.

Notation: By convention, we denote random variables by upper case letters towards the end of the alphabet; e.g., W, X, Y, Z, etc. A possible value of Y (i.e., a value in the support) is denoted generically by the lower case version y. In words,

$$P(Y = y)$$

is read, "the probability that the random variable Y equals the value y."

Terminology: If a random variable Y can assume only a finite (or countable) number of values, we call Y a **discrete** random variable. If it makes more sense to envision Y as assuming values in an interval of numbers, we call Y a **continuous** random variable.

Example 2.8. Classify the following random variables as **discrete** or **continuous** and specify the support of each random variable.

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

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

$$\{v : v = 0, 1, 2, ..., 12\}.$$

- The random variable W is **continuous**. It most certainly assumes values in

$$\{w : -\infty < w < \infty\}.$$

  Of course, 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 assume values in

$$\{x : x > 0\}.$$

  The key feature 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 assume values in

$$\{y : y = 0, 1\},$$

  where I have arbitrarily labeled "1" for passing and "0" for failing. Random variables that can assume exactly 2 values (e.g., 0, 1) are called **binary**.

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

$$\{z : z = 0, 1, 2, ..., \}.$$

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

3 Discrete Distributions

3.1 Introduction

Terminology: Suppose that Y is a **discrete** random variable. The function

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

is called the **probability mass function** (**pmf**) for Y. The pmf $p_Y(y)$ is a function that assigns probabilities to each possible value of Y.

Properties: A pmf $p_Y(y)$ for a discrete random variable Y satisfies the following:

1. $0 < p_Y(y) < 1$, for all possible values of y
2. The sum of the probabilities, taken over all possible values of Y, must equal 1; i.e.,

$$\sum_{\text{all } y} p_Y(y) = 1.$$

Example 3.1. A mail-order computer business has six telephone lines. Let Y denote the number of lines in use at a specific time. Suppose that the probability mass function (pmf) of Y is given by

y	0	1	2	3	4	5	6
$p_Y(y)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

- Figure 3.1 (left) displays $p_Y(y)$, the **probability mass function** (pmf) of Y.
  - The height of the bar above y is equal to $p_Y(y) = P(Y = y)$.
  - If y is not equal to 0, 1, 2, 3, 4, 5, 6, then $p_Y(y) = 0$.
- Figure 3.1 (right) displays the **cumulative distribution function** (cdf) of Y.

$$F_Y(y) = P(Y \leq y).$$

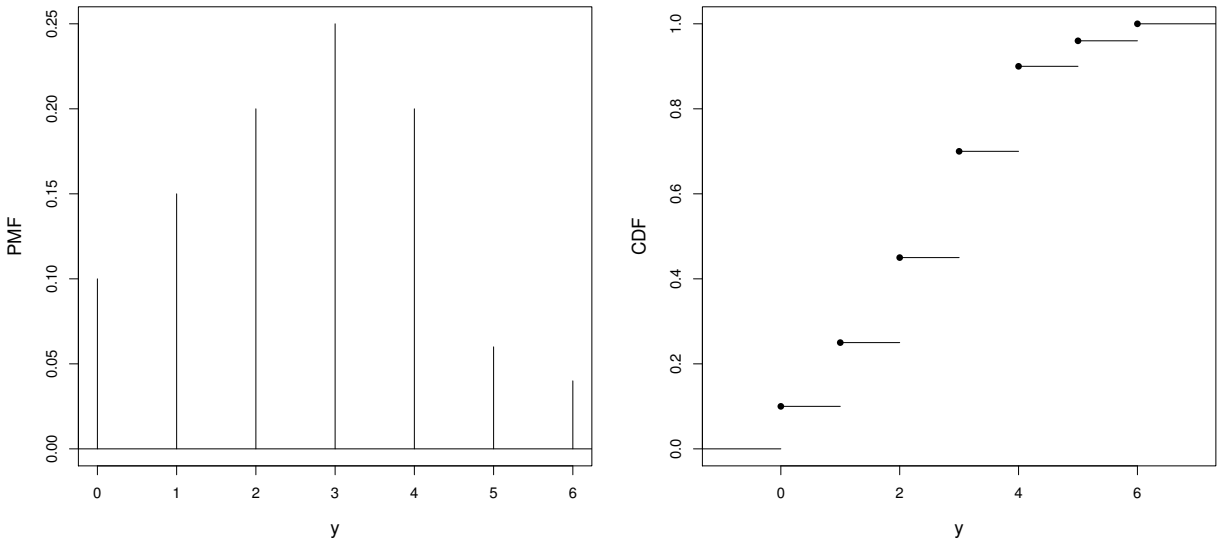

Figure 3.1: PMF (left) and CDF (right) of Y in Example 3.1.

- The cdf $F_Y(y)$ is a nondecreasing function.
- $0 \leq F_Y(y) \leq 1$; this makes sense since $F_Y(y) = P(Y \leq y)$ is a probability.
- The cdf $F_Y(y)$ in this example (Y is discrete) takes a "step" at each possible value of Y and stays constant otherwise.
- The **height** of the step at a particular y is equal to $p_Y(y) = P(Y = y)$.

(a) What is the probability that **exactly two** lines are in use?

$$p_Y(2) = P(Y = 2) = 0.20.$$

(b) What is the probability that **at most two** lines are in use?

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= p_Y(0) + p_Y(1) + p_Y(2) \\ &= 0.10 + 0.15 + 0.20 = 0.45. \end{aligned}$$

Note: This is also equal to $F_Y(2) = 0.45$ (see graph above).

(c) What is the probability that **at least five** lines are in use?

$$\begin{aligned} P(Y \geq 5) &= P(Y = 5) + P(Y = 6) \\ &= p_Y(5) + p_Y(6) = 0.06 + 0.04 = 0.10. \end{aligned}$$

We could have also computed

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y \leq 4) \\ &= 1 - F_Y(4) = 1 - 0.90 = 0.10. \end{aligned}$$

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

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

The expected value for a discrete random variable Y is a weighted average of the possible values of Y. Each value y is weighted by its probability $p_Y(y)$. In statistical applications, $\mu = E(Y)$ is called the **population mean**.

Example 3.1 (continued). In Example 3.1, we examined the distribution of Y, the number of lines in use at a specified time. The probability mass function (pmf) of Y is

y	0	1	2	3	4	5	6
$p_Y(y)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

The expected value of Y is

$$\begin{aligned} \mu = E(Y) &= \sum_{\text{all } y} y p_Y(y) \\ &= 0(0.10) + 1(0.15) + 2(0.20) + 3(0.25) + 4(0.20) + 5(0.06) + 6(0.04) \\ &= 2.64. \end{aligned}$$

Interpretation: On average, we would expect 2.64 calls at the specified time.

Interpretation: Over the long run, if we observed many values of Y at this specified time, then the average of these Y observations would be close to 2.64.

Interpretation: Place an "×" at $\mu = 2.64$ in Figure 3.1 (left). This represents the "balance point" of the probability mass function.

Result: Let Y be a discrete random variable with pmf $p_Y(y)$. Suppose that g is a real-valued function. Then, $g(Y)$ is a random variable and

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

Properties: Let Y be a discrete random variable with pmf $p_Y(y)$. Suppose that $g, g_1, g_2, ..., g_k$ are real-valued functions, and let c be any real constant. Expectations satisfy the following (linearity) properties:

(a) $E(c) = c$

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

(c) $E[\sum_{j=1}^{k} g_j(Y)] = \sum_{j=1}^{k} E[g_j(Y)]$.

Note: These rules are also applicable if Y is continuous (next chapter).

Example 3.2. In a one-hour period, the number of gallons of a certain toxic chemical that is produced at a local plant, say Y, has the following pmf:

y	0	1	2	3
$p_Y(y)$	0.2	0.3	0.3	0.2

(a) Compute the expected number of gallons produced during a one-hour period.
SOLUTION: The expected value of Y is

$$\mu = E(Y) = \sum_{\text{all } y} yp_Y(y) = 0(0.2) + 1(0.3) + 2(0.3) + 3(0.2) = 1.5.$$

Therefore, we would expect 1.5 gallons of the toxic chemical to be produced per hour (on average).

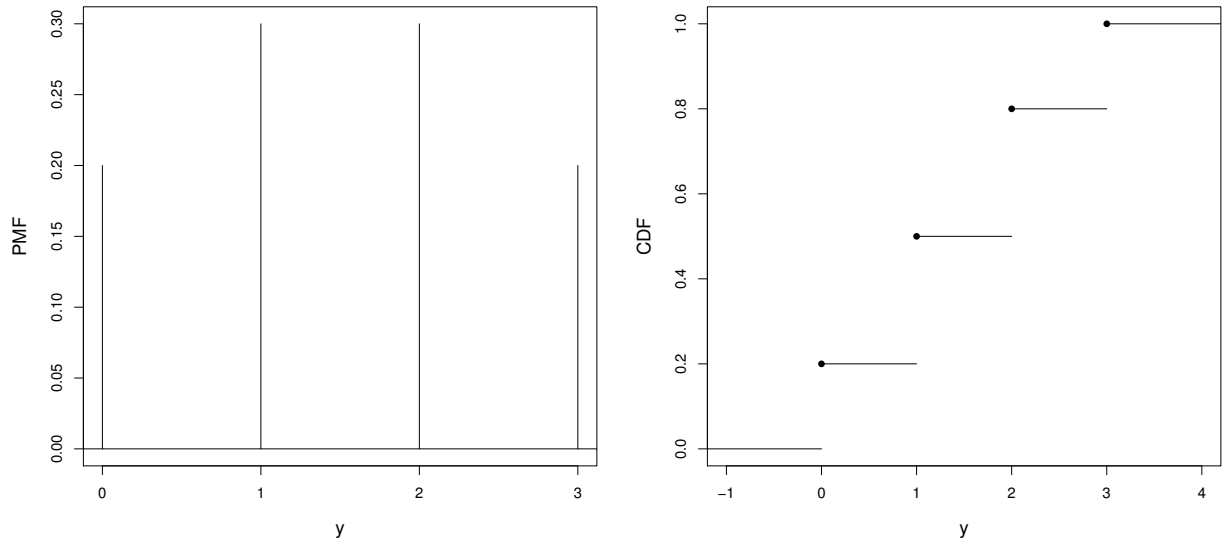

Figure 3.2: PMF (left) and CDF (right) of Y in Example 3.2.

(b) The cost (in \$100s) to produce Y gallons of this chemical per hour is

$$g(Y) = 3 + 12Y + 2Y^2.$$

What is the expected cost in a one-hour period?

SOLUTION: We want to compute $E[g(Y)]$. We first compute $E(Y^2)$:

$$E(Y^2) = \sum_{\text{all } y} y^2 p_Y(y) = 0^2(0.2) + 1^2(0.3) + 2^2(0.3) + 3^2(0.2) = 3.3.$$

Therefore,

$$\begin{aligned} E[g(Y)] = E(3 + 12Y + 2Y^2) &= 3 + 12E(Y) + 2E(Y^2) \\ &= 3 + 12(1.5) + 2(3.3) = 27.6. \end{aligned}$$

The expected hourly cost is $\$2,760.00$.

Terminology: Let Y be a discrete random variable with pmf $p_Y(y)$ and expected value $E(Y) = \mu$. The **variance** of Y is given by

$$\begin{aligned} \sigma^2 = \text{var}(Y) &= E[(Y - \mu)^2] \\ &= \sum_{\text{all } y} (y - \mu)^2 p_Y(y). \end{aligned}$$

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

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

Facts: The variance σ^2 satisfies the following:

(a) $\sigma^2 \geq 0$. $\sigma^2 = 0$ if and only if the random variable Y has a **degenerate distribution**; i.e., all the probability mass is located at one support point.

(b) The larger (smaller) σ^2 is, the more (less) spread in the possible values of Y about the population mean $\mu = E(Y)$.

(c) σ^2 is measured in (units)^2 and σ is measured in the original units.

Computing Formula: Let Y be a random variable with mean $E(Y) = \mu$. An alternative "computing formula" for the variance is

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

This formula is easy to remember and can make calculations easier.

Example 3.2 (continued). In Example 3.2, we examined the pmf for Y, the number of gallons of a toxic chemical that is produced per hour. We computed

$$\begin{aligned} E(Y) &= 1.5 \\ E(Y^2) &= 3.3. \end{aligned}$$

The variance of Y is

$$\begin{aligned} \sigma^2 = \text{var}(Y) &= E(Y^2) - [E(Y)]^2 \\ &= 3.3 - (1.5)^2 = 1.05 \text{ (gallons)}^2 \end{aligned}$$

The standard deviation of Y is

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} \approx 1.025 \text{ gallons}.$$

3.2 Binomial distribution

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

1. each trial results in a "success" or a "failure,"
2. the trials are independent, and
3. the probability of "success," denoted by p, $0 < p < 1$, is the same on every trial.

Examples:

- When circuit boards used in the manufacture of Blue Ray players are tested, the long-run percentage of defective boards is 5 percent.
  - circuit board = "trial"
  - defective board is observed = "success"
  - $p = P(\text{"success"}) = P(\text{defective board}) = 0.05$.
- 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"
  - $p = P(\text{"success"}) = P(\text{correct interpretation}) = 0.98$.
- Albino rats used to study the hormonal regulation of a metabolic pathway are injected with a drug that inhibits body synthesis of protein. The probability that a rat will die from the drug before the study is complete is 0.20.
  - rat = "trial"
  - dies before study is over = "success"
  - $p = P(\text{"success"}) = P(\text{dies early}) = 0.20$.

Terminology: Suppose that n Bernoulli trials are performed. Define

$$Y = \text{ the number of successes (out of } n \text{ trials performed).}$$

We say that Y has a **binomial distribution** with number of trials n and success probability p. **Notation:** $Y \sim b(n, p)$.

PMF: If $Y \sim b(n, p)$, then the probability mass function of Y is given by

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

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

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

Example 3.3. In an agricultural study, it is determined that 40 percent of all plots respond to a certain treatment. Four plots are observed. In this situation, we interpret

- plot of land = "trial"
- plot responds to treatment = "success"
- $p = P(\text{"success"}) = P(\text{responds to treatment}) = 0.4$.

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

$$Y = \text{ the number of plots which respond} \sim b(n = 4, p = 0.4).$$

(a) What is the probability that **exactly two** plots respond?

$$\begin{aligned} P(Y = 2) = p_Y(2) &= \binom{4}{2} (0.4)^2 (1-0.4)^{4-2} \\ &= 6(0.4)^2(0.6)^2 = 0.3456. \end{aligned}$$

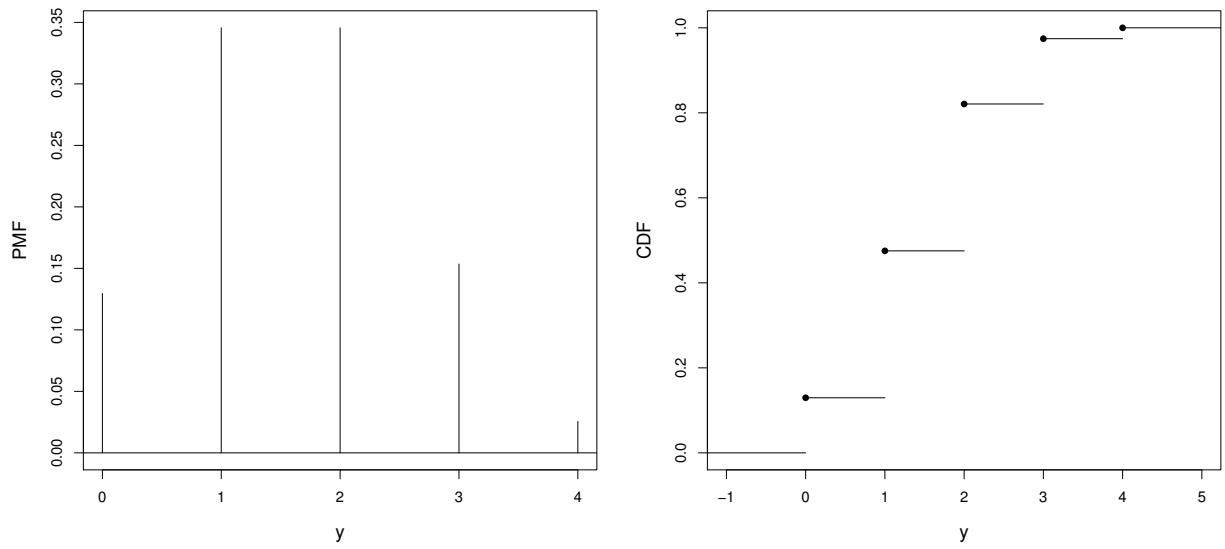

Figure 3.3: PMF (left) and CDF (right) of $Y \sim b(n = 4, p = 0.4)$ in Example 3.3.

(b) What is the probability that **at least one** plot responds?

$$\begin{aligned} P(Y \geq 1) = 1 - P(Y = 0) &= 1 - \binom{4}{0}(0.4)^0(1-0.4)^{4-0} \\ &= 1 - 1(1)(0.6)^4 = 0.8704. \end{aligned}$$

(c) What are $E(Y)$ and $\text{var}(Y)$?

$$\begin{aligned} E(Y) &= np = 4(0.4) = 1.6 \\ \text{var}(Y) &= np(1-p) = 4(0.4)(0.6) = 0.96. \end{aligned}$$

Example 3.4. An electronics manufacturer claims that 10 percent of its power supply units need servicing during the warranty period. Technicians at a testing laboratory purchase 30 units and simulate usage during the warranty period. We interpret

- power supply unit = "trial"
- supply unit needs servicing during warranty period = "success"
- $p = P(\text{"success"}) = P(\text{supply unit needs servicing}) = 0.1$.

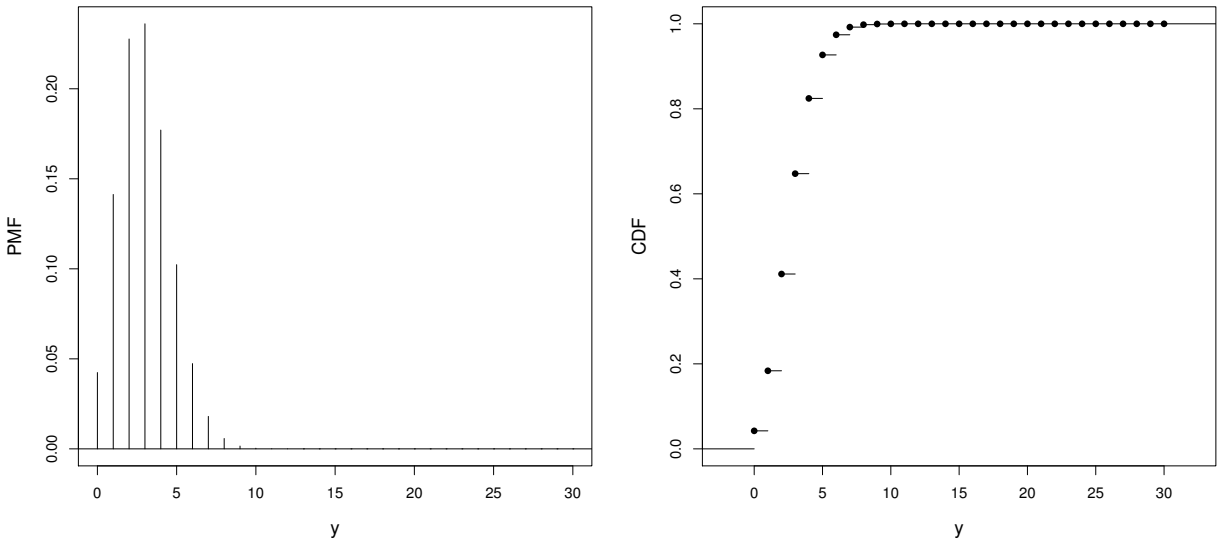

Figure 3.4: PMF (left) and CDF (right) of $Y \sim b(n = 30, p = 0.1)$ in Example 3.4.

If the Bernoulli trial assumptions hold (independent units, same probability of needing service for each unit), then

$$\begin{aligned} Y &= \text{the number of units requiring service during warranty period} \\ &\sim b(n = 30, p = 0.1). \end{aligned}$$

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

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
`dbinom(y,n,p)`	`pbinom(y,n,p)`

(a) What is the probability that **exactly five** of the 30 power supply units require servicing during the warranty period?

$$\begin{aligned} p_Y(5) = P(Y = 5) &= \binom{30}{5}(0.1)^5(1-0.1)^{30-5} \\ \texttt{dbinom(5,30,0.1)} &= 0.1023048. \end{aligned}$$

(b) What is the probability **at most five** of the 30 power supply units require service?

$$\begin{aligned} F_Y(5) = P(Y \leq 5) &= \sum_{y=0}^{5} \binom{30}{y} (0.1)^y (1-0.1)^{30-y} \\ \texttt{pbinom(5,30,0.1)} &= 0.9268099. \end{aligned}$$

(c) What is the probability **at least five** of the 30 power supply units require service?

$$\begin{aligned} P(Y \geq 5) = 1 - P(Y \leq 4) &= 1 - \sum_{y=0}^{4} \binom{30}{y} (0.1)^y (1-0.1)^{30-y} \\ \texttt{1-pbinom(4,30,0.1)} &= 0.1754949. \end{aligned}$$

(d) What is $P(2 \leq Y \leq 8)$?

$$P(2 \leq Y \leq 8) = \sum_{y=2}^{8} \binom{30}{y} (0.1)^y (1-0.1)^{30-y}.$$

One way to get this in R is to use the command:

```
> sum(dbinom(2:8,30,0.1))
[1] 0.8142852
```

The `dbinom(2:8,30,0.1)` command creates a vector containing $p_Y(2), p_Y(3), ..., p_Y(8)$, and the `sum` command adds them. Another way to calculate this probability in R is

```
> pbinom(8,30,0.1)-pbinom(1,30,0.1)
[1] 0.8142852
```

3.3 Geometric distribution

Note: The geometric distribution also arises in experiments involving Bernoulli trials:

1. Each trial results in a "success" or a "failure."

2. The trials are independent.

3. The probability of "success," denoted by p, $0 < p < 1$, is the same on every trial.

Terminology: Suppose that Bernoulli trials are continually observed. Define

$$Y = \text{ the number of trials to observe the } \textbf{first} \text{ success.}$$

We say that Y has a **geometric distribution** with success probability p. **Notation:** $Y \sim \text{geom}(p)$.

PMF: If $Y \sim \text{geom}(p)$, then the probability mass function of Y is given by

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

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

$$\begin{aligned} E(Y) &= \frac{1}{p} \\ \text{var}(Y) &= \frac{1-p}{p^2}. \end{aligned}$$

Example 3.5. Biology students are checking the eye color of fruit flies. For each fly, the probability of observing white eyes is $p = 0.25$. We interpret

- fruit fly = "trial"
- fly has white eyes = "success"
- $p = P(\text{"success"}) = P(\text{white eyes}) = 0.25$.

If the Bernoulli trial assumptions hold (independent flies, same probability of white eyes for each fly), then

$$\begin{aligned} Y &= \text{ the number of flies needed to find the } \textbf{first} \text{ white-eyed} \\ &\sim \text{geom}(p = 0.25). \end{aligned}$$

(a) What is the probability the first white-eyed fly is observed on the fifth fly checked?

$$\begin{aligned} p_Y(5) = P(Y = 5) &= (1 - 0.25)^{5-1}(0.25) \\ &= (0.75)^4(0.25) \approx 0.079. \end{aligned}$$

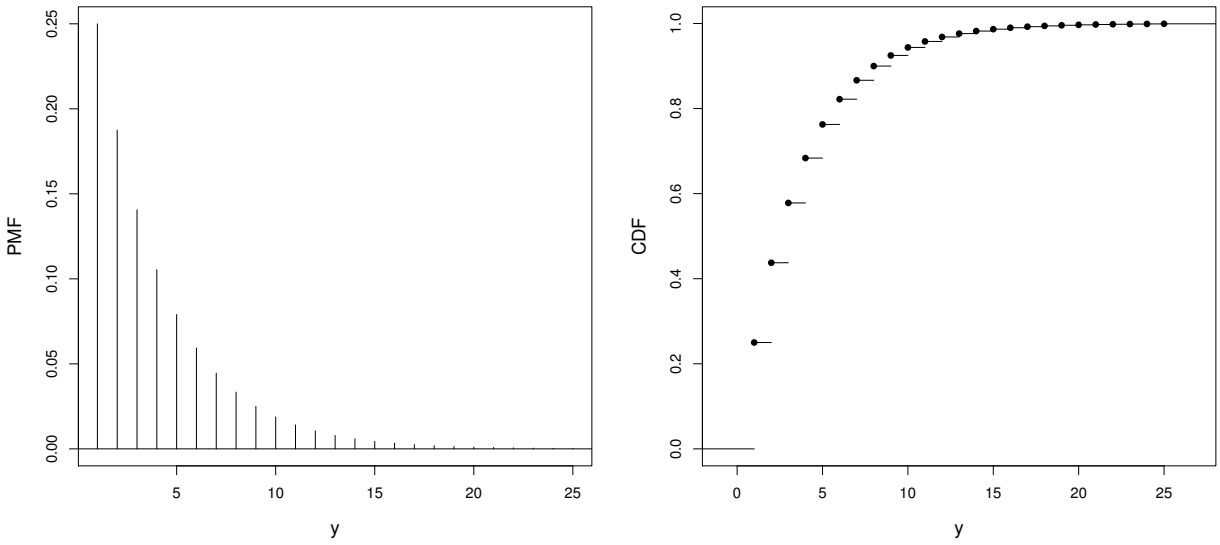

Figure 3.5: PMF (left) and CDF (right) of $Y \sim \text{geom}(p = 0.25)$ in Example 3.5.

(b) What is the probability the first white-eyed fly is observed before the fourth fly is examined? **Note:** For this to occur, we must observe the first white-eyed fly (success) on either the first, second, or third fly.

$$\begin{aligned} F_Y(3) = P(Y \leq 3) &= P(Y=1) + P(Y=2) + P(Y=3) \\ &= (1-0.25)^{1-1}(0.25) + (1-0.25)^{2-1}(0.25) + (1-0.25)^{3-1}(0.25) \\ &= 0.25 + 0.1875 + 0.140625 \approx 0.578. \end{aligned}$$

GEOMETRIC R CODE: Suppose that $Y \sim \text{geom}(p)$.

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
`dgeom(y-1,p)`	`pgeom(y-1,p)`

```
> dgeom(5-1,0.25) ## Part (a)
[1] 0.07910156
> pgeom(3-1,0.25) ## Part (b)
[1] 0.578125
```

3.4 Negative binomial distribution

Note: The negative binomial distribution also arises in experiments involving Bernoulli trials:

1. Each trial results in a "success" or a "failure."
2. The trials are independent.
3. The probability of "success," denoted by p, $0 < p < 1$, is the same on every trial.

Terminology: Suppose that Bernoulli trials are continually observed. Define

$$Y = \text{ the number of trials to observe the } r\textbf{th} \text{ success.}$$

We say that Y has a **negative binomial distribution** with waiting parameter r and success probability p. **Notation:** $Y \sim \text{nib}(r, p)$.

Remark: The negative binomial distribution is a generalization of the geometric. If $r = 1$, then the nib(r, p) distribution reduces to the geom(p).

PMF: If $Y \sim \text{nib}(r, p)$, then the probability mass function of Y is given by

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

MEAN/VARIANCE: If $Y \sim \text{nib}(r, p)$, then

$$\begin{aligned} E(Y) &= \frac{r}{p} \\ \text{var}(Y) &= \frac{r(1-p)}{p^2}. \end{aligned}$$

Example 3.6. At an automotive paint plant, 15 percent of all batches sent to the lab for chemical analysis do not conform to specifications. In this situation, we interpret

- batch = "trial"
- batch does not conform = "success"

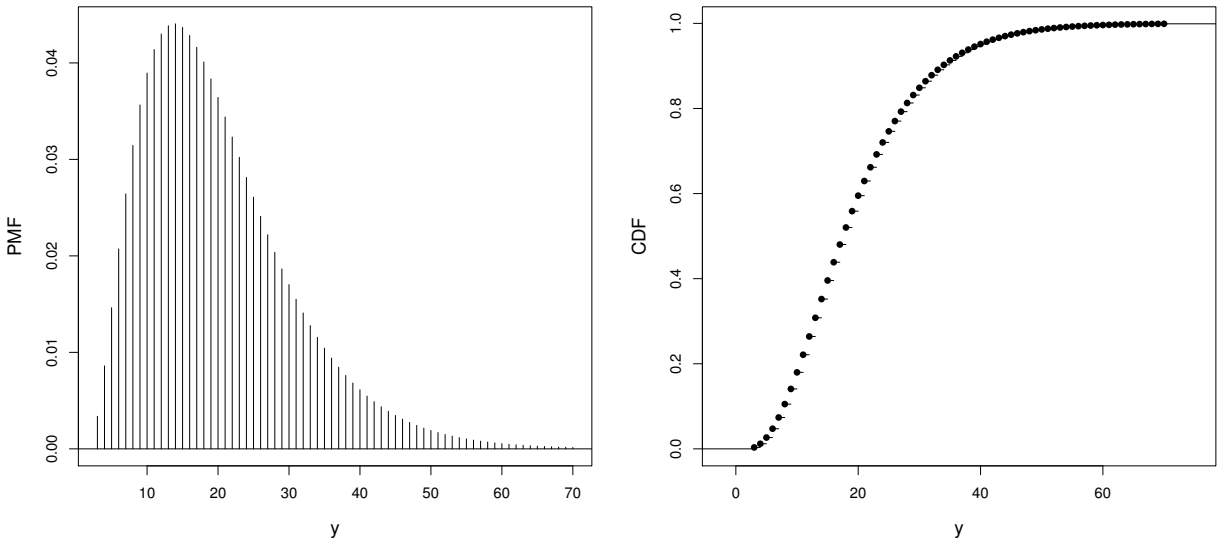

Figure 3.6: PMF (left) and CDF (right) of $Y \sim \text{nib}(r = 3, p = 0.15)$ in Example 3.6.

- $p = P(\text{“success”}) = P(\text{not conforming}) = 0.15$.

If the Bernoulli trial assumptions hold (independent batches, same probability of nonconforming for each batch), then

$$\begin{aligned} Y &= \text{the number of batches needed to find the } \textbf{third} \text{ nonconforming} \\ &\sim \text{nib}(r = 3, p = 0.15). \end{aligned}$$

(a) What is the probability the third nonconforming batch is observed on the tenth batch sent to the lab?

$$\begin{aligned} p_Y(10) = P(Y = 10) &= \binom{10-1}{3-1}(0.15)^3(1-0.15)^{10-3} \\ &= \binom{9}{2}(0.15)^3(0.85)^7 \approx 0.039. \end{aligned}$$

(b) What is the probability **no more than two** nonconforming batches will be observed among the first 30 batches sent to the lab? **Note:** This means the third nonconforming

batch must be observed on the 31st batch tested, the 32nd, the 33rd, etc.

$$\begin{aligned} P(Y \geq 31) &= 1 - P(Y \leq 30) \\ &= 1 - \sum_{y=3}^{30} \binom{y-1}{3-1} (0.15)^3 (0.85)^{y-3} \approx 0.151. \end{aligned}$$

NEGATIVE BINOMIAL R CODE: Suppose that $Y \sim \text{nib}(r, p)$.

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
`dnbinom(y-r,r,p)`	`pnbinom(y-r,r,p)`

```
> dnbinom(10-3,3,0.15) ## Part (a)
[1] 0.03895012
> 1-pnbinom(30-3,3,0.15) ## Part (b)
[1] 0.1514006
```

3.5 Hypergeometric distribution

Setting: Consider a population of N objects and suppose that each object belongs to one of two dichotomous classes: Class 1 and Class 2. For example, the objects (classes) might be people (infected/not), parts (conforming/not), plots of land (respond to treatment/not), etc. In the population of interest, we have

$$\begin{aligned} N &= \text{total number of objects} \\ r &= \text{number of objects in Class 1} \\ N - r &= \text{number of objects in Class 2.} \end{aligned}$$

Envision taking a sample n objects from the population (objects are selected at random and without replacement). Define

$$Y = \text{ the number of objects in Class 1 (out of the } n \text{ selected).}$$

We say that Y has a **hypergeometric distribution**. **Notation:** $Y \sim \text{hyper}(N, n, r)$.

PMF: If $Y \sim \text{hyper}(N, n, r)$, then the probability mass function of Y is given by

$$p_Y(y) = \begin{cases} \dfrac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & y \leq r \text{ and } n-y \leq N-r \\ 0, & \text{otherwise.} \end{cases}$$

MEAN/VARIANCE: If $Y \sim \text{hyper}(N, n, r)$, then

$$\begin{aligned} E(Y) &= n\left(\frac{r}{N}\right) \\ \text{var}(Y) &= n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right). \end{aligned}$$

Example 3.7. A supplier ships parts to a company in lots of 100 parts. The company has an **acceptance sampling plan** which adopts the following acceptance rule:

> "....sample 5 parts at random and without replacement. If there are no defectives in the sample, accept the entire lot; otherwise, reject the entire lot."

The population size is $N = 100$. The sample size is $n = 5$. Define the random variable

$$\begin{aligned} Y &= \text{the number of defectives in the sample} \\ &\sim \text{hyper}(N = 100, n = 5, r). \end{aligned}$$

(a) If $r = 10$, what is the probability that the lot will be accepted? **Note:** The lot will be accepted only if $Y = 0$.

$$p_Y(0) = P(Y=0) = \frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}} = \frac{1(43949268)}{75287520} \approx 0.584.$$

(b) If $r = 10$, what is the probability that **at least 3** of the 5 parts sampled are defective?

$$\begin{aligned} P(Y \geq 3) &= 1 - P(Y \leq 2) \\ &= 1 - [P(Y=0) + P(Y=1) + P(Y=2)] \\ &= 1 - \left[\frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}} + \frac{\binom{10}{1}\binom{90}{4}}{\binom{100}{5}} + \frac{\binom{10}{2}\binom{90}{3}}{\binom{100}{5}}\right] \\ &\approx 1 - (0.584 + 0.339 + 0.070) = 0.007. \end{aligned}$$

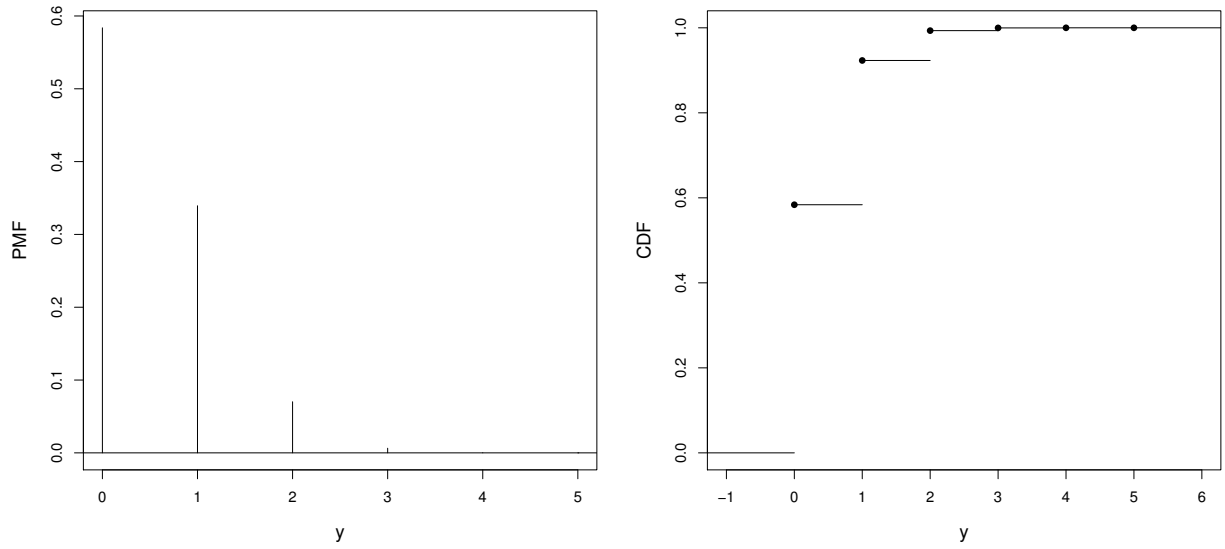

Figure 3.7: PMF (left) and CDF (right) of $Y \sim \text{hyper}(N = 100, n = 5, r = 10)$ in Example 3.7.

HYPERGEOMETRIC R CODE: Suppose that $Y \sim \text{hyper}(N, n, r)$.

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
`dhyper(y,r,N-r,n)`	`phyper(y,r,N-r,n)`

```
> dhyper(0,10,100-10,5) ## Part (a)
[1] 0.5837524
> 1-phyper(2,10,100-10,5) ## Part (b)
[1] 0.006637913
```

3.6 Poisson distribution

Note: The Poisson distribution is commonly used to model **counts**, such as

1. the number of customers entering a post office in a given hour
2. the number of machine breakdowns per month

3. the number of insurance claims received per day

4. the number of defects on a piece of raw material.

Terminology: In general, we define

$$Y = \text{ the number of "occurrences" over a } \textbf{unit interval} \text{ of time (or space).}$$

A Poisson distribution for Y emerges if "occurrences" obey the following postulates:

P1. The number of occurrences in non-overlapping intervals are independent.

P2. The probability of an occurrence is proportional to the length of the interval.

P3. The probability of 2 or more occurrences in a sufficiently short interval is zero.

We say that Y has a **Poisson distribution**. **Notation:** $Y \sim \text{Poisson}(\lambda)$. A process that produces occurrences according to these postulates is called a **Poisson process**.

PMF: If $Y \sim \text{Poisson}(\lambda)$, then the probability mass function of Y is given by

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

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

$$\begin{aligned} E(Y) &= \lambda \\ \text{var}(Y) &= \lambda. \end{aligned}$$

Example 3.8. Let Y denote the number of times per month that a detectable amount of radioactive gas is recorded at a nuclear power plant. Suppose that Y follows a Poisson distribution with mean $\lambda = 2.5$ times per month.

(a) What is the probability that there are **exactly three** times a detectable amount of gas is recorded in a given month?

$$\begin{aligned} P(Y = 3) = p_Y(3) &= \frac{(2.5)^3 e^{-2.5}}{3!} \\ &= \frac{15.625 e^{-2.5}}{6} \approx 0.214. \end{aligned}$$

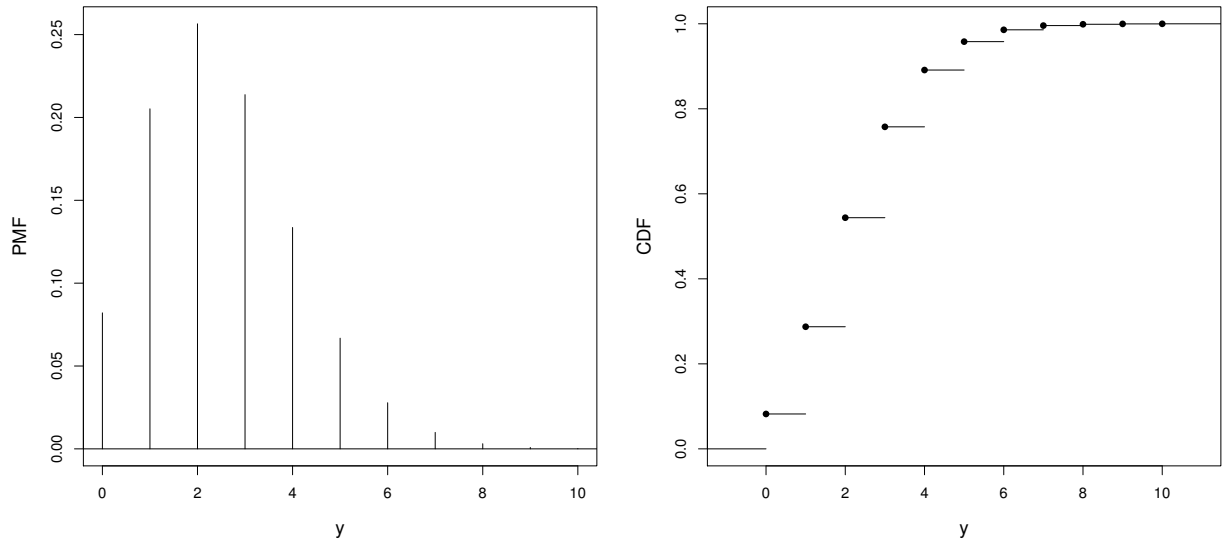

Figure 3.8: PMF (left) and CDF (right) of $Y \sim \text{Poisson}(\lambda = 2.5)$ in Example 3.8.

(b) What is the probability that there are **no more than four** times a detectable amount of gas is recorded in a given month?

$$\begin{aligned} P(Y \leq 4) &= P(Y=0) + P(Y=1) + P(Y=2) + P(Y=3) + P(Y=4) \\ &= \frac{(2.5)^0 e^{-2.5}}{0!} + \frac{(2.5)^1 e^{-2.5}}{1!} + \frac{(2.5)^2 e^{-2.5}}{2!} + \frac{(2.5)^3 e^{-2.5}}{3!} + \frac{(2.5)^4 e^{-2.5}}{4!} \\ &\approx 0.891. \end{aligned}$$

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

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

```
> dpois(3,2.5) ## Part (a)
[1] 0.213763
> ppois(4,2.5) ## Part (b)
[1] 0.891178
```

4 Continuous Distributions

4.1 Introduction

Recall: A random variable Y is called **continuous** if it can assume any value in an interval of real numbers.

- Contrast this with a discrete random variable whose values can be "counted."
- For example, if Y = time (seconds), then the set of all possible values of Y is

$$\{y : y > 0\}.$$

  If Y = temperature (deg C), the set of all possible values of Y (ignoring absolute zero and physical upper bounds) might be described as

$$\{y : -\infty < y < \infty\}.$$

  Neither of these sets of values can be "counted."

Important: Assigning probabilities to events involving continuous random variables is different than in discrete models. We do not assign positive probability to **specific values** (e.g., $Y = 3$, etc.) like we did with discrete random variables. Instead, we assign positive probability to events which are **intervals** (e.g., $2 < Y < 4$, etc.).

Terminology: Every continuous random variable we will discuss in this course has a **probability density function** (**pdf**), denoted by $f_Y(y)$. This function has the following characteristics:

1. $f_Y(y) \geq 0$, that is, $f_Y(y)$ is nonnegative.
2. The area under any pdf is equal to 1, that is,

$$\int_{-\infty}^{\infty} f_Y(y)dy = 1.$$

Terminology: The **cumulative distribution function** (**cdf**) of Y is given by

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(t)dt.$$

Result: If a and b are specific values of interest ($a \leq b$), then

$$\begin{aligned} P(a \leq Y \leq b) &= \int_a^b f_Y(y)dy \\ &= F_Y(b) - F_Y(a). \end{aligned}$$

Result: If a is a specific value, then $P(Y = a) = 0$. In other words, in continuous probability models, specific points are assigned zero probability. An immediate consequence of this is that if Y is **continuous**,

$$P(a \leq Y \leq b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a < Y < b)$$

and each is equal to

$$\int_a^b f_Y(y)dy.$$

This is not true if Y has a discrete distribution because positive probability is assigned to specific values of Y.

Remark: Evaluating a pdf at a specific value a, that is, computing $f_Y(a)$, does not give you a probability. That is,

$$f_Y(a) \neq \text{ a probability (of any type).}$$

Compare this to calculating $F_Y(a)$, which gives the cumulative probability $P(Y \leq a)$.

Example 4.1. Suppose that Y has the pdf

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

(a) Find the cumulative distribution function (cdf) of Y.

SOLUTION. Before you do any calculations, first note that

- if $y \leq 0$, then $F_Y(y) = 0$.
- if $y \geq 1$, then $F_Y(y) = 1$.

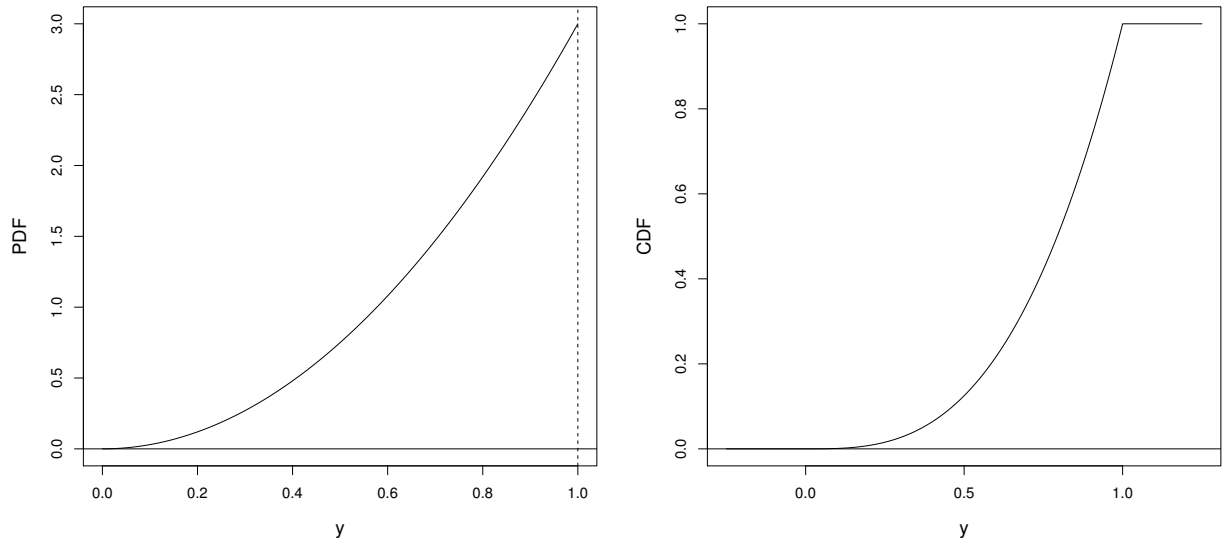

Figure 4.1: PDF (left) and CDF (right) of Y in Example 4.1.

For $0 < y < 1$,

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \int_0^y 3t^2 dt = t^3\Big|_0^y = y^3.$$

Therefore, the cdf of Y is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y^3, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

The cdf is plotted in Figure 4.1 (right).

(b) Calculate $P(Y < 0.3)$.

SOLUTION. Using the pdf of Y, we calculate

$$P(Y < 0.3) = \int_0^{0.3} 3y^2 dy = y^3\Big|_0^{0.3} = (0.3)^3 - 0^3 = 0.027.$$

Note also that, using the cdf of Y,

$$\begin{aligned} P(Y < 0.3) &= P(Y \leq 0.3) \\ &= F_Y(0.3) = (0.3)^3 = 0.027. \end{aligned}$$

We get the same answer, as we should.

(c) Calculate $P(0.3 \leq Y \leq 0.8)$.

SOLUTION. Using the pdf of Y, we calculate

$$P(0.3 \leq Y \leq 0.8) = \int_{0.3}^{0.8} 3y^2 dy = y^3\Big|_{0.3}^{0.8} = (0.8)^3 - (0.3)^3 = 0.485.$$

Note also that, using the cdf of Y,

$$\begin{aligned} P(0.3 \leq Y \leq 0.8) &= F_Y(0.8) - F_Y(0.3) \\ &= (0.8)^3 - (0.3)^3 = 0.485. \end{aligned}$$

Terminology: Let Y be a continuous random variable with pdf $f_Y(y)$. The **expected value** (or **mean**) of Y is given by

$$\mu = E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

The limits of the integral in this definition, while technically correct, will always be the lower and upper limits corresponding to the nonzero part of the pdf.

Result: Let Y be a continuous random variable with pdf $f_Y(y)$. Suppose that g is a real-valued function. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y) dy.$$

Terminology: Let Y be a continuous random variable with pdf $f_Y(y)$ and expected value $E(Y) = \mu$. The **variance** of Y is given by

$$\begin{aligned} \sigma^2 = \text{var}(Y) &= E[(Y-\mu)^2] \\ &= \int_{-\infty}^{\infty} (y-\mu)^2 f_Y(y) dy. \end{aligned}$$

The "computing formula" is still

$$\text{var}(Y) = E(Y^2) - [E(Y)]^2.$$

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

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

Exercise: Calculate $E(Y)$ and $\text{var}(Y)$ in Example 4.1.

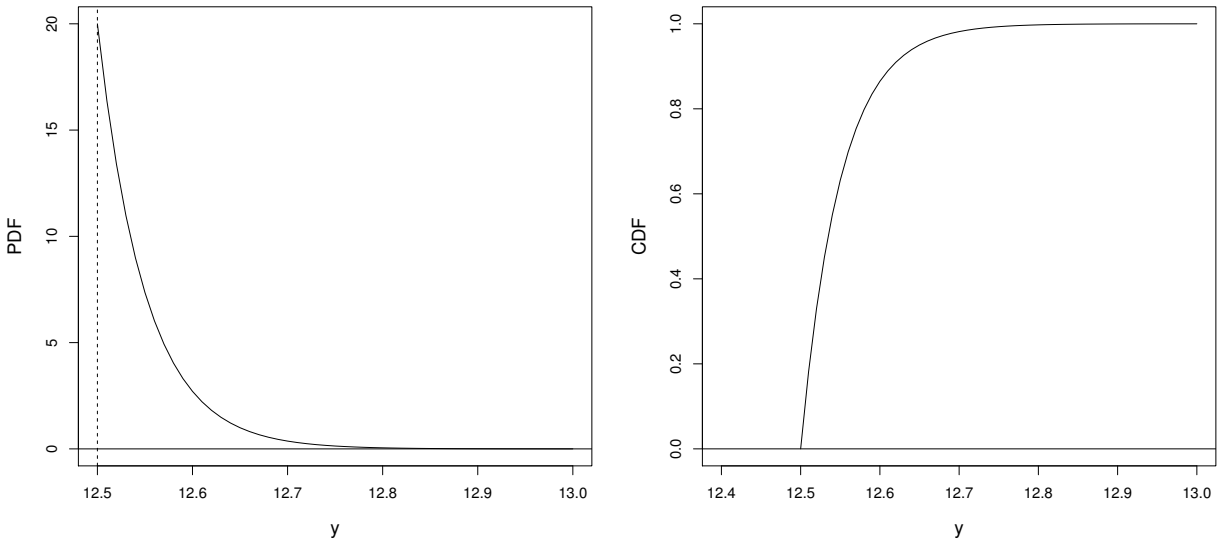

Figure 4.2: PDF (left) and CDF (right) of Y in Example 4.2.

Example 4.2. Let Y denote the diameter of a hole drilled in a sheet metal component. The target diameter is 12.5 mm and can never be lower than this. However, minor random disturbances to the drilling process always result in larger diameters. Suppose that Y is modeled using the pdf

$$f_Y(y) = \begin{cases} 20e^{-20(y-12.5)}, & y > 12.5 \\ 0, & \text{otherwise.} \end{cases}$$

The cdf of Y (verify) is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 12.5 \\ 1 - e^{-20(y-12.5)}, & y > 12.5. \end{cases}$$

The pdf (left) and cdf (right) of Y are graphed in Figure 4.2. The expected value of Y is given by

$$\begin{aligned} \mu = E(Y) &= \int_{12.5}^{\infty} y f_Y(y) dy \\ &= \int_{12.5}^{\infty} 20y e^{-20(y-12.5)} dy = 12.55. \end{aligned}$$

To do this integral, I used the `integrate` function in R:

```
> # Calculate E(Y)
> integrand.1 <- function(y){y*20*exp(-20*(y-12.5))}
> integrate(integrand.1,lower=12.5,upper=Inf)
12.55 with absolute error < 1.3e-07
```

The variance of Y is given by

$$\begin{aligned}\sigma^2 = \text{var}(Y) &= \int_{12.5}^{\infty}(y-\mu)^2 f_Y(y)dy \\ &= \int_{12.5}^{\infty} 20(y-12.55)^2 e^{-20(y-12.5)}dy = 0.0025.\end{aligned}$$

```
> # Calculate var(Y)
> integrand.2 <- function(y){(y-12.55)^2*20*exp(-20*(y-12.5))}
> integrate(integrand.2,lower=12.5,upper=Inf)
0.0025 with absolute error < 4.1e-08
```

Exercise: In Example 4.2, what proportion of diameters will exceed 12.65 mm?

Terminology: Suppose Y is a continuous random variable and let $0 < p < 1$. The p**th quantile** of the distribution of Y, denoted by ϕ_p, solves

$$F_Y(\phi_p) = P(Y \leq \phi_p) = \int_{-\infty}^{\phi_p} f_Y(y)dy = p.$$

The **median** of Y is the $p = 0.5$ quantile. That is, the median $\phi_{0.5}$ solves

$$F_Y(\phi_{0.5}) = P(Y \leq \phi_{0.5}) = \int_{-\infty}^{\phi_{0.5}} f_Y(y)dy = 0.5.$$

Another name for the pth quantile is the $100p$**th percentile**.

Example 4.2 (continued). Find the median diameter $\phi_{0.5}$ in Example 4.2.

SOLUTION. We set

$$F_Y(\phi_{0.5}) = 1 - e^{-20(\phi_{0.5}-12.5)} \stackrel{\text{set}}{=} 0.5$$

and solve for $\phi_{0.5}$. We obtain $\phi_{0.5} \approx 12.535$. Therefore, 50 percent of the diameters will be less than 12.535 mm (of course, 50 percent will be greater than this value too).

4.2 Exponential distribution

Terminology: A random variable Y is said to have an **exponential distribution** with parameter $\lambda > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Notation: $Y \sim$ exponential(λ). The exponential distribution is used to model the distribution of positive quantities (e.g., lifetimes, etc.).

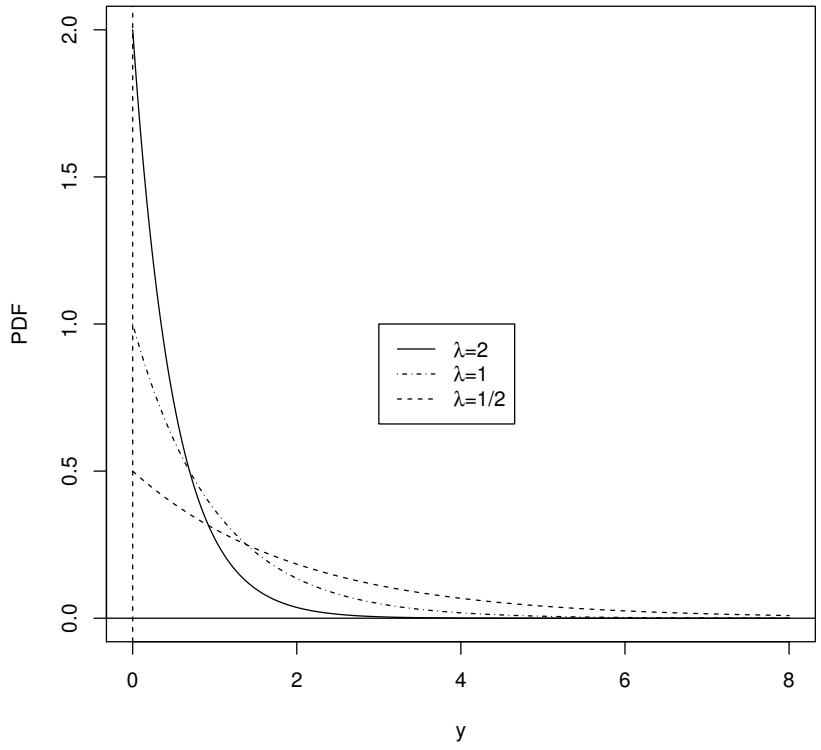

Figure 4.3: Exponential pdfs with different values of λ.

MEAN/VARIANCE: If $Y \sim$ exponential(λ), then

$$\begin{aligned} E(Y) &= \frac{1}{\lambda} \\ \text{var}(Y) &= \frac{1}{\lambda^2}. \end{aligned}$$

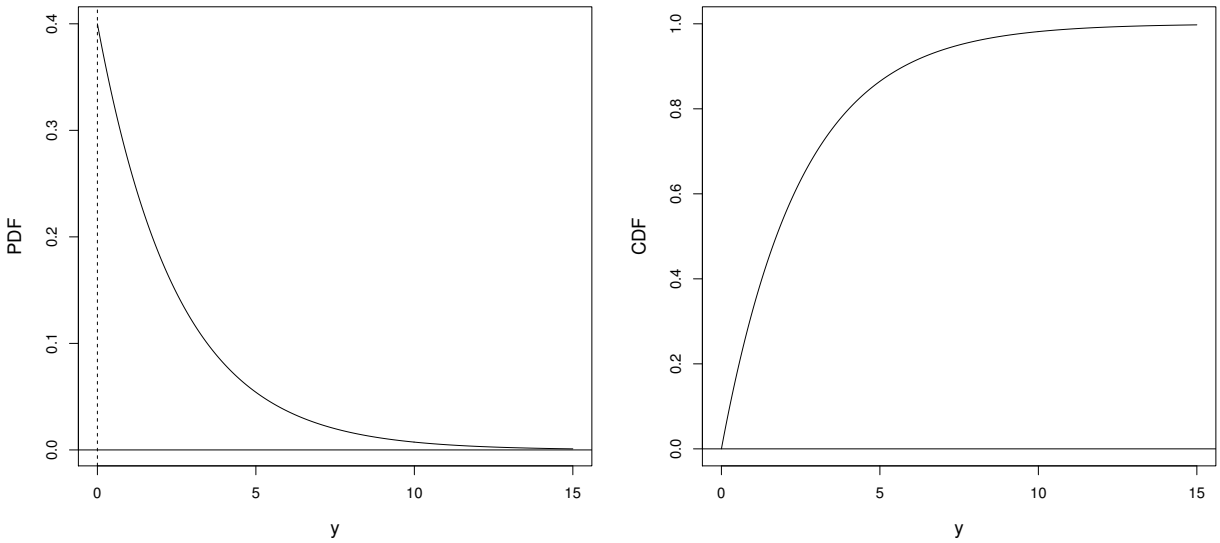

Figure 4.4: PDF (left) and CDF (right) of $Y \sim$ exponential($\lambda = 0.4$) in Example 4.3.

CDF: If $Y \sim$ exponential(λ), then the cdf of Y exists in closed form and is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-\lambda y}, & y > 0. \end{cases}$$

Example 4.3. Experience with fans used in diesel engines has suggested that the exponential distribution provides a good model for time until failure (i.e., lifetime). Suppose that the lifetime of a fan, denoted by Y (measured in 10000s of hours), follows an exponential distribution with $\lambda = 0.4$.

(a) What is the probability that a fan lasts longer than 30,000 hours?

SOLUTION. Using the pdf of Y, we calculate

$$\begin{aligned} P(Y > 3) = \int_3^\infty 0.4e^{-0.4y}dy &= 0.4\left(-\frac{1}{0.4}e^{-0.4y}\Big|_3^\infty\right) \\ &= -e^{-0.4y}\Big|_3^\infty \\ &= e^{-1.2} \approx 0.301. \end{aligned}$$

Using the cdf of Y,

$$\begin{aligned} P(Y > 3) = 1 - P(Y \leq 3) &= 1 - F_Y(3) \\ &= 1 - [1 - e^{-0.4(3)}] \\ &= e^{-1.2} \approx 0.301. \end{aligned}$$

We get the same answer, as we should.

(b) What is the probability that a fan will last between 20,000 and 50,000 hours?

SOLUTION. Using the pdf of Y, we calculate

$$\begin{aligned} P(2 < Y < 5) = \int_2^5 0.4e^{-0.4y}dy &= 0.4\left(-\frac{1}{0.4}e^{-0.4y}\Big|_2^5\right) \\ &= -e^{-0.4y}\Big|_2^5 \\ &= -[e^{-0.4(5)} - e^{-0.4(2)}] \\ &= e^{-0.8} - e^{-2} \approx 0.314. \end{aligned}$$

Using the cdf of Y,

$$\begin{aligned} P(2 < Y < 5) &= F_Y(5) - F_Y(2) \\ &= [1 - e^{-0.4(5)}] - [1 - e^{-0.4(2)}] \\ &= e^{-0.8} - e^{-2} \approx 0.314. \end{aligned}$$

MEMORYLESS PROPERTY: Suppose that $Y \sim$ exponential(λ), and let r and s be positive constants. Then

$$P(Y > r + s|Y > r) = P(Y > s).$$

If Y measures time (e.g., time to failure, etc.), then the memoryless property says that the distribution of additional lifetime (s time units beyond time r) is the same as the original distribution of the lifetime. In other words, the fact that Y has "made it" to time r has been "forgotten." For example, in Example 4.3,

$$P(Y > 5|Y > 2) = P(Y > 3) \approx 0.301.$$

POISSON RELATIONSHIP: Suppose that we are observing "occurrences" over time according to a Poisson distribution with rate λ. Define the random variable

$$Y = \text{ the } \textbf{time} \text{ until the first occurrence.}$$

Then, $Y \sim$ exponential(λ).

Example 4.4. Suppose customers arrive at a check-out according to a Poisson process with mean $\lambda = 12$ per hour.

(a) What is the probability that we will have to wait longer than 10 minutes to see the first customer? NOTE: 10 minutes is 1/6th of an hour.

SOLUTION. The **time** until the first arrival, say Y, follows an exponential distribution with $\lambda = 12$. The cdf of Y, for $y > 0$, is $F_Y(y) = 1 - e^{-12y}$. The desired probability is

$$\begin{aligned} P(Y > 1/6) = 1 - P(Y \leq 1/6) &= 1 - F_Y(1/6) \\ &= 1 - [1 - e^{-12(1/6)}] = e^{-2} \approx 0.135. \end{aligned}$$

(b) Ninety percent of all first-customer waiting times will be less than what value?

SOLUTION. We want $\phi_{0.9}$, the 90th percentile ($p = 0.9$ quantile) of the distribution of Y. We set

$$F_Y(\phi_{0.9}) = 1 - e^{-12\phi_{0.9}} \stackrel{\text{set}}{=} 0.9$$

and solve for $\phi_{0.9}$. Doing so gives $\phi_{0.9} \approx 0.192$. This means that 90 percent of all first-customer waiting times will be less than 0.192 hours (only 10 percent will exceed).

EXPONENTIAL R CODE: Suppose that $Y \sim$ exponential(λ).

$F_Y(y) = P(Y \leq y)$	ϕ_p
`pexp(y,`λ`)`	`qexp(p,`λ`)`

```
> 1-pexp(1/6,12) ## P(Y>1/6) in Example 4.4
[1] 0.1353353
> qexp(0.9,12) ## 0.9 quantile in Example 4.4
[1] 0.1918821
```

4.3 Gamma distribution

Terminology: The **gamma function** is a real function defined by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy,$$

for all $\alpha > 0$. The gamma function satisfies the recursive relationship

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1),$$

for $\alpha > 1$. Therefore, if α is an integer, then

$$\Gamma(\alpha) = (\alpha-1)!.$$

Terminology: A random variable Y is said to have a **gamma distribution** with parameters $\alpha > 0$ and $\lambda > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\lambda y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Notation: $Y \sim \text{gamma}(\alpha, \lambda)$.

- By changing the values of α and λ, the gamma pdf can assume many shapes; see examples in Figure 4.5.
- This makes the gamma distribution popular for modeling positive random variables (it is more **flexible** than the exponential).
- Note that when $\alpha = 1$, the gamma pdf reduces to the exponential(λ) pdf.

MEAN/VARIANCE: If $Y \sim \text{gamma}(\alpha, \lambda)$, then

$$\begin{aligned} E(Y) &= \frac{\alpha}{\lambda} \\ \text{var}(Y) &= \frac{\alpha}{\lambda^2}. \end{aligned}$$

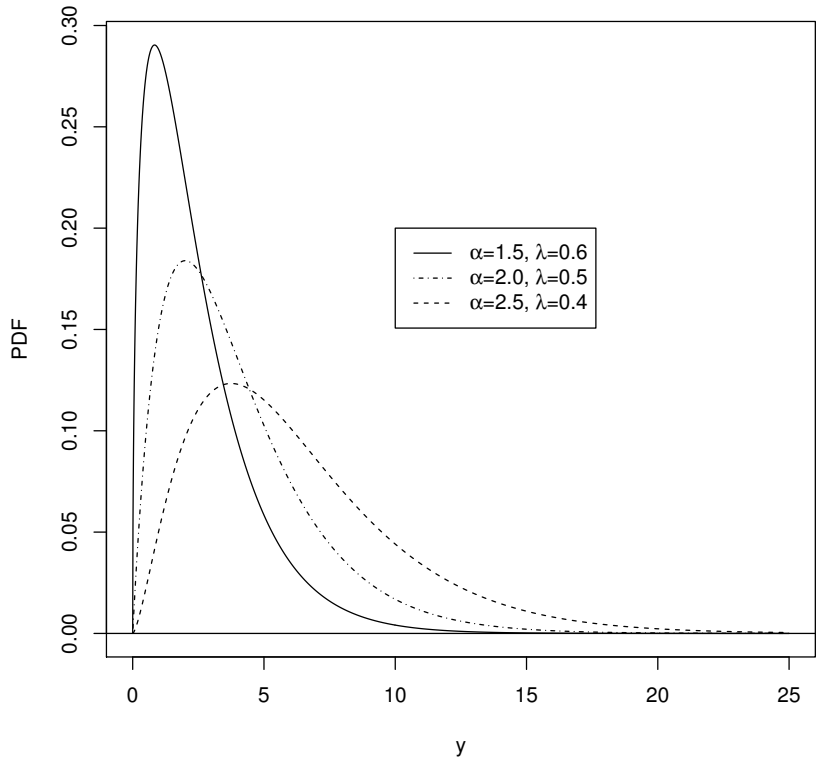

Figure 4.5: Gamma pdfs with different values of α and λ.

CDF: The cdf of a gamma random variable does not exist in closed form. Therefore, probabilities involving gamma random variables and gamma quantiles **must** be computed numerically (e.g., using R, etc.).

GAMMA R CODE: Suppose that $Y \sim \text{gamma}(\alpha, \lambda)$.

$F_Y(y) = P(Y \leq y)$	ϕ_p
`pgamma(y,`α`,`λ`)`	`qgamma(p,`α`,`λ`)`

Example 4.5. Accelerated life testing is the process of testing a product by subjecting it to conditions (stress, strain, temperatures, voltage, vibration rate, pressure etc.) in excess of its normal service conditions in an effort to uncover potential modes of failure in a short amount of time. When a certain component is subjected to this type of test, the lifetime Y (in weeks) is modeled by a gamma distribution with $\alpha = 4$ and $\lambda = 1/6$.

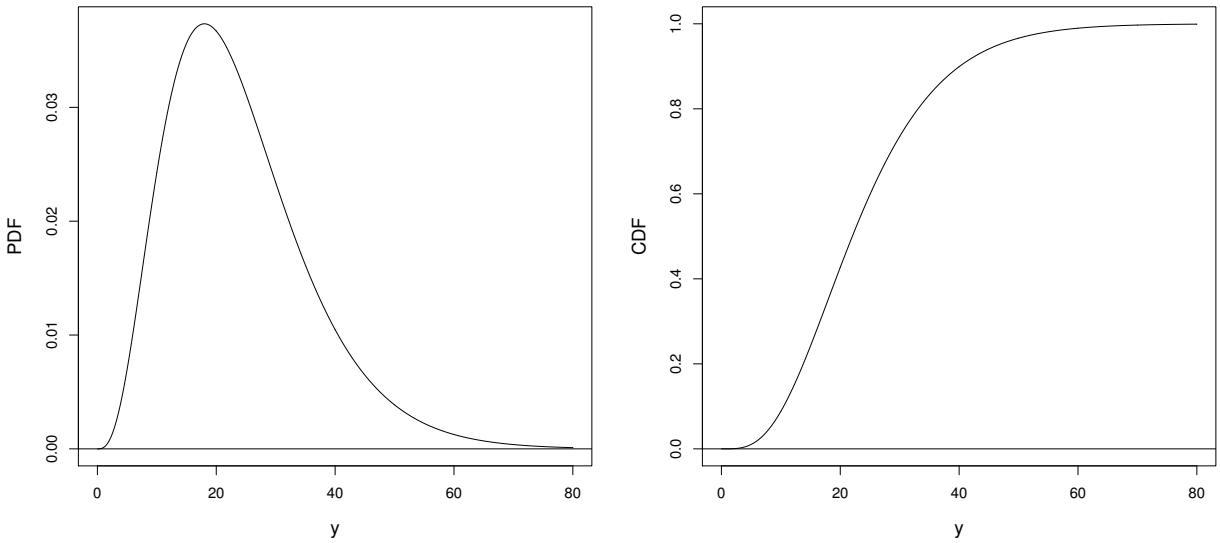

Figure 4.6: PDF (left) and CDF (right) of $Y \sim \text{gamma}(\alpha = 4, \lambda = 1/6)$ in Example 4.5.

(a) Find the probability that a component will last **at least** 50 weeks.

$$\begin{aligned} P(Y \geq 50) = 1 - P(Y < 50) &= 1 - F_Y(50) \\ &= \texttt{1-pgamma(50,4,1/6)} \\ &= \texttt{0.03377340}. \end{aligned}$$

(b) Find the probability that a component will last **between** 12 and 24 weeks.

$$\begin{aligned} P(12 \leq Y \leq 24) &= F_Y(24) - F_Y(12) \\ &= \texttt{pgamma(24,4,1/6)-pgamma(12,4,1/6)} \\ &= \texttt{0.4236533}. \end{aligned}$$

(c) Twenty percent of the component lifetimes will be **below** which time? **Note:** I am asking for the 0.20 quantile (20th percentile) of the lifetime distribution.

```
> qgamma(0.2,4,1/6)
[1] 13.78072
```

Therefore, 20 percent of the components will fail before 13.78 weeks.

POISSON RELATIONSHIP: Suppose that we are observing "occurrences" over time according to a Poisson distribution with rate λ. Define the random variable

$$Y = \text{ the } \textbf{time} \text{ until the } \alpha\text{th occurrence.}$$

Then, $Y \sim \text{gamma}(\alpha, \lambda)$.

Exercise: In Example 4.4 (pp 48), what is the distribution of the time until the 2nd customer arrives? the 3rd? Find the probability that we have to wait longer than 30 minutes for the 5th customer to arrive.

4.4 Normal distribution

Terminology: A random variable Y is said to have a **normal distribution** if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Notation: $Y \sim \mathcal{N}(\mu, \sigma^2)$. This is also known as the **Guassian distribution**.

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

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

Remark: The normal distribution serves as a very good model for a wide range of measurements; e.g., reaction times, fill amounts, part dimensions, weights/heights, measures of intelligence/test scores, economic indicators, etc.

CDF: The cdf of a normal random variable does not exist in closed form. Probabilities involving normal random variables and normal quantiles can be computed numerically (e.g., using R, etc.). There are other antiquated methods of calculating normal probabilities/quantiles using probability tables (we will avoid like the plague).

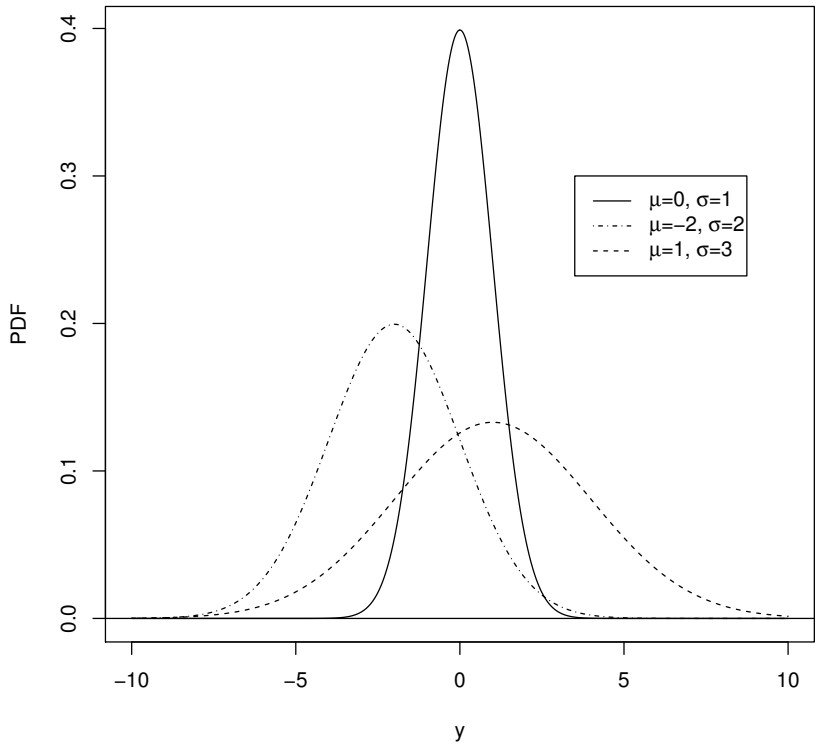

Figure 4.7: Normal pdfs with different values of μ and σ^2.

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

$F_Y(y) = P(Y \leq y)$	ϕ_p
`pnorm(y,`μ`,`σ`)`	`qnorm(p,`μ`,`σ`)`

Example 4.6. The time it takes for a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions. For a population of drivers (e.g., drivers in SC, etc.), suppose that

$Y =$ the reaction time to brake during in-traffic driving (measured in seconds),

follows a normal distribution with mean $\mu = 1.5$ and variance $\sigma^2 = 0.16$.

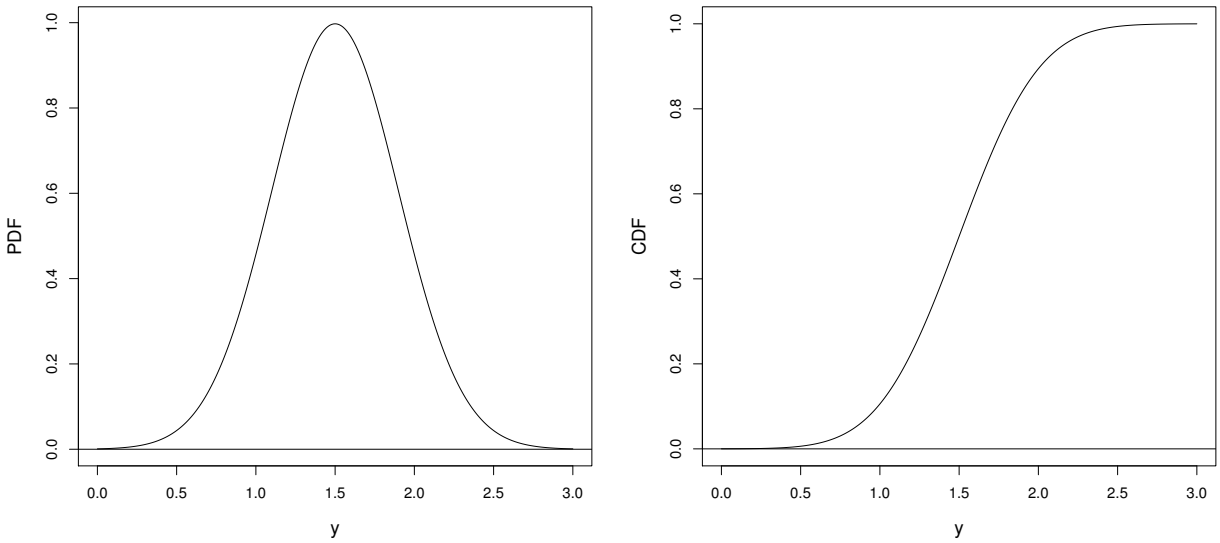

Figure 4.8: PDF (left) and CDF (right) of $Y \sim \mathcal{N}(\mu = 1.5, \sigma^2 = 0.16)$ in Example 4.6.

(a) What is the probability that reaction time is **less than** 1 second?

$$\begin{aligned} P(Y < 1) &= F_Y(1) \\ &= \texttt{pnorm(1,1.5,sqrt(0.16))} \\ &= \texttt{0.1056498}. \end{aligned}$$

(b) What is the probability that reaction time is **between** 1.1 and 2.5 seconds?

$$\begin{aligned} P(1.1 < Y < 2.5) &= F_Y(2.5) - F_Y(1.1) \\ &= \texttt{pnorm(2.5,1.5,sqrt(0.16))-pnorm(1.1,1.5,sqrt(0.16))} \\ &= \texttt{0.835135}. \end{aligned}$$

(c) Five percent of all reaction times will **exceed** which time? **Note:** I am asking for the 0.95 quantile (95th percentile) of the reaction time distribution.

```
> qnorm(0.95,1.5,sqrt(0.16))
[1] 2.157941
```

Therefore, the slowest 5 percent of the population will have reaction times larger than 2.16 seconds.

Empirical Rule: For any $\mathcal{N}(\mu, \sigma^2)$ distribution,

- about 68% of the distribution is between $\mu - \sigma$ and $\mu + \sigma$
- about 95% of the distribution is between $\mu - 2\sigma$ and $\mu + 2\sigma$
- about 99.7% of the distribution is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

This is also called the **68-95-99.7% rule.** This rule allows for us to make statements like this (referring to Example 4.6, where $\mu = 1.5$ and $\sigma = 0.4$):

> "About 68 percent of all reaction times will be between 1.1 and 1.9 seconds."
> "About 95 percent of all reaction times will be between 0.7 and 2.3 seconds."
> "About 99.7 percent of all reaction times will be between 0.3 and 2.7 seconds."

Terminology: A random variable Z is said to have a **standard normal distribution** if its pdf is given by

$$f_Z(z) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, & -\infty < z < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Notation: $Z \sim \mathcal{N}(0, 1)$. A standard normal distribution is simply a "special" normal distribution, that is, a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. The variable Z is called a **standard normal random variable**.

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

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

The result says that Z follows a standard normal distribution; i.e., $Z \sim \mathcal{N}(0, 1)$. In this context, Z is called the **standardized value** of Y. For example,

- if the standardized value of y is $z = 1.5$, this means that y is 1.5 standard deviations **above** the mean μ.
- if the standardized value of y is $z = -1.5$, this means that y is 1.5 standard deviations **below** the mean μ.

5 Reliability and Lifetime Distributions

Terminology: Reliability analysis deals with failure time (i.e., lifetime, time-to-event) data. For example,

$$\begin{aligned} T &= \text{time from start of product service until failure} \\ T &= \text{time until a warranty claim} \\ T &= \text{number of hours in use/cycles until failure.} \end{aligned}$$

We call T a **lifetime random variable** if it measures the time to an "event;" e.g., failure, death, eradication of some infection/condition, etc. Engineers are often involved with reliability studies, because reliability is strongly related to product quality.

Note: There are many well known **lifetime distributions**, including

- exponential
- Weibull
- **Others:** gamma, lognormal, inverse Gaussian, Gompertz-Makeham, Birnbaum-Sanders, extreme value, log-logistic, etc.

5.1 Weibull distribution

Terminology: A random variable T is said to have a **Weibull distribution** with parameters $\beta > 0$ and $\eta > 0$ if its pdf is given by

$$f_T(t) = \begin{cases} \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Notation: $T \sim \text{Weibull}(\beta, \eta)$. We call

$$\begin{aligned} \beta &= \textbf{shape} \text{ parameter} \\ \eta &= \textbf{scale} \text{ parameter.} \end{aligned}$$

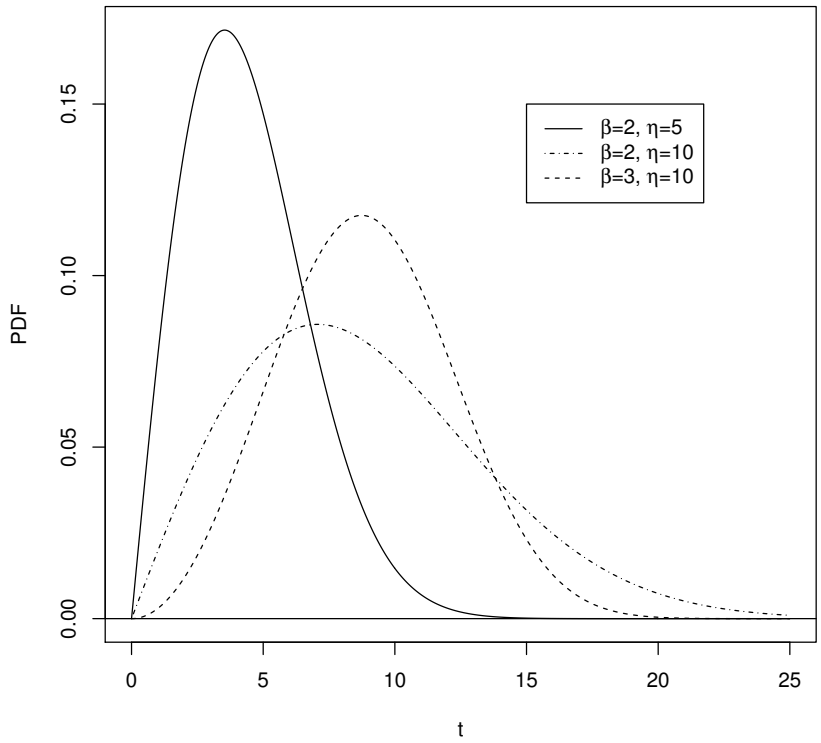

Figure 5.1: Weibull pdfs with different values of β and η.

- As you can see in Figure 5.1, by changing the values of β and η, the Weibull pdf can assume many shapes. Because of this flexibility (and for other reasons), the Weibull distribution is very popular among engineers in reliability applications.
- When $\beta = 1$, the Weibull pdf reduces to the exponential($\lambda = 1/\eta$) pdf. In other words, the exponential model is a special case of the more general Weibull distribution.

MEAN/VARIANCE: If $T \sim \text{Weibull}(\beta, \eta)$, then

$$\begin{aligned} E(T) &= \eta\Gamma\left(1+\frac{1}{\beta}\right) \\ \text{var}(T) &= \eta^2\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}. \end{aligned}$$

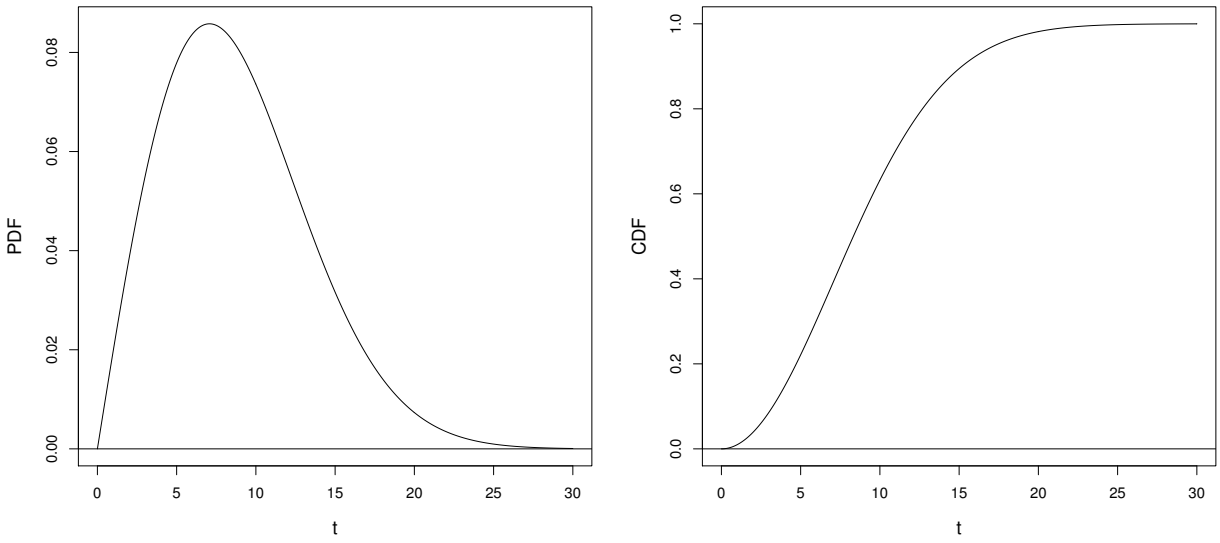

Figure 5.2: PDF (left) and CDF (right) of $T \sim \text{Weibull}(\beta = 2, \eta = 10)$ in Example 5.1.

CDF: Suppose that $T \sim \text{Weibull}(\beta, \eta)$. The cdf of T exists in closed form and is given by

$$F_T(t) = \begin{cases} 0, & t \leq 0 \\ 1 - e^{-(t/\eta)^\beta}, & t > 0. \end{cases}$$

Example 5.1. The lifetime of a rechargeable battery under constant usage conditions, denoted by T (measured in hours), follows a Weibull distribution with parameters $\beta = 2$ and $\eta = 10$.

(a) What is the **mean** time to failure?

$$E(T) = 10\Gamma\left(\frac{3}{2}\right) \approx 8.862 \text{ hours.}$$

(b) What is the probability that a battery is still functional at time $t = 20$?

$$\begin{aligned} P(T \geq 20) = 1 - P(T < 20) &= 1 - F_T(20) \\ &= 1 - [1 - e^{-(20/10)^2}] \\ &\approx 0.018. \end{aligned}$$

(c) What is the probability that a battery is still functional at time $t = 20$ **given** that the

battery is functional at time $t = 10$?

SOLUTION. This is a conditional probability. We are given that the battery has "survived" to at least 10 hours.

$$\begin{aligned} P(T \geq 20|T \geq 10) = \frac{P(T \geq 20 \text{ and } T \geq 10)}{P(T \geq 10)} &= \frac{P(T \geq 20)}{P(T \geq 10)} \\ &= \frac{1 - F_T(20)}{1 - F_T(10)} \\ &= \frac{e^{-(20/10)^2}}{e^{-(10/10)^2}} \approx 0.050. \end{aligned}$$

Remark: Note that

$$0.050 \approx P(T \geq 20|T \geq 10) \neq P(T \geq 10) = e^{-(10/10)^2} \approx 0.368.$$

Therefore, the Weibull distribution does **not** satisfy the memoryless property.

(d) What is the 99th percentile of this lifetime distribution? We set

$$F_T(\phi_{0.99}) = 1 - e^{-(\phi_{0.99}/10)^2} \stackrel{\text{set}}{=} 0.99.$$

Solving for $\phi_{0.99}$ gives $\phi_{0.99} \approx 21.460$ hours. Only one percent of the battery lifetimes will exceed this value.

WEIBULL R CODE: Suppose that $T \sim \text{Weibull}(\beta, \eta)$.

$F_T(t) = P(T \leq t)$	ϕ_p
`pweibull(t,`β`,`η`)`	`qweibull(p,`β`,`η`)`

```
> 10*gamma(3/2) ## Part (a)
[1] 8.86227
> 1-pweibull(20,2,10) ## Part (b)
[1] 0.01831564
> (1-pweibull(20,2,10))/(1-pweibull(10,2,10)) ## Part (c)
[1] 0.04978707
> qweibull(0.99,2,10) ## Part (d)
[1] 21.45966
```

5.2 Reliability functions

Description: We now describe some different, but equivalent, ways of defining the distribution of a (continuous) lifetime random variable T.

- The **cumulative distribution function** (**cdf**)

$$F_T(t) = P(T \leq t).$$

  This can be interpreted as the proportion of units that have **failed** by time t.

- The **survivor function**

$$S_T(t) = P(T > t) = 1 - F_T(t).$$

  This can be interpreted as the proportion of units that have **not failed** by time t; e.g., the unit is still functioning, a warranty claim has not been made, etc.

- The **probability density function** (**pdf**)

$$f_T(t) = \frac{d}{dt}F_T(t) = -\frac{d}{dt}S_T(t).$$

  Also, recall that

$$F_T(t) = \int_0^t f_T(u)du$$

  and

$$S_T(t) = \int_t^\infty f_T(u)du.$$

Terminology: The **hazard function** of a lifetime random variable T is

$$h_T(t) = \lim_{\epsilon \to 0} \frac{P(t \leq T < t + \epsilon | T \geq t)}{\epsilon}.$$

The hazard function is not a probability; rather, it is a **probability rate**. Therefore, it is possible that a hazard function may exceed one.

Interpretation: The hazard function offers a very useful interpretation of a lifetime distribution. **It indicates how the rate of failure varies with time**.

- Distributions with increasing hazard functions are seen in units where some kind of aging or "wear out" takes place. The population gets **weaker** over time.

- Distributions with decreasing hazard functions correspond to the population getting **stronger** over time. For example, certain types of units (e.g., electronic devices, etc.) may display a decreasing hazard function, at least in the early stages.

- In some populations, the hazard function decreases initially, stays constant for a period of time, and then increases. This corresponds to a population whose units get stronger initially (defective individuals "die out" early), exhibit random failures for a period of time (constant hazard), and then eventually the population weakens (e.g., due to old age, etc.). These hazard functions are **bathtub-shaped**.

Result: The hazard function can be calculated if we know the pdf $f_T(t)$ and the survivor function $S_T(t)$. To see why, note that

$$\begin{aligned} h_T(t) &= \lim_{\epsilon \to 0} \frac{P(t \leq T < t+\epsilon | T \geq t)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{P(t \leq T < t+\epsilon)}{\epsilon \, P(T \geq t)} \\ &= \frac{1}{P(T \geq t)} \underbrace{\lim_{\epsilon \to 0} \frac{F_T(t+\epsilon) - F_T(t)}{\epsilon}}_{=\frac{d}{dt}F_T(t)} = \frac{f_T(t)}{S_T(t)}. \end{aligned}$$

We can therefore describe the distribution of T by using either $f_T(t)$, $F_T(t)$, $S_T(t)$, or $h_T(t)$. If we know one of these functions, we can retrieve the other three.

Example 5.2. In this example, we find the hazard function for $T \sim \text{Weibull}(\beta, \eta)$. Recall that the pdf of T is

$$f_T(t) = \begin{cases} \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The cdf of T is

$$F_T(t) = \begin{cases} 0, & t \leq 0 \\ 1 - e^{-(t/\eta)^\beta}, & t > 0. \end{cases}$$

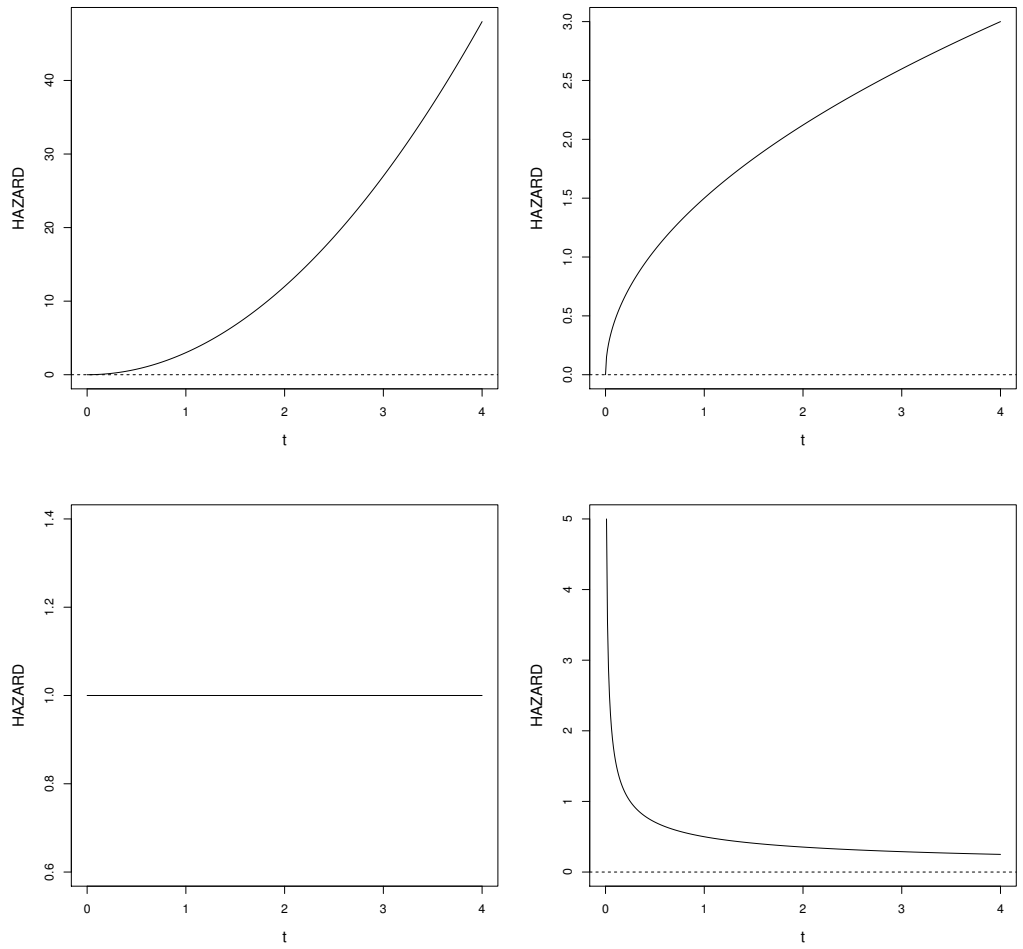

Figure 5.3: Weibull hazard functions with $\eta = 1$. Upper left: $\beta = 3$. Upper right: $\beta = 1.5$. Lower left: $\beta = 1$. Lower right: $\beta = 0.5$.

The survivor function of T is

$$S_T(t) = 1 - F_T(t) = \begin{cases} 1, & t \leq 0 \\ e^{-(t/\eta)^\beta}, & t > 0. \end{cases}$$

Therefore, the hazard function of T is

$$h_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}}{e^{-(t/\eta)^\beta}} = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}.$$

This function describes how the rate of failure changes over time (under the Weibull model assumption).

Interpretation: Plots of Weibull hazard functions are given in Figure 5.3. It is easy to show that for a Weibull distribution:

- $h_T(t)$ is **increasing** if $\beta > 1$ (wear out; population gets weaker)
- $h_T(t)$ is **constant** if $\beta = 1$ (random failures; exponential distribution)
- $h_T(t)$ is **decreasing** if $\beta < 1$ (infant mortality; population gets stronger).

5.3 Example: Weibull analysis

Example 5.3. The data below are times, denoted by T (measured in months), to the first failure for 20 electric carts used for internal delivery and transportation in a large manufacturing facility.

3.9	4.2	5.4	6.5	7.0	8.8	9.2	11.4	14.3	15.1
15.3	15.5	17.9	18.0	19.0	19.0	23.9	24.8	26.0	34.2

In this example, we will assume a Weibull(β, η) model for

$$T = \text{ time to cart failure (in months).}$$

Because the model parameters β and η are not given to us, our first task is to estimate them using the data above. We would like to find the values of β and η that "most closely agree" with the data. To do this, we form the **likelihood function**

$$\begin{aligned} L(\beta, \eta) = \prod_{i=1}^{20} f_T(t_i) &= \prod_{i=1}^{20} \frac{\beta}{\eta}\left(\frac{t_i}{\eta}\right)^{\beta-1} e^{-(t_i/\eta)^\beta} \\ &= \left(\frac{\beta}{\eta^\beta}\right)^{20}\left(\prod_{i=1}^{20} t_i\right)^{\beta-1} e^{-\sum_{i=1}^{20}(t_i/\eta)^\beta}, \end{aligned}$$

where $t_1, t_2, ..., t_{20}$ are the 20 times above. The values of β and η that "most closely agree" with the data are the values of β and η that **maximize** $L(\beta, \eta)$.

Definition: Let $\widehat{\beta}$ and $\widehat{\eta}$ denote the values of β and η, respectively, that maximize $L(\beta, \eta)$. We call $\widehat{\beta}$ and $\widehat{\eta}$ the **maximum likelihood estimates** of β and η.

Calculation: We can use R to find the maximum likelihood estimates:

```
> cart.data = c(3.9,4.2,5.4,6.5,7.0,8.8,9.2,11.4,14.3,15.1,15.3,15.5,17.9,
    18.0,19.0,19.0,23.9,24.8,26.0,34.2)
> fitdist(cart.data,"weibull")
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
       estimate Std. Error
shape  1.988746  0.3503796
scale 16.935158  2.0081780
```

For the cart data, the maximum likelihood estimates of β and η are

$$\begin{aligned} \widehat{\beta} &\approx 1.99 \\ \widehat{\eta} &\approx 16.94. \end{aligned}$$

The $\widehat{\beta} \approx 1.99$ estimate suggests that there is "wear out" taking place among the carts; that is, the population of carts gets weaker as time passes (see the estimated hazard function on the next page).

Note: I have plotted the estimated pdf, the estimated cdf, the estimated survivor function, and the estimated hazard function in Figure 5.4. We use the term "estimated," because these functions are constructed using the estimates $\widehat{\beta} \approx 1.99$ and $\widehat{\eta} \approx 16.94$.

(a) Using the estimated Weibull($\widehat{\beta} \approx 1.99, \widehat{\eta} \approx 16.94$) distribution as a model for cart lifetimes, find the probability that a future cart will "survive" past 20 months.

$$\begin{aligned} P(T > 20) = 1 - P(T \leq 20) &= 1 - F_T(20) \\ &= 1 - [1 - e^{-(20/16.94)^{1.99}}] \approx 0.249. \end{aligned}$$

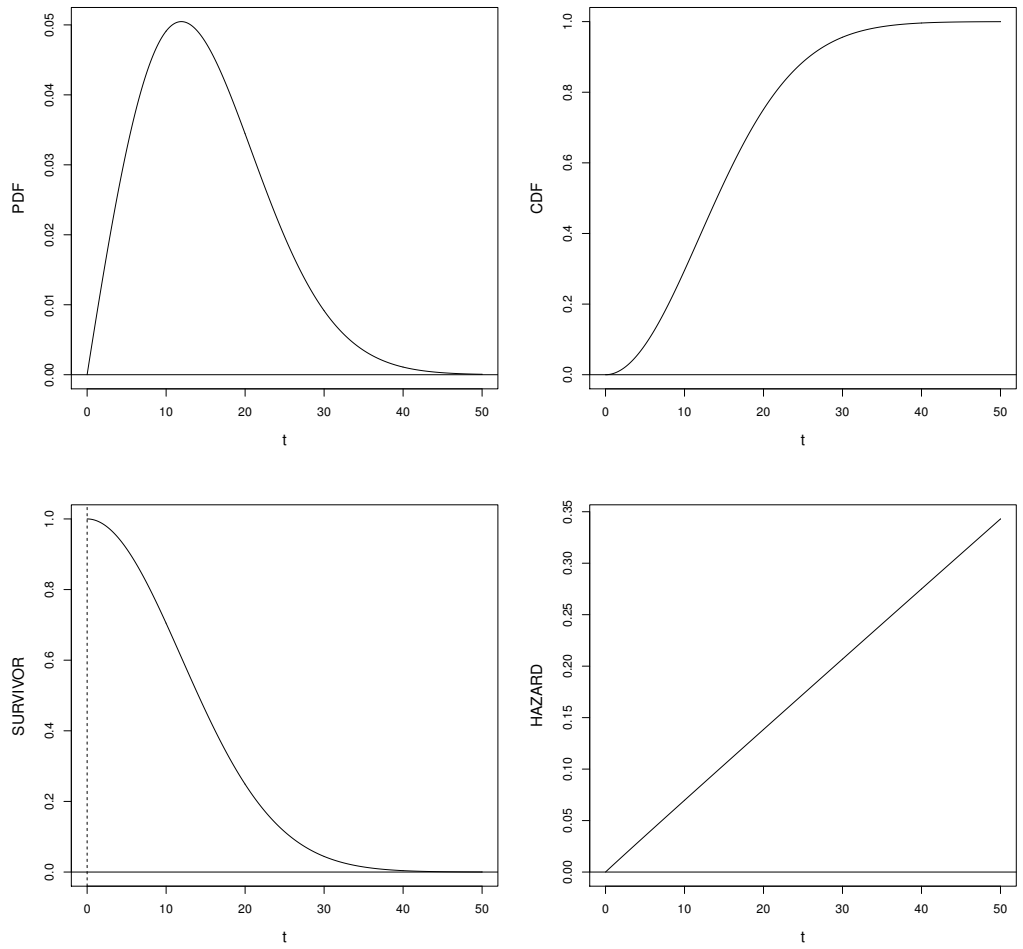

Figure 5.4: Cart data in Example 5.3. Estimated Weibull functions with $\widehat{\beta} = 1.99$ and $\widehat{\eta} \approx 16.94$. Upper left: PDF. Upper right: CDF. Lower left: Survivor function. Lower right: Hazard function.

(b) Use the estimated distribution to find the 90th percentile of the cart lifetimes.

$$F_T(\phi_{0.90}) = 1 - e^{-(\phi_{0.90}/16.94)^{1.99}} \stackrel{\text{set}}{=} 0.90.$$

Solving this equation for $\phi_{0.90}$ gives

$$\phi_{0.90} \approx 25.75 \text{ months}.$$

Only ten percent of the cart lifetimes will exceed this value.

Note: Here is the R code for answering parts (a) and (b) on the last two pages:

```
> 1-pweibull(20,1.99,16.94) ## Part (a)
[1] 0.248679
> qweibull(0.9,1.99,16.94) ## Part (b)
[1] 25.75914
```

5.4 Quantile-quantile plots

Terminology: A **quantile-quantile plot** (**qq plot**) is a graphical display that can help assess the appropriateness of a model (distribution). 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 (ordered) theoretical quantiles from the distribution (model) assumed for the observed data.

Our intuition should suggest the following:

- If the observed data "agree" with the distribution's theoretical quantiles, then the qq plot should look like a **straight line** (the distribution is a good choice).
- If the observed data do not "agree" with the theoretical quantiles, then the qq plot should have **curvature** in it (the distribution is not a good choice).

Important: When you interpret qq plots, you are looking for **general agreement**. The observed data will never line up perfectly with the model's quantiles (due to natural variability). In other words, don't be "too picky" when interpreting these plots, especially with small sample sizes (like $n = 20$).

Cart data: I constructed the Weibull qq plot for the cart lifetime data in Example 5.3; see Figure 5.5 on the next page.

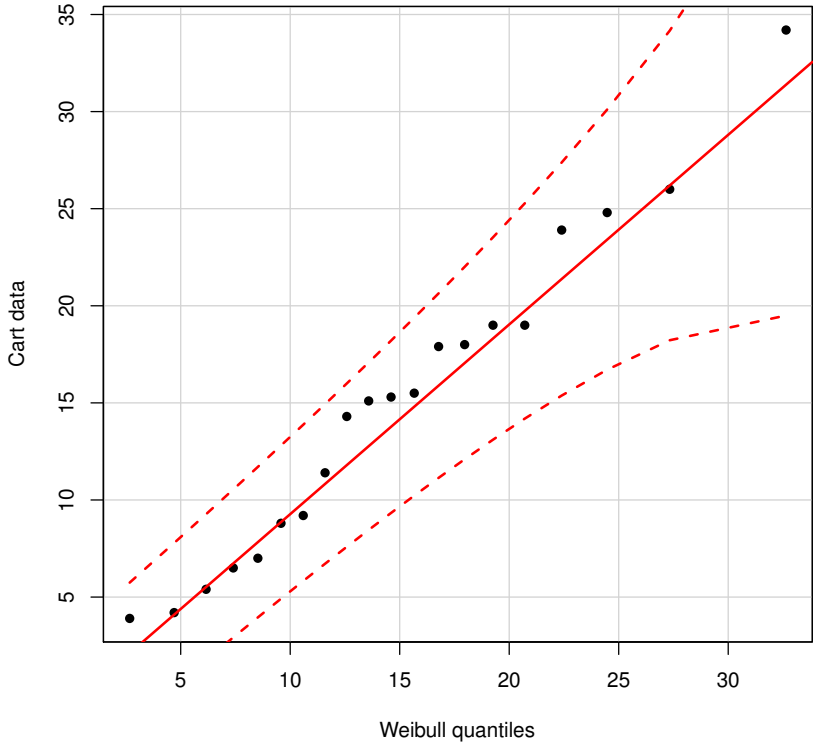

Figure 5.5: Cart data: **Weibull** qq plot. The observed data are plotted versus the theoretical quantiles from a Weibull distribution with $\widehat{\beta} \approx 1.99$ and $\widehat{\eta} \approx 16.94$.

- There is a **general agreement** with the observed data and the quantiles from the Weibull distribution. This suggests that the Weibull model is reasonable for the cart data.
- The straight line is formed from the 25th and 75th percentiles of the observed data and the assumed model (here, a Weibull model with $\widehat{\beta} \approx 1.99$ and $\widehat{\eta} \approx 16.94$).
- The bands about the line can be used to
  - get an idea of the variability "allowed" in the plot. If all of the data fall within the bands, then there is no reason to suspect the model.
  - detect **outlier** observations (i.e., observations that are grossly inconsistent with the assumed model).

6 Statistical Inference

6.1 Populations and samples

Overview: We now focus on **statistical inference**. This deals with making (probabilistic) statements about a population of individuals based on information that is contained in a sample taken from the population.

Example 6.1. Suppose we are studying the performance of lithium batteries used in a certain calculator. We would like to learn about the **lifetime** of these batteries so we can place a limited warranty on them in the future. Because this type of battery has not been used in this calculator before, no one (except the Oracle) can tell us the distribution of Y, the battery's lifetime. In fact, not only is the distribution not known, but all parameters which index this distribution aren't known either.

Terminology: A **population** refers to the entire group of "individuals" (e.g., parts, people, batteries, etc.) about which we would like to make a statement (e.g., proportion defective, median IQ score, mean lifetime, etc.).

- It is generally accepted that the entire population can not be measured. It is too large and/or it would be too time consuming to do so.
- To draw inferences (make statements) about a population, we therefore observe a **sample** of individuals from the population.
- We will assume that the sample of individuals constitutes a **random sample**. Mathematically, this means that all observations are independent and follow the same probability distribution. Informally, this means that each sample (of the same size) has the same chance of being selected.
- Taking a random sample of individuals is our best hope of obtaining individuals that are "representative" of the entire population.

Notation: We will denote a random sample of observations by

$$Y_1, Y_2, ..., Y_n.$$

That is, Y_1 is the value of Y for the first individual in the sample, Y_2 is the value of Y for the second individual in the sample, and so on. The **sample size** tells us how many individuals are in the sample and is denoted by n. We refer to the set of observations $Y_1, Y_2, ..., Y_n$ generically as **data**. Lower case notation $y_1, y_2, ..., y_n$ is used when citing numerical values.

Example 6.1 (continued). Consider the following random sample of $n = 50$ battery lifetimes $y_1, y_2, ..., y_{50}$ measured in hours:

4285	2066	2584	1009	318	1429	981	1402	1137	414
564	604	14	4152	737	852	1560	1786	520	396
1278	209	349	478	3032	1461	701	1406	261	83
205	602	3770	726	3894	2662	497	35	2778	1379
3920	1379	99	510	582	308	3367	99	373	454

In Figure 6.1, we display a **histogram** and **boxplot** of the battery lifetime data. We see that the (empirical) distribution of the battery lifetimes is skewed towards the high side.

- Which continuous probability distribution seems to display the same type of pattern that we see in the histogram?
- An exponential(λ) model seems reasonable here (based on the histogram shape). What is λ?
- In this example, λ is called a (population) **parameter**. It describes the distribution which is used to model the entire population of batteries.
- In general, (population) parameters which index probability distributions (like the exponential) are unknown.

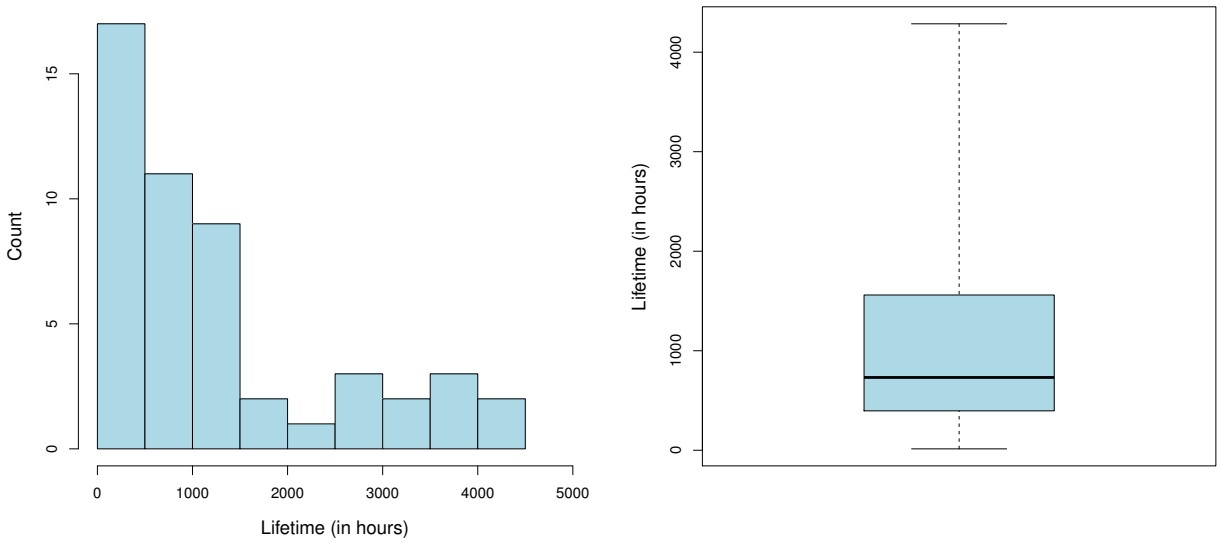

Figure 6.1: Histogram (left) and boxplot (right) of the battery lifetime data (measured in hours) in Example 6.1.

- All the probability distributions we have discussed so far are meant to describe population-level behavior.

6.2 Parameters and statistics

Terminology: A **parameter** is a numerical quantity that describes a population. In general, population parameters are unknown. Some very common examples are:

$$\begin{aligned} \mu &= \text{population mean} \\ \sigma^2 &= \text{population variance} \\ p &= \text{population proportion.} \end{aligned}$$

Connection: All of the probability distributions that we talked about in Chapters 3-5 were indexed by population (model) parameters.

For example,

- the $\mathcal{N}(\mu, \sigma^2)$ distribution is indexed by two parameters, the population mean μ and the population variance σ^2.
- the Poisson(λ) distribution is indexed by one parameter, the population mean λ.
- the Weibull(β, η) distribution is indexed by two parameters, the shape parameter β and the scale parameter η.
- the $b(n, p)$ distribution is indexed by one parameter, the population proportion of successes p.

Terminology: Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from a population. The **sample mean** is

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

The **sample variance** is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

The **sample standard deviation** is the positive square root of the sample variance; i.e.,

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

Important: Unlike their population analogues (which are unknown), these quantities can be computed from the sample $Y_1, Y_2, ..., Y_n$.

Terminology: A **statistic** is a numerical quantity that can be calculated from a sample of data. Some very common examples are:

$$\begin{aligned} \overline{Y} &= \text{sample mean} \\ S^2 &= \text{sample variance} \\ \widehat{p} &= \text{sample proportion.} \end{aligned}$$

For example, with the battery lifetime data (a random sample of $n = 50$ lifetimes),

$$\begin{aligned} \overline{y} &= 1274.14 \text{ hours} \\ s^2 &= 1505156 \text{ (hours)}^2 \\ s &\approx 1226.85 \text{ hours.} \end{aligned}$$

```
> mean(battery) ## sample mean
[1] 1274.14
> var(battery) ## sample variance
[1] 1505156
> sd(battery) ## sample standard deviation
[1] 1226.848
```

- $\overline{y} = 1274.14$ is an **estimate** of the population mean μ.
- $s^2 = 1505156$ is an **estimate** of the population variance σ^2.
- $s = 1226.848$ is an **estimate** of the population standard deviation σ.

Summary: The table below succinctly summarizes the differences between a population and a sample (a parameter and a statistic):

Group of individuals	Numerical quantity	Status
Population (Not observed)	Parameter	Unknown
Sample (Observed)	Statistic	Calculated from sample data

Statistical inference deals with making (probabilistic) statements about a population of individuals based on information that is contained in a sample taken from the population. We do this by

(a) **estimating** unknown population parameters with sample statistics

(b) quantifying the **uncertainty** (variability) that arises in the estimation process.

6.3 Point estimators and sampling distributions

Remark: To keep our discussion as general as possible (as the material in this subsection can be applied to many situations), we will let θ denote a **population parameter**.

- For example, θ could denote a population mean, a population variance, a population proportion, a Weibull or gamma model parameter, etc. It could also denote a parameter in a regression context (Chapters 10-12).

Terminology: A **point estimator** $\widehat{\theta}$ is a statistic that is used to estimate a population parameter θ. Common examples of point estimators are:

$$\begin{aligned} \overline{Y} &\longrightarrow \text{ a point estimator for } \mu \text{ (population mean)} \\ S^2 &\longrightarrow \text{ a point estimator for } \sigma^2 \text{ (population variance)} \\ S &\longrightarrow \text{ a point estimator for } \sigma \text{ (population standard deviation).} \end{aligned}$$

Critical point: A point estimator $\widehat{\theta}$ is a statistic, so it depends on the sample of data $Y_1, Y_2, ..., Y_n$.

- The data $Y_1, Y_2, ..., Y_n$ come from the sampling process; e.g., different random samples will yield different data sets $Y_1, Y_2, ..., Y_n$.
- In this light, because the sample values $Y_1, Y_2, ..., Y_n$ will vary from sample to sample, the value of $\widehat{\theta}$ will too. It therefore makes perfect sense to think about the **distribution** of $\widehat{\theta}$ itself.

Terminology: The distribution of an estimator $\widehat{\theta}$ is called its **sampling distribution**. A sampling distribution describes how the estimator $\widehat{\theta}$ varies in repeated sampling.

Terminology: We say that $\widehat{\theta}$ is an **unbiased estimator** of θ if and only if

$$E(\widehat{\theta}) = \theta.$$

In other words, the mean of the sampling distribution of $\widehat{\theta}$ is equal to θ. Unbiasedness is a characteristic describing the center of a sampling distribution. This deals with **accuracy**.

Result: Mathematics shows that when $Y_1, Y_2, ..., Y_n$ is a random sample,

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

That is, $\overline{Y}$ and S^2 are unbiased estimators of their population analogues.

Goal: Not only do we desire to use point estimators $\widehat{\theta}$ which are unbiased, but we would also like for them to have small variability. In other words, when $\widehat{\theta}$ "misses" θ, we would like for it to "not miss by much." This deals with **precision**.

Main point: Accuracy and precision are the two main mathematical characteristics that arise when evaluating the quality of a point estimator $\widehat{\theta}$. We desire point estimators $\widehat{\theta}$ which are **unbiased** (perfectly accurate) and have **small variance** (highly precise).

Terminology: The **standard error** of a point estimator $\widehat{\theta}$ is equal to

$$\text{se}(\widehat{\theta}) = \sqrt{\text{var}(\widehat{\theta})}.$$

In other words, the standard error is equal to the standard deviation of the sampling distribution of $\widehat{\theta}$. An estimator's standard error measures the amount of variability in the point estimator $\widehat{\theta}$. Therefore,

$$\text{smaller se}(\widehat{\theta}) \iff \widehat{\theta} \text{ more precise.}$$

Illustration: In Example 5.3 (last chapter), we fit a Weibull model to the cart data:

```
> fitdist(cart.data,"weibull")
Parameters:
       estimate Std. Error
shape  1.988746  0.3503796
scale 16.935158  2.0081780
```

- The values $\widehat{\beta} = 1.99$ and $\widehat{\eta} = 16.94$ are point estimates of β and η, respectively.
- The associated standard errors (0.35 and 2.01, respectively) are provided.
- Point estimates and standard errors are used to construct **confidence intervals** (which we will start in the next chapter).

6.4 Sampling distribution of $\overline{Y}$

Result 1: Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. The sample mean $\overline{Y}$ has the following **sampling distribution**:

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

- This result reminds us that

$$E(\overline{Y}) = \mu.$$

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

- This result also shows that the **standard error** of $\overline{Y}$ (as a point estimator) is

$$\text{se}(\overline{Y}) = \sqrt{\text{var}(\overline{Y})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Example 6.2. In Example 4.6 (pp 53), we examined the distribution of

$$Y = \text{ time (in seconds) to react to brake lights during in-traffic driving.}$$

We assumed that $Y \sim \mathcal{N}(\mu = 1.5, \sigma^2 = 0.16)$. We call this the **population distribution**, because it describes the distribution of values of Y for all individuals in the population (here, in-traffic drivers).

(a) Suppose that we take a random sample of $n = 5$ drivers from the population with times $Y_1, Y_2, ..., Y_5$. What is the distribution of the sample mean $\overline{Y}$?

SOLUTION. If the sample size is $n = 5$, then with $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \overline{Y} \sim \mathcal{N}(1.5, 0.032).$$

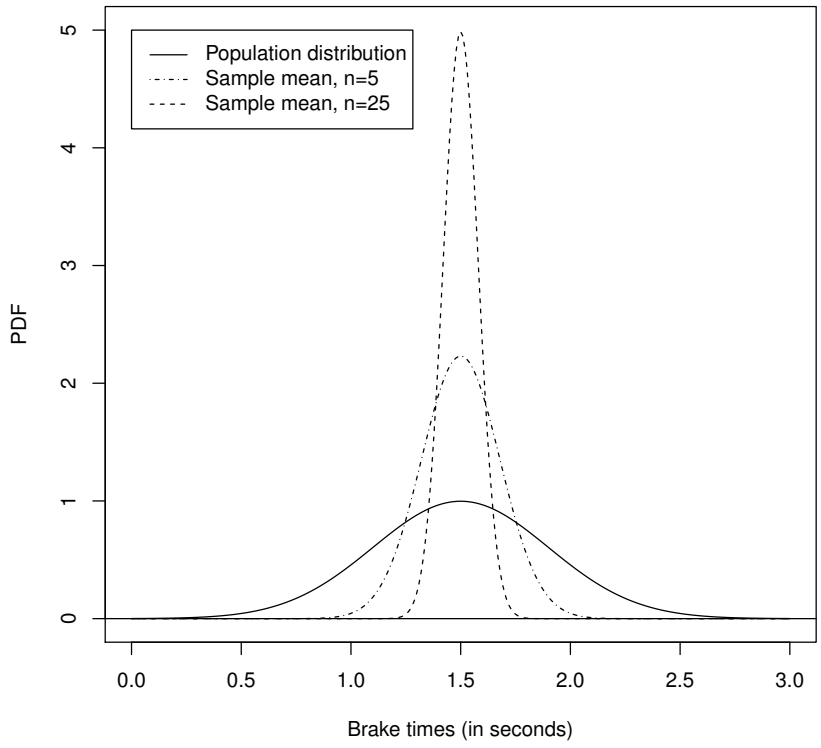

Figure 6.2: Braking time example. Population distribution: $Y \sim \mathcal{N}(\mu = 1.5, \sigma^2 = 0.16)$. Also depicted are the sampling distributions of $\overline{Y}$ when $n = 5$ and $n = 25$.

This distribution describes the values of $\overline{Y}$ we would expect to see in repeated sampling, that is, if we repeatedly sampled $n = 5$ individuals from this population of in-traffic drivers and calculated the sample mean $\overline{Y}$ each time.

(b) Suppose that we take a random sample of $n = 25$ drivers from the population with times $Y_1, Y_2, ..., Y_{25}$. What is the distribution of the sample mean $\overline{Y}$?

SOLUTION. If the sample size is $n = 25$, then with $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \implies \overline{Y} \sim \mathcal{N}(1.5, 0.0064).$$

The sampling distribution of $\overline{Y}$ when $n = 5$ and when $n = 25$ is shown in Figure 6.2.

6.5 Central Limit Theorem

Result 2: Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from a population distribution with mean μ and variance σ^2 (not necessarily a normal distribution). 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)**.

- Result 1 guarantees that when the underlying population distribution is $\mathcal{N}(\mu, \sigma^2)$, the sample mean

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

- The Central Limit Theorem (Result 2) says that even if the population distribution is not normal (Guassian), the sampling distribution of the sample mean $\overline{Y}$ will be **approximately** normal (Gaussian) when the sample size is sufficiently large.

Example 6.3. The time to death for rats injected with a toxic substance, denoted by Y (measured in days), follows an exponential distribution with $\lambda = 1/5$. That is,

$$Y \sim \text{exponential}(\lambda = 1/5).$$

This is the **population distribution**. It describes the time to death for all individual rats in the population.

- In Figure 6.3, I have shown the exponential($\lambda = 1/5$) population distribution (solid curve). I have also depicted the theoretical sampling distributions of $\overline{Y}$ when $n = 5$ and when $n = 25$.
- Notice how the sampling distribution of $\overline{Y}$ begins to (albeit distantly) resemble a normal distribution when $n = 5$. When $n = 25$, the sampling distribution of $\overline{Y}$ looks very much to be normal (Gaussian).

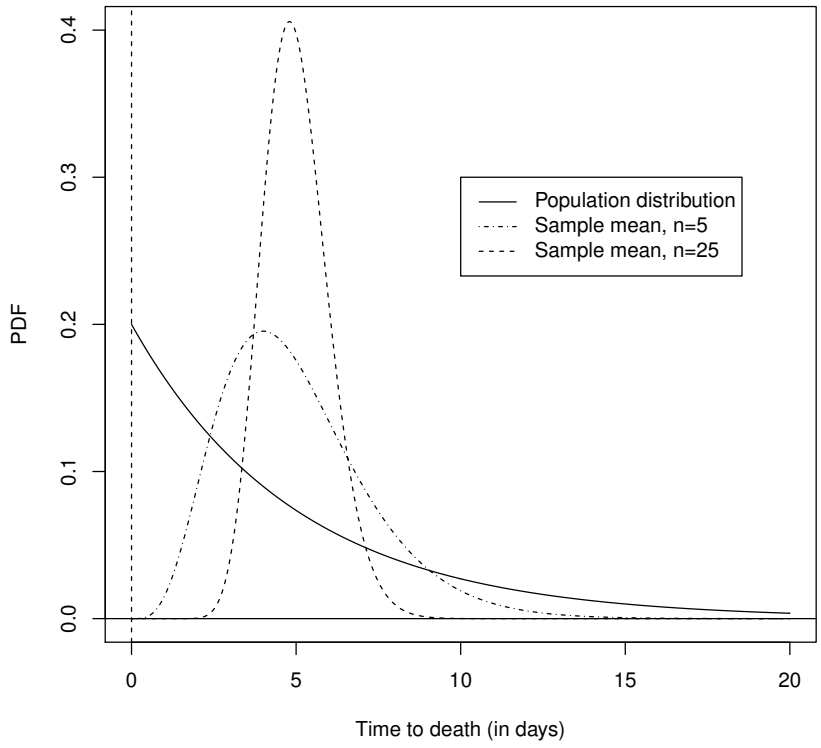

Figure 6.3: Rat death times. Population distribution: $Y \sim$ exponential($\lambda = 1/5$). Also depicted are the sampling distributions of $\overline{Y}$ when $n = 5$ and $n = 25$.

- This is a consequence of the CLT. The larger the sample size n, the better a normal (Gaussian) distribution represents the sampling distribution of $\overline{Y}$.

Example 6.4. When a batch of a chemical product is prepared, the amount of an impurity in the batch (measured in grams) is a random variable Y with

$$\begin{aligned} \mu &= 4.0\text{g} \\ \sigma^2 &= (1.5\text{g})^2. \end{aligned}$$

Suppose that $n = 50$ batches are prepared independently. What is the probability that the sample mean impurity amount $\overline{Y}$ will be greater than 4.2 grams?

SOLUTION. With $n = 50$, $\mu = 4$, and $\sigma^2 = (1.5)^2$, the CLT says that

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right) \implies \overline{Y} \sim \mathcal{AN}(4, 0.045).$$

Therefore,

$$\begin{aligned} P(\overline{Y} > 4.2) &= 1 - P(\overline{Y} \leq 4.2) \\ &\approx \texttt{1-pnorm(4.2,4,sqrt(0.045))} \\ &= \texttt{0.1728893}. \end{aligned}$$

Important: Note that in making this (approximate) probability calculation, we never made an assumption about the underlying population distribution. A sample of size $n = 50$ is probably large enough for the Central Limit Theorem to "take effect" regardless of what the population distribution is.

6.6 The t distribution

Result 3: Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Recall that Result 1 says the sample mean $\overline{Y}$ has the following sampling distribution:

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

If we standardize $\overline{Y}$ (see the last result on page 55), we obtain

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

If we replace the population standard deviation σ with the sample standard deviation S, we get a new sampling distribution:

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

a t **distribution** with degrees of freedom $\nu = n - 1$.

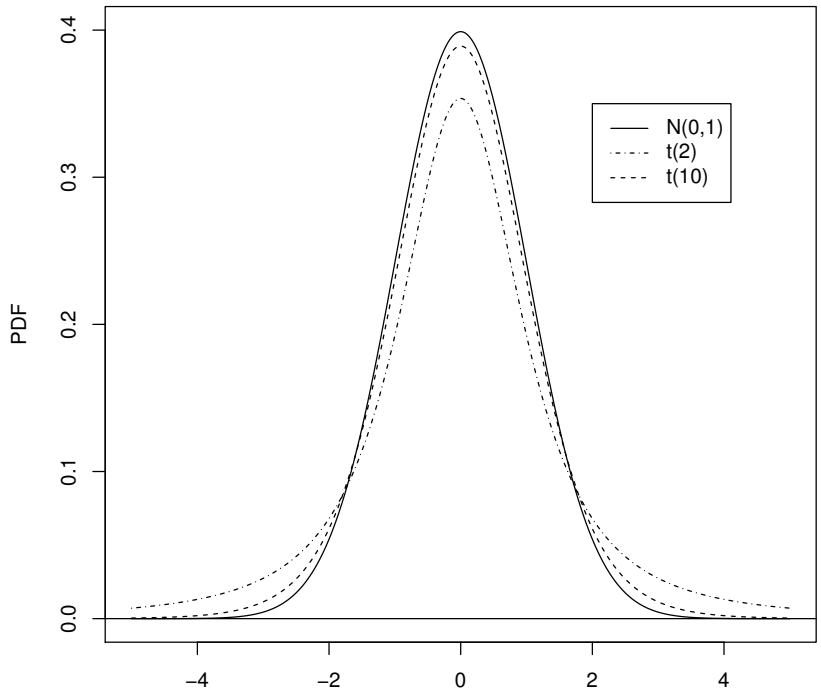

Figure 6.4: Probability density functions of $\mathcal{N}(0,1)$, $t(2)$, and $t(10)$.

Facts: The t pdf has the following characteristics:

- It is continuous and symmetric about 0 (just like the standard normal pdf).
- It is indexed by a value ν called the **degrees of freedom**. In practice, ν is often an integer (related to the sample size).
- As $\nu \to \infty$, $t(\nu) \to \mathcal{N}(0,1)$; thus, when ν becomes larger, the $t(\nu)$ pdf and the $\mathcal{N}(0,1)$ pdf look more alike.
- When compared to the standard normal pdf, the t pdf, in general, is less peaked and has more probability (area) in the tails.

Remark: The t pdf formula is complicated and is unnecessary for our purposes. R will compute probabilities and quantiles from the t distribution.

t **R CODE:** Suppose that $T \sim t(\nu)$.

$F_T(t) = P(T \leq t)$	ϕ_p
`pt(t,`ν`)`	`qt(p,`ν`)`

Example 6.5. Hollow pipes are to be used in an electrical wiring project. In testing "1-inch" pipes, the data below were collected by a design engineer. The data are measurements of Y, the **outside diameter** of this type of pipe (measured in inches). These $n = 25$ pipes were randomly selected and measured−all in the same location.

1.296	1.320	1.311	1.298	1.315
1.305	1.278	1.294	1.311	1.290
1.284	1.287	1.289	1.292	1.301
1.298	1.287	1.302	1.304	1.301
1.313	1.315	1.306	1.289	1.291

The manufacturers of this pipe claim that the population distribution is normal (Gaussian) and that the mean outside diameter is $\mu = 1.29$ inches. **Under this assumption** (which may or may not be true), calculate the value of

$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}}.$$

SOLUTION. We use R to find the sample mean $\overline{y}$ and the sample standard deviation s:

```
> mean(pipes) ## sample mean
[1] 1.29908
> sd(pipes) ## sample standard deviation
[1] 0.01108272
```

We compute

$$t = \frac{1.299 - 1.29}{0.011/\sqrt{25}} \approx 4.096.$$

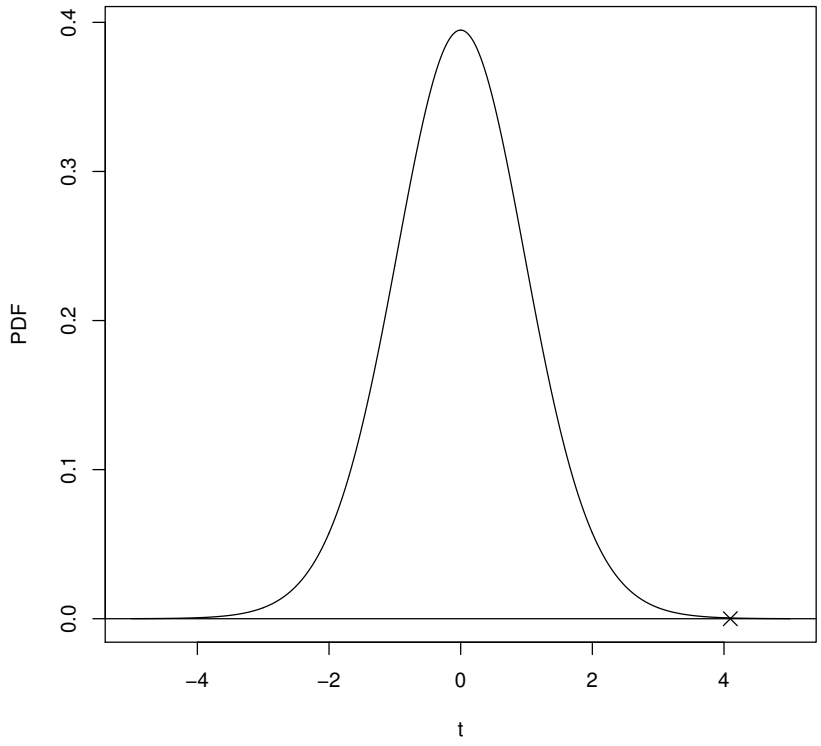

Figure 6.5: $t(24)$ probability density function. An "$\times$" at $t = 4.096$ has been added.

ANALYSIS. If the manufacturer's claim is true (that is, if $\mu = 1.29$ inches), then

$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}}$$

should come from a $t(24)$ distribution. The $t(24)$ pdf is displayed above in Figure 6.5. I placed an "$\times$" at the value $t = 4.096$.

Discussion: Does $t = 4.096$ seem like a value you would expect to see from this distribution? Recall that t was computed under the assumption that $\mu = 1.29$ inches (the manufacturer's claim). Therefore, if the manufacturer's claim is true, we would expect to see a value of t around the center of this distribution.

This isn't what we see here. This value of t is somewhat extreme (way out in the tail) and ultimately looks to be more consistent with a value of μ that is larger than 1.29 inches.

6.7 Normal quantile-quantile plots

Recall: Result 3 says that if $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

An obvious question therefore arises:

> "What if $Y_1, Y_2, ..., Y_n$ are non-normal (i.e., non-Gaussian)? Does the sampling distribution result above still hold?"

Answer: The t distribution result still approximately holds, even if the underlying population distribution is not perfectly normal. The approximation is best when

- the sample size is larger
- the population distribution is more symmetric (not highly skewed).

Because normality (for the population distribution) is not absolutely critical for the t sampling distribution, we say that this sampling distribution is **robust** to the normality assumption.

Note: Robustness is a nice property. Here, it assures us that the underlying assumption of normality is not an absolute requirement for Result 3 to hold. Other sampling distribution results (coming up) are not always robust to normality departures.

Terminology: Just as we used Weibull qq plots to assess the Weibull model assumption in the last chapter, we can use a normal **quantile-quantile (qq) plot** to assess the normal distribution assumption. The plot is constructed as follows:

- On the vertical axis, we plot the observed data, ordered from low to high.
- On the horizontal axis, we plot the (ordered) theoretical quantiles from the distribution (model) assumed for the observed data (here, normal).

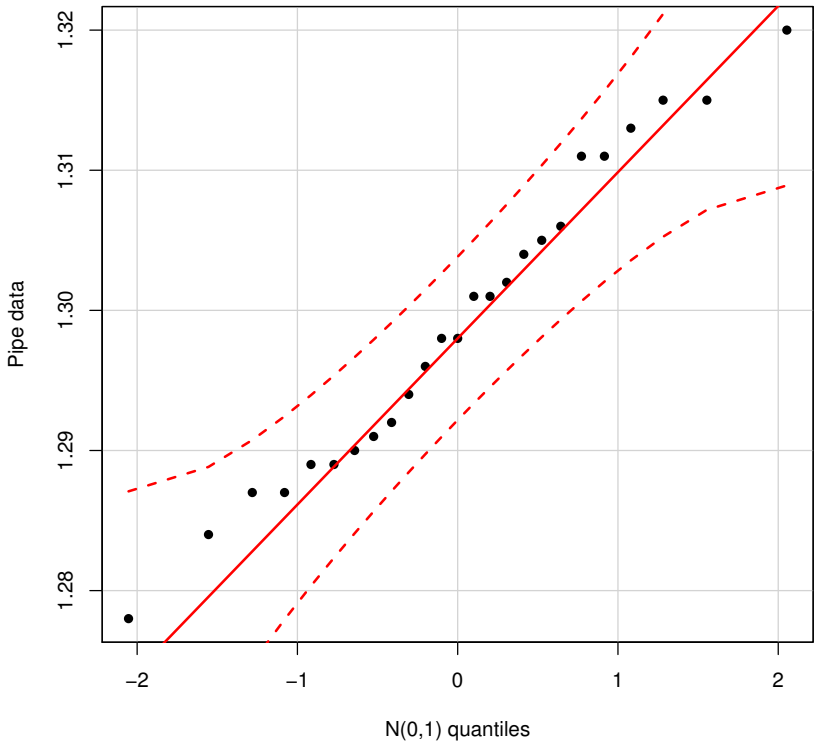

Figure 6.6: Pipe diameter data in Example 6.5. **Normal** qq plot. The observed data are plotted versus the theoretical quantiles from a standard normal distribution. The line added passes through the first and third quartiles.

- In the `qqPlot` function, it is "default" to take the assumed model to be the $\mathcal{N}(0,1)$ distribution (i.e., a normal distribution with mean 0 and variance 1).
- Therefore, we are really comparing the standardized values of the observed data (here, the pipe diameter data) to the $\mathcal{N}(0,1)$ quantiles.
- Linearity in the plot supports the normal assumption. Departures from linearity refute it.

Pipe diameter data: Figure 6.6 shows the normal qq plot for the pipe diameter data in Example 6.5. The plot supports the normal distribution assumption for the population of pipe diameters.

7 One-Sample Inference

Preview: In this chapter, we discuss one-sample inference procedures for three population parameters:

- A population **mean** μ (Section 7.1)
- A population **variance** σ^2 (Section 7.2)
- A population **proportion** p (Section 7.3).

Remember that these are population-level quantities, so they are unknown. Our goal is to use sample information to estimate these quantities.

Relevance: To begin our discussion, suppose that we would like to estimate a **population mean** μ. To do so, suppose we have a random sample $Y_1, Y_2, ..., Y_n$ from a population distribution (e.g., normal, exponential, Weibull, Poisson, etc.). Regardless of what the population distribution is, we know that $\overline{Y}$ is an **unbiased estimator** for μ, that is,

$$E(\overline{Y}) = \mu.$$

However, reporting $\overline{Y}$ alone does not acknowledge that there is **variability** attached to this estimator. For example, in Example 6.5 (pp 81), with the $n = 25$ measured pipes, reporting

$$\overline{y} \approx 1.299 \text{ in}$$

as an estimate of the population mean μ does not account for the fact that

- the 25 pipes measured were drawn randomly from a population of all pipes, and
- different samples would give different sets of pipes (and different values of $\overline{y}$).

In other words, using $\overline{Y}$ only **ignores important information**; namely, how variable the population of pipes is.

Remedy: To address this problem, we therefore pursue the topic of **interval estimation** (also known as **confidence intervals**). The main difference between a point estimate (like $\overline{y} \approx 1.299$) and an interval estimate is that

- a **point estimate** is a "one-shot guess" at the value of the parameter; this ignores the variability in the estimate.
- an **interval estimate** (i.e., **confidence interval**) is an interval of values. It is formed by taking the point estimate and then adjusting it downwards and upwards to account for the point estimate's variability. The end result is an "interval estimate."

7.1 Confidence interval for a population mean μ

Recall: We start our discussion by revisiting Result 3 in the last chapter (pp 79). Recall that if $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then the quantity

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

a t distribution with $n-1$ degrees of freedom.

Goal: We will use this sampling distribution to create an interval estimate (i.e., a **confidence interval**) for the population mean μ.

Notation: We introduce new notation that identifies quantiles from a t distribution with $n-1$ degrees of freedom. Define

$$\begin{aligned} t_{n-1,\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile from } t(n-1) \text{ pdf} \\ -t_{n-1,\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile from } t(n-1) \text{ pdf} \end{aligned}$$

Because the $t(n-1)$ pdf is symmetric about zero, these two quantiles are equal in absolute value (the upper quantile is positive; the lower quantile is negative); see Figure 7.1.

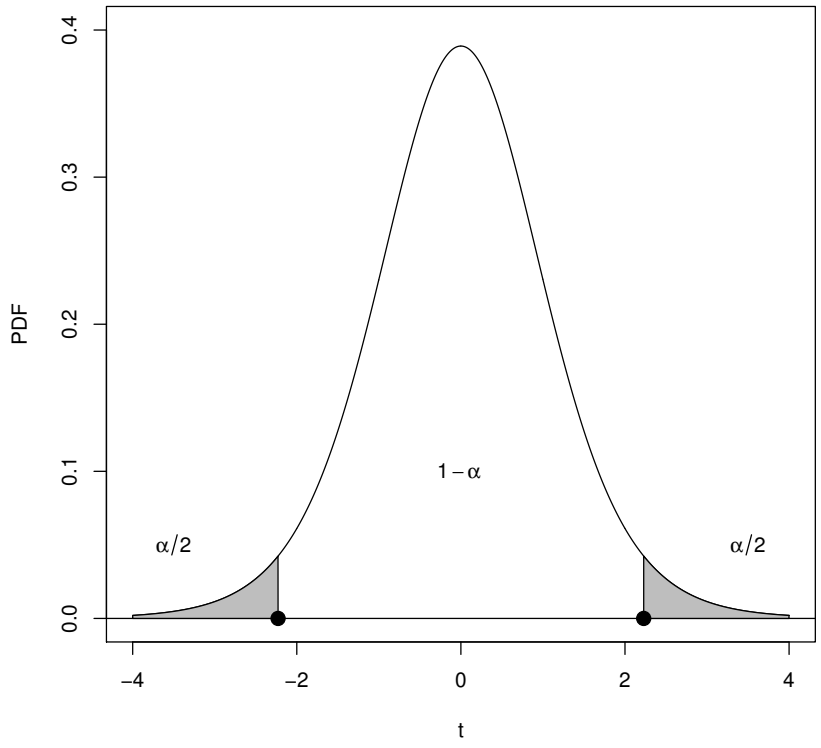

Figure 7.1: A t pdf with $n-1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $t_{n-1,\alpha/2}$ (upper) and $-t_{n-1,\alpha/2}$ (lower), respectively.

Illustration: If $n = 11$ and $\alpha = 0.05$ then

$$\begin{aligned} t_{n-1,\alpha/2} &= t_{10,0.025} \approx 2.23 \\ -t_{n-1,\alpha/2} &= -t_{10,0.025} \approx -2.23 \end{aligned}$$

```
> qt(0.975,10) ## upper 0.025 quantile
[1] 2.228139
> qt(0.025,10) ## lower 0.025 quantile
[1] -2.228139
```

Derivation: In general, for any value of α, $0 < \alpha < 1$, we can write

$$\begin{aligned}
1-\alpha &= P\left(-t_{n-1,\alpha/2} < \frac{\overline{Y}-\mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) \\
&= P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \overline{Y}-\mu < t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\left(t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu-\overline{Y} > -t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\left(\overline{Y}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu > \overline{Y}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\left(\overline{Y}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{Y}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right).
\end{aligned}$$

We call

$$\left(\overline{Y}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{Y}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

a $100(1-\alpha)$ **percent confidence interval** for the population mean μ. This is written more succinctly as

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

Discussion: Before we do an example, let's discuss relevant issues about this confidence interval.

- Note the form of the interval:

$$\underbrace{\text{point estimate}}_{\overline{Y}} \ \pm\ \underbrace{\text{quantile}}_{t_{n-1,\alpha/2}} \ \times\ \underbrace{\text{standard error}}_{S/\sqrt{n}}.$$

  Many confidence intervals we will study follow this same general form.

- Here is how we interpret this interval: We say

  > "We are $100(1-\alpha)$ percent confident that the population mean μ is in this interval."

- Unfortunately, the word "confident" does not mean "probability."
  - The term "confidence" means that if we were able to sample from the population over and over again, each time computing a $100(1-\alpha)$ percent confidence interval

for μ, then $100(1-\alpha)$ percent of the intervals we would compute would contain the population mean μ.

- In other words, "confidence" refers to "long term behavior" of many intervals; not probability for a single interval. Because of this, we call $100(1-\alpha)$ the **confidence level**. Typical confidence levels are
  - 90 percent ($\alpha = 0.10$)
  - 95 percent ($\alpha = 0.05$)
  - 99 percent ($\alpha = 0.01$).
- The **length** of the $100(1-\alpha)$ percent confidence interval

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

is equal to

$$2 \times t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

Therefore, other things being equal,

  - the larger the sample size n, the smaller the interval length.
  - the smaller the population variance σ^2, the smaller the interval length. Recall that S^2 is an **unbiased estimator** for σ^2.
  - the larger the confidence level $100(1-\alpha)$, the larger the interval length.

Remark: Clearly, shorter confidence intervals are preferred. They are more informative. Lower confidence levels will produce shorter intervals; however, you pay a price. You have less confidence that your interval contains μ.

Example 7.1. Acute exposure to cadmium produces respiratory distress and kidney and liver damage (and possibly death). For this reason, the level of airborne cadmium dust and cadmium oxide fume in the air, denoted by Y (measured in milligrams of cadmium per m^3 of air), is closely monitored. A random sample of $n = 35$ measurements from a large factory are given on the next page.

0.044	0.030	0.052	0.044	0.046	0.020	0.066
0.052	0.049	0.030	0.040	0.045	0.039	0.039
0.039	0.057	0.050	0.056	0.061	0.042	0.055
0.037	0.062	0.062	0.070	0.061	0.061	0.058
0.053	0.060	0.047	0.051	0.054	0.042	0.051

Based on past experience, engineers assume a normal population distribution (for the population of all cadmium measurements). Based on the data above, find a 99 percent confidence interval for μ, the population mean level of airborne cadmium.

SOLUTION. The interval is

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

We can use R to calculate the sample mean $\overline{y}$ and the sample standard deviation s:

```
> mean(cadmium) ## sample mean
[1] 0.04928571
> sd(cadmium) ## sample standard deviation
[1] 0.0110894
```

For a 99 percent confidence level; i.e., with $\alpha = 0.01$, we use

$$t_{34,0.01/2} = t_{34,0.005} \approx 2.728.$$

```
> qt(0.995,34) ## upper 0.005 quantile
[1] 2.728394
```

A 99 percent confidence interval for the population mean level of airborne cadmium μ is

$$\overline{y} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \implies 0.049 \pm 2.728\left(\frac{0.011}{\sqrt{35}}\right) \implies (0.044, 0.054) \text{ mg/m}^3.$$

Interpretation: We are 99 percent confident that the population mean level of airborne cadmium μ is between 0.044 and 0.054 mg/m^3.

Note: It is possible to implement the t interval procedure entirely in R using the `t.test` function:

```
> t.test(cadmium,conf.level=0.99)$conf.int
[1] 0.04417147 0.05439996
```

Assumptions: The confidence interval

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

for the population mean μ was created based on the following assumptions:

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

For the confidence interval for the population mean μ to be meaningful, the random sample assumption must be satisfied. However, recall from the last chapter that the t sampling distribution result

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

does still hold approximately even if the underlying population distribution is not perfectly normal. Therefore, the confidence interval (which was derived from this sampling distribution) is also "robust to normality departures."

- This means that even if the population distribution is **mildly** non-normal, the confidence interval formula

  $$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

  can still be used to estimate the population mean μ.
- However, if there is strong evidence that the population distribution is **grossly** non-normal, then you should exercise caution in using this confidence interval, **especially** when the sample size n is small.
- Recall that you can use qq plots to check the normality assumption.

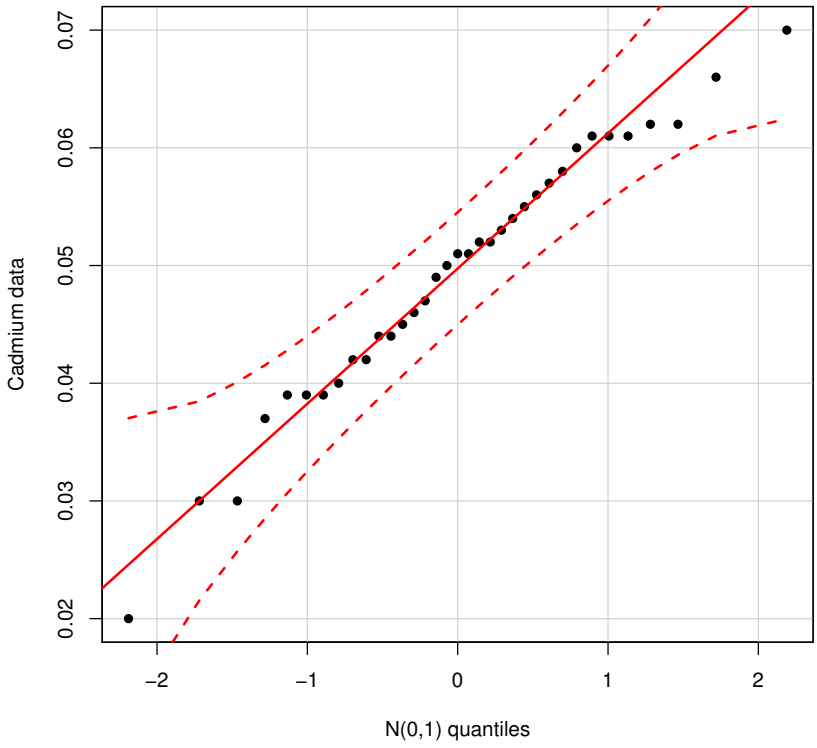

Figure 7.2: Normal qq plot for the cadmium data in Example 7.1. The observed data are plotted versus the theoretical quantiles from a normal distribution. The line added passes through the first and third theoretical quartiles.

Cadmium data: The qq plot for the cadmium data in Figure 7.2 does not reveal any serious departures from the normality assumption. We can feel comfortable reporting

$$(0.044, 0.054) \text{ mg/m}^3$$

as a 99 percent confidence interval for the population mean cadmium level μ.

Remark: As we have just seen, statistical inference procedures are derived from specific assumptions. Going forward, it is important to know what these assumptions are, how critical they are, and how to check them.

7.2 Confidence interval for a population variance σ^2

Relevance: In many situations, we are concerned not with the mean of a population, but with the variance σ^2 instead. If the population variance σ^2 is excessively large, this could point to a potential problem with a manufacturing process, for example, where there is too much variation in the measurements produced. Elsewhere,

- in a laboratory setting, engineers might wish to estimate the variance σ^2 attached to a measurement system (e.g., scale, caliper, etc.).
- in field trials, agronomists are often interested in comparing the variability levels for different cultivars or genetically-altered varieties.
- in clinical trials, physicians are often concerned if there are substantial differences in the variation levels of patient responses at different clinic sites.

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

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

a χ^2 **distribution** with $n-1$ degrees of freedom.

Goal: We will use this sampling distribution to create a **confidence interval** for the population variance σ^2.

Facts: The χ^2 pdf has the following characteristics:

- It is continuous, skewed to the right, and always positive; see Figure 7.3.
- It is indexed by a value ν called the **degrees of freedom**. In practice, ν is often an integer (related to the sample size).
- The χ^2 pdf formula is unnecessary for our purposes. R will compute χ^2 probabilities and quantiles from the χ^2 distribution.

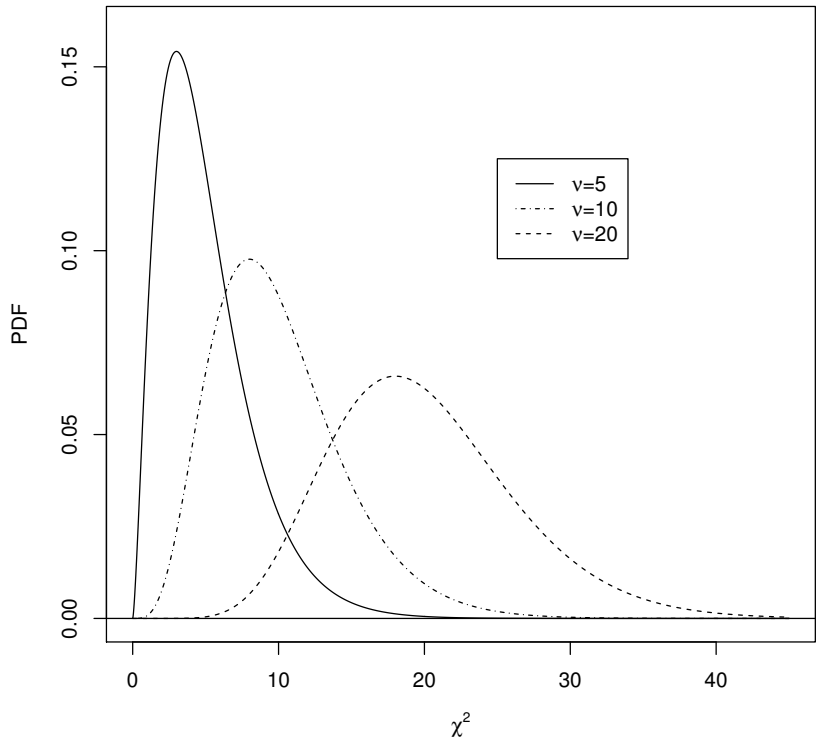

Figure 7.3: χ^2 pdfs with different degrees of freedom.

χ^2 **R CODE:** Suppose that $Q \sim \chi^2(\nu)$.

$F_Q(q) = P(Q \leq q)$	ϕ_p
`pchisq(q,`ν`)`	`qchisq(p,`ν`)`

Notation: We introduce new notation that identifies quantiles from a χ^2 distribution with $n-1$ degrees of freedom. Define

$$\chi^2_{n-1,1-\alpha/2} = \textbf{upper } \alpha/2 \text{ quantile from } \chi^2(n-1) \text{ pdf}$$
$$\chi^2_{n-1,\alpha/2} = \textbf{lower } \alpha/2 \text{ quantile from } \chi^2(n-1) \text{ pdf}$$

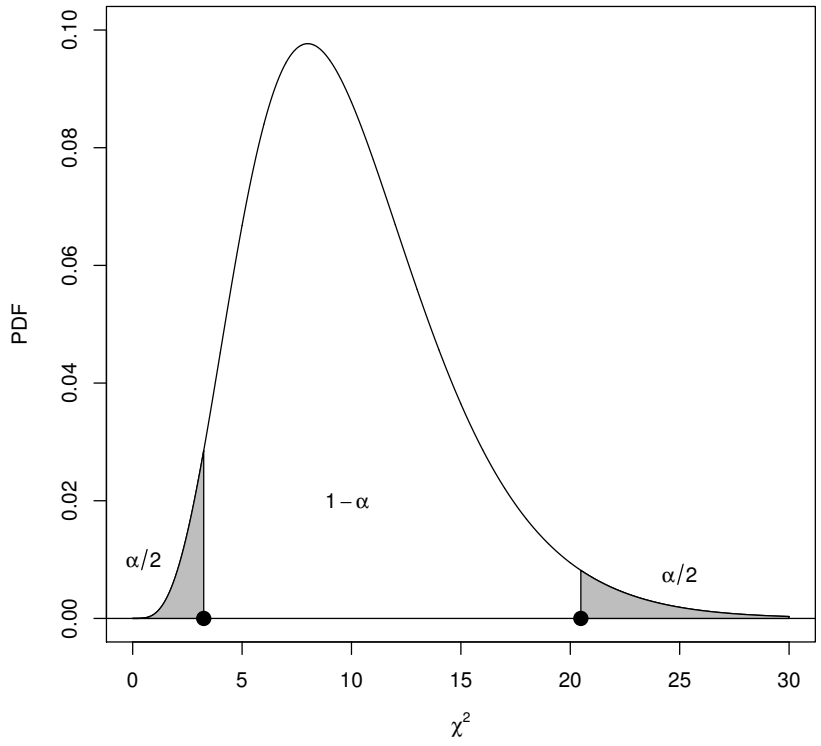

Figure 7.4: A χ^2 pdf with $n-1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $\chi^2_{n-1,1-\alpha/2}$ (upper) and $\chi^2_{n-1,\alpha/2}$ (lower), respectively.

Illustration: If $n = 11$ and $\alpha = 0.05$ then

$$\begin{aligned} \chi^2_{n-1,1-\alpha/2} &= \chi^2_{10,0.975} \approx 20.48 \\ \chi^2_{n-1,\alpha/2} &= \chi^2_{10,0.025} \approx 3.25 \end{aligned}$$

```
> qchisq(0.975,10) ## upper 0.025 quantile
[1] 20.48318
> qchisq(0.025,10) ## lower 0.025 quantile
[1] 3.246973
```

Derivation: In general, for any value of α, $0 < \alpha < 1$, we can write

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) \\ &= P\left(\frac{1}{\chi^2_{n-1,\alpha/2}} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{n-1,1-\alpha/2}}\right) \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right). \end{aligned}$$

This argument shows that

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right)$$

is a $100(1-\alpha)$ **percent confidence interval** for the population variance σ^2. We interpret the interval in the same way:

> "We are $100(1-\alpha)$ percent confident that the population variance σ^2 is in this interval."

Note: A $100(1-\alpha)$ percent confidence interval for the **population standard deviation** σ arises from simply taking the square root of the endpoints of the σ^2 interval.

- That is,
$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}\right)$$
is a $100(1-\alpha)$ percent confidence interval for the population standard deviation σ.
- In practice, this interval may be preferred over the σ^2 interval, because standard deviation is a measure of variability in terms of the original units (e.g., dollars, inches, days, etc.).
- The variance is measured in squared units (e.g., dollars^2, in^2, days^2, etc.) and is in general harder to interpret.

Example 7.2. Industrial engineers at IKEA observed a random sample of $n = 36$ rivet-head screws used in the Billy Bookcase system. The observed diameters of the top of the screws (measured in cm) are given below:

1.206	1.190	1.200	1.195	1.201	1.200	1.198	1.196	1.195	1.202	1.203	1.210
1.206	1.193	1.207	1.201	1.199	1.200	1.199	1.204	1.194	1.203	1.194	1.199
1.203	1.200	1.197	1.208	1.199	1.205	1.199	1.204	1.202	1.196	1.211	1.204

The IKEA manufactured specifications dictate that the population standard deviation diameter for these screws should be **no larger than** $\sigma = 0.003$. Otherwise, there is too much variability in the screws (which could lead to difficulty in construction and hence customer dissatisfaction). Based on the data above, find a 95 percent confidence interval for the population standard deviation σ.

SOLUTION. We first calculate a 95 percent confidence interval for the population variance σ^2 using

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right).$$

There is no internal function in R to calculate this interval, so I wrote one:

```
var.interval = function(data,conf.level=0.95){
    df = length(data)-1
    chi.lower = qchisq((1-conf.level)/2,df)
    chi.upper = qchisq((1+conf.level)/2,df)
    s2 = var(data)
    c(df*s2/chi.upper,df*s2/chi.lower)
    }
```

With this function, I calculated

```
> var.interval(diameters)
[1] 1.545590e-05 3.997717e-05
```

Interpretation: We are 95 percent confident that the population variance σ^2 for the screw diameters is between 0.0000155 and 0.0000400 cm^2.

A 95 percent confidence interval for the population standard deviation σ (which is what we originally wanted) is calculated here:

```
> sd.interval = sqrt(var.interval(diameters))
> sd.interval
[1] 0.003931399 0.006322751
```

Interpretation: We are 95 percent confident that the population standard deviation σ for the screw diameters is between 0.0039 and 0.0063 cm. This interval suggests that the population standard deviation is larger than 0.003 cm, which indicates that there is excessive variability in the diameters of the screws.

Assumptions: The confidence interval

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right)$$

for the population variance σ^2 was created based on the following assumptions:

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

For the confidence interval for the population variance σ^2 to be meaningful, the random sample assumption must be satisfied.

Warning: Unlike the t confidence interval for a population mean μ, the χ^2 interval for a population variance σ^2 **is not robust** to normality departures. This is true because the sampling distribution

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

depends critically on the $\mathcal{N}(\mu, \sigma^2)$ population distribution assumption.

- If the underlying population distribution is non-normal (non-Guassian), then the confidence interval formulas for σ^2 (and σ) are not to be used.

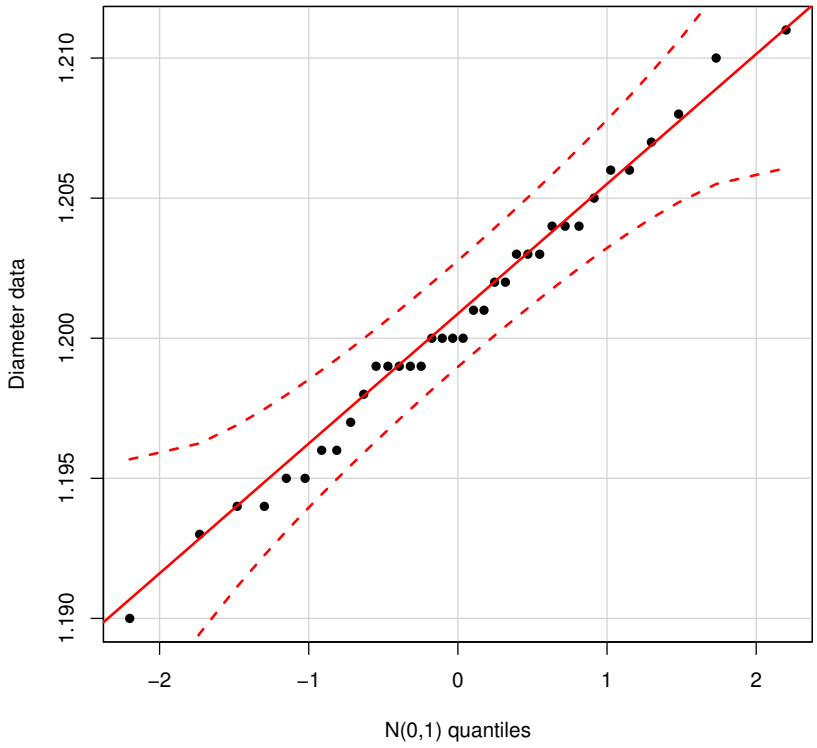

Figure 7.5: Normal qq plot for IKEA screw diameter data in Example 7.2. The observed data are plotted versus the theoretical quantiles from a normal distribution. The line added passes through the first and third theoretical quartiles.

- In the presence of non-normality, these confidence intervals may give results which are misleading (and hence potentially dangerous).
- Therefore, it is very important to check the normality assumption when you construct a confidence interval for a population variance σ^2 (or for a population standard deviation σ).

Screw diameter data: Fortunately, the qq plot for the IKEA screw diameter data in Figure 7.5 shows that there is no cause for concern. The normality assumption for these data is not in doubt.

7.3 Confidence interval for a population proportion p

Situation: We now switch gears and focus on a new population-level parameter: the **population proportion** p. This parameter is relevant when the characteristic we measure on each individual is **binary** (i.e., only 2 outcomes possible). Here are some examples:

$$\begin{aligned} p &= \text{proportion of defective circuit boards} \\ p &= \text{proportion of customers who are "satisfied"} \\ p &= \text{proportion of payments received on time} \\ p &= \text{proportion of HIV positives in SC.} \end{aligned}$$

To start our discussion, we need to recall the **Bernoulli trial** assumptions for each individual in the sample:

1. each individual results in a "success" or a "failure,"
2. the individuals are independent, and
3. the probability of "success" p is the same for every individual.

In our examples above,

$$\begin{aligned} \text{"success"} &\longrightarrow \text{circuit board defective} \\ \text{"success"} &\longrightarrow \text{customer satisfied} \\ \text{"success"} &\longrightarrow \text{payment received on time} \\ \text{"success"} &\longrightarrow \text{HIV positive individual.} \end{aligned}$$

Recall: If the individual success/failure statuses in the sample adhere to the Bernoulli trial assumptions, then

$$Y = \text{ the number of successes out of } n \text{ sampled individuals}$$

follows a binomial distribution, that is, $Y \sim b(n, p)$. The statistical problem at hand is to use the information in Y to **estimate** p.

Note: A natural point estimator for p, the **population proportion**, is

$$\widehat{p} = \frac{Y}{n},$$

the **sample proportion**. This statistic is simply the proportion of "successes" in the sample (out of n individuals).

Properties: Mathematical arguments can be used to show the following results:

$$\begin{aligned} E(\widehat{p}) &= p \\ \text{se}(\widehat{p}) &= \sqrt{\frac{p(1-p)}{n}}. \end{aligned}$$

The first result says that the sample proportion $\widehat{p}$ is an **unbiased estimator** of the population proportion p. The second (standard error) result quantifies the precision of $\widehat{p}$ as an estimator of p.

Result: Knowing the sampling distribution of $\widehat{p}$ is critical if we are going to formalize statistical inference procedures for p. In this situation, we appeal to an approximate result (conferred by the Central Limit Theorem) which says that

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

when the sample size n is large. Standardizing $\widehat{p}$, we get

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AN}(0,1),$$

an approximate **standard normal distribution**.

Notation: We introduce new notation that identifies quantiles from a $\mathcal{N}(0,1)$ distribution. Define

$$\begin{aligned} z_{\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile from } \mathcal{N}(0,1) \text{ pdf} \\ -z_{\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile from } \mathcal{N}(0,1) \text{ pdf} \end{aligned}$$

Because the $\mathcal{N}(0,1)$ pdf is symmetric about zero, these two quantiles are equal in absolute value (the upper quantile is positive; the lower quantile is negative); see Figure 7.6.

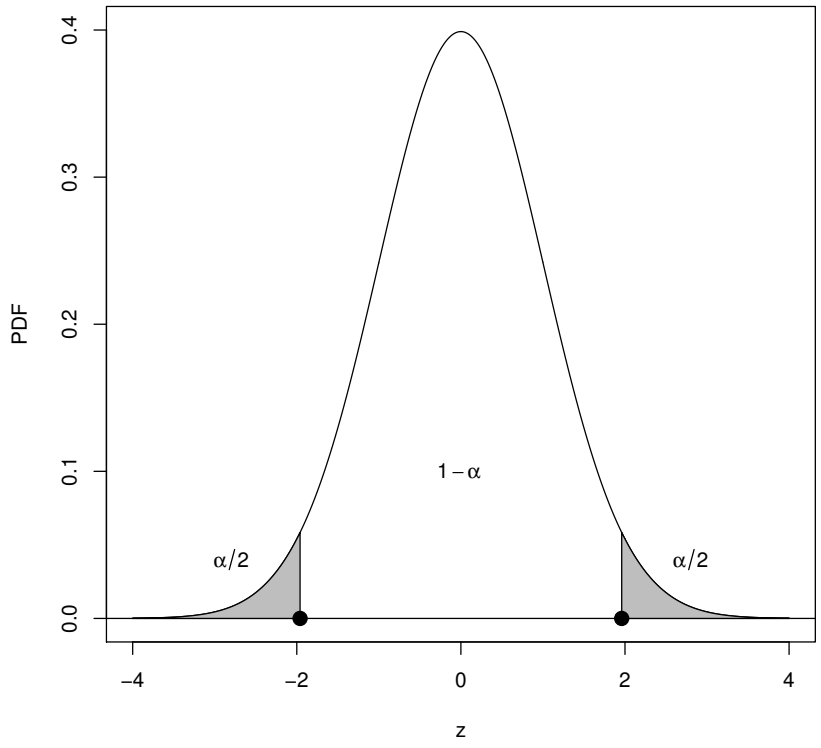

Figure 7.6: The $\mathcal{N}(0,1)$ pdf. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $z_{\alpha/2}$ (upper) and $-z_{\alpha/2}$ (lower), respectively.

Illustration: If $\alpha = 0.05$ then

$$\begin{aligned} z_{\alpha/2} &= z_{0.025} \approx 1.96 \\ -z_{\alpha/2} &= -z_{0.025} \approx -1.96 \end{aligned}$$

```
> qnorm(0.975,0,1) ## upper 0.025 quantile
[1] 1.959964
> qnorm(0.025,0,1) ## lower 0.025 quantile
[1] -1.959964
```

Derivation: In general, for any value of α, $0 < \alpha < 1$, we can write

$$\begin{aligned}
1-\alpha &\approx P\left(-z_{\alpha/2} < \frac{\widehat{p}-p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < \widehat{p}-p < z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \\
&= P\left(z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} > p-\widehat{p} > -z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \\
&= P\left(\widehat{p}+z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} > p > \widehat{p}-z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \\
&= P\left(\widehat{p}-z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < \widehat{p}+z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right).
\end{aligned}$$

We call

$$\left(\widehat{p}-z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p}+z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

a $100(1-\alpha)$ **percent confidence interval** for the population proportion p. This is written more succinctly as

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

- Note the form of the interval:

$$\underbrace{\text{point estimate}}_{\widehat{p}} \ \pm\ \underbrace{\text{quantile}}_{z_{\alpha/2}} \ \times\ \underbrace{\text{standard error}}_{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}.$$

- We interpret the interval in the same way:

> "We are $100(1-\alpha)$ percent confident that the population proportion p is in this interval."

Note: This interval should be used only when the sample size n is "large." A common **rule of thumb** is to require

$$\begin{aligned}
n\widehat{p} &\geq 5 \\
n(1-\widehat{p}) &\geq 5.
\end{aligned}$$

Under these conditions, the CLT should adequately describe the sampling distribution of $\widehat{p}$, thereby making the confidence interval formula above approximately valid.

Example 7.3. One source of water pollution is gasoline leakage from underground storage tanks. In Pennsylvania, a random sample of $n = 74$ gasoline stations is selected from the state and the tanks are inspected; 10 stations are found to have at least one leaking tank. Calculate a 95 percent confidence interval for p, the population proportion of gasoline stations with at least one leaking tank.

SOLUTION. In this situation, we interpret

- gasoline station = individual "trial"
- at least one leaking tank = "success"
- p = population proportion of stations with at least one leaking tank.

For 95 percent confidence, we need $z_{0.05/2} = z_{0.025} \approx 1.96$.

```
> qnorm(0.975,0,1) ## upper 0.025 quantile
[1] 1.959964
```

The sample proportion of stations with at least one leaking tank is

$$\widehat{p} = \frac{10}{74} \approx 0.135.$$

Therefore, an approximate 95 percent confidence interval for p is

$$0.135 \pm 1.96\sqrt{\frac{0.135(1-0.135)}{74}} \implies (0.057, 0.213).$$

Interpretation: We are 95 percent confident that the population proportion of stations in Pennsylvania with at least one leaking tank is between 0.057 and 0.213.

CLT approximation check: We have

$$\begin{aligned} n\widehat{p} &= 74\left(\frac{10}{74}\right) = 10 \\ n(1-\widehat{p}) &= 74\left(1-\frac{10}{74}\right) = 64. \end{aligned}$$

Both of these are larger than $5 \implies$ we can feel comfortable in using this confidence interval formula.

7.4 Sample size determination

Motivation: In the planning stages of an experiment or investigation, we need to first determine **how many individuals** are needed to write a confidence interval with a given level of precision. For example, we might want to construct a

- 95 percent confidence interval to estimate the population mean time needed for patients to recover from infection. How many patients should we recruit?
- 99 percent confidence interval to estimate the population proportion of defective parts. How many parts should be sampled?

Of course, collecting data almost always costs money. Therefore, one must be cognizant not only of the statistical issues associated with **sample size determination**, but also of the practical issues like cost, time spent in data collection, personnel training, etc.

7.4.1 Inference for a population mean

Setting: Suppose that $Y_1, Y_2, ..., Y_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population. In Section 7.1, we derived a $100(1-\alpha)$ percent confidence interval for μ to be

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

Suppose, for the moment, that the population variance σ^2 was **known**. In real life, this is rarely the case (i.e., "rarely" = "never"). However, if σ^2 was known, then a $100(1-\alpha)$ percent confidence interval for μ could be calculated as

$$\overline{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Denote the **margin of error** by

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

This is the quantity that is added/subtracted to the point estimate $\overline{Y}$ to form the confidence interval for the population mean μ.

Formula: In the setting described above, it is possible to determine the sample size n necessary once we specify these three pieces of information:

- the value of σ^2 (e.g., an educated guess at its value; e.g., from historical data, etc.)
- the confidence level, $100(1-\alpha)$
- the margin of error, B.

This is true because

$$B = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \iff n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2.$$

This is the necessary sample size to guarantee a prescribed level of confidence $100(1-\alpha)$ and margin of error B.

Example 7.4. In a biomedical experiment, we would like to estimate the population mean remaining life μ of healthy rats that are given a certain dose of a toxic substance. Suppose that we would like to write a 95 percent confidence interval for μ with a margin of error equal to $B = 2$ days. From past studies, remaining rat lifetimes have been approximated by a normal distribution with standard deviation $\sigma = 8$ days. How many rats should we use for the experiment?

SOLUTION. With $z_{0.05/2} = z_{0.025} \approx 1.96$, $B = 2$, and $\sigma = 8$, the desired sample size to estimate μ is

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.96 \times 8}{2}\right)^2 \approx 61.46.$$

We would sample $n = 62$ rats to achieve these goals.

Extension: Suppose it is determined that this is too many rats. We could weaken our requirements to, say, $B = 3$ (a less informative interval) and 90 percent confidence (less confidence). The desired sample size is now

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.645 \times 8}{3}\right)^2 \approx 19.24.$$

We would need to sample only $n = 20$ rats to meet these weaker requirements.

7.4.2 Inference for a population proportion

Setting: Suppose we would like to write a $100(1-\alpha)$ percent confidence interval for p, a population proportion. We know that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for p. What sample size n should we use?

Note: To determine the necessary sample size n, we first need to specify two pieces of information:

- the confidence level $100(1-\alpha)$
- the margin of error:

$$B = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

A small problem arises. Note that the margin of error B depends on $\widehat{p}$. Unfortunately, $\widehat{p}$ can only be calculated once we know the sample size n. We overcome this problem by replacing $\widehat{p}$ with p_0, an **a priori guess** at its value. The last expression becomes

$$B = z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}.$$

Solving this equation for n, we get

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p_0(1-p_0).$$

This is the desired sample size n to write a $100(1-\alpha)$ percent confidence interval for the population proportion p with a prescribed margin of error (roughly) equal to B. I say "roughly," because there may be additional uncertainty arising from our use of p_0 (our best guess).

Remark: If there is no sensible guess for p available, use $p_0 = 0.5$. In this situation, the resulting value for n will be as large as possible. Put another way, using $p_0 = 0.5$ gives the

most **conservative** solution (i.e., the largest sample size, n). This is true because

$$n = n(p_0) = \left(\frac{z_{\alpha/2}}{B}\right)^2 p_0(1-p_0),$$

when viewed as a function of p_0, is maximized when $p_0 = 0.5$. However, the resulting sample size could be very large (perhaps much larger than is practical to use).

Example 7.5. You have been asked to estimate the proportion of raw material (in a certain manufacturing process) that is being "scrapped;" e.g., the material is so defective that it can not be reworked. If this proportion is larger than 10 percent, this will be deemed (by management) to be an unacceptable continued operating cost and a substantial process overhaul will be performed. Past experience suggests that the scrap rate is about 5 percent, but recent information suggests that this rate may be increasing.

You would like to write a 95 percent confidence interval for p, the population proportion of raw material that is to be scrapped, with a margin of error equal to $B = 0.02$. How many pieces of material should you ask to be sampled?

SOLUTION. For 95 percent confidence, we need $z_{0.05/2} = z_{0.025} \approx 1.96$. In providing an initial guess, we have options; we could use

$$\begin{aligned} p_0 &= 0.05 \text{ (historical scrap rate)} \\ p_0 &= 0.10 \text{ ("critical mass" value)} \\ p_0 &= 0.50 \text{ (most conservative choice).} \end{aligned}$$

For these choices, we have

$$\begin{aligned} n &= \left(\frac{1.96}{0.02}\right)^2 0.05(1-0.05) \approx 457 \\ n &= \left(\frac{1.96}{0.02}\right)^2 0.10(1-0.10) \approx 865 \\ n &= \left(\frac{1.96}{0.02}\right)^2 0.50(1-0.50) \approx 2401. \end{aligned}$$

As we can see, the "guessed" value of p_0 has a substantial impact on the final sample size calculation. Furthermore, it may not be practical to sample 2,401 parts, as would be required by the most conservative approach.

8 Two-Sample Inference

Preview: In this chapter, we discuss two-sample inference procedures for the following population parameters:

- The difference of two population **means** $\mu_1 - \mu_2$ (Section 8.1)
- The ratio of two population **variances** σ_2^2/σ_1^2 (Section 8.2)
- The difference of two population **proportions** $p_1 - p_2$ (Section 8.3).

Remember that these are population-level quantities (now involving two different populations), so they are unknown. Our goal is to use sample information (now with two samples) to estimate these quantities.

Usefulness: In practice, it is very common to compare the same characteristic (e.g., mean, variance, proportion, etc.) on two different populations. For example, we may wish to compare

- the population **mean** starting salaries of male and female engineers (compare μ_1 and μ_2). Is there evidence that males have a larger mean starting salary?
- the population **variance** of sound levels from two indoor swimming pool designs (compare σ_1^2 and σ_2^2). Are the sound-level acoustics of a new design more variable than the standard design?
- the population **proportion** of defectives produced from two different suppliers (compare p_1 and p_2). Are there differences between the two suppliers?

Note: Our methods in the last chapter are applicable only for a **single population** (i.e., a population mean μ, a population variance σ^2, and a population proportion p). We therefore extend these methods to two populations. We start with comparing two population means.

8.1 Confidence interval for the difference of two population means

$\mu_1 - \mu_2$

Setting: Suppose that we have two **independent** random samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2). \end{aligned}$$

Point estimators: Define the statistics

$$\begin{aligned} \overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} &= \text{sample mean for sample 1} \\ \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} &= \text{sample mean for sample 2} \\ S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 &= \text{sample variance for sample 1} \\ S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2 &= \text{sample variance for sample 2.} \end{aligned}$$

Goal: Our goal is to construct a $100(1-\alpha)$ percent confidence interval for the **difference** of two population means $\mu_1 - \mu_2$.

Important: How we construct this interval depends on our assumptions on the population variances σ_1^2 and σ_2^2. In particular, we consider two cases:

- $\sigma_1^2 = \sigma_2^2$; that is, the two population variances are **equal**
- $\sigma_1^2 \neq \sigma_2^2$; that is, the two population variances are **not equal**.

8.1.1 Independent samples: Equal population variances

Result: Under the assumptions above and when $\sigma_1^2 = \sigma_2^2$, the quantity

$$t = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2),$$

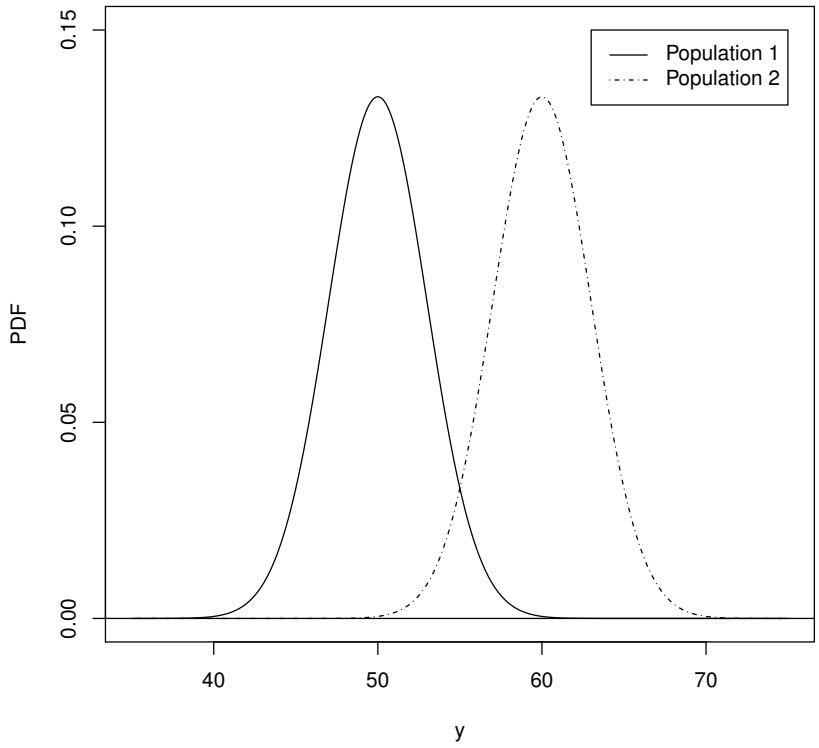

Figure 8.1: Two normal distributions with $\sigma_1^2 = \sigma_2^2$.

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

- For this sampling distribution to hold exactly, we need
  - the two random samples to be independent
  - the two population distributions to be normal (Gaussian)
  - the two population distributions to have the same variance; i.e., $\sigma_1^2 = \sigma_2^2$.
- The statistic S_p^2 is called the **pooled sample variance estimator** of the common population variance, say, σ^2. It is a weighted average of the two sample variances S_1^2 and S_2^2 (where the weights are functions of the sample sizes n_1 and n_2).

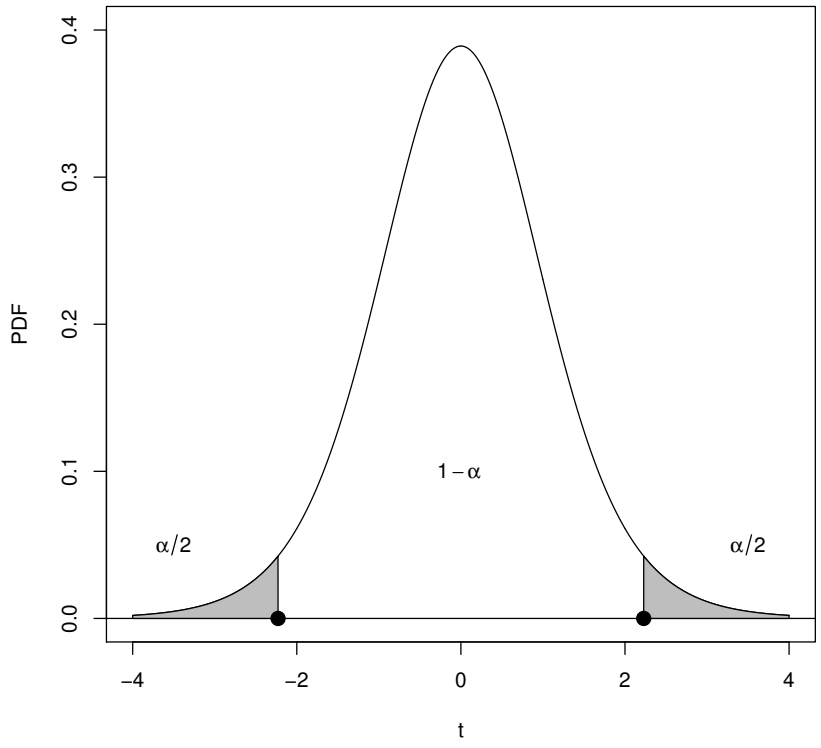

Figure 8.2: A t pdf with $n_1 + n_2 - 2$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $t_{n_1+n_2-2,\alpha/2}$ (upper) and $-t_{n_1+n_2-2,\alpha/2}$ (lower), respectively.

- The sampling distribution $t \sim t(n_1 + n_2 - 2)$ suggests that confidence interval quantiles will come from this t distribution; note that this distribution depends on the **sample sizes** from both samples.

- In particular, because $t \sim t(n_1 + n_2 - 2)$, the upper quantile $t_{n_1+n_2-2,\alpha/2}$ satisfies

$$P\left(-t_{n_1+n_2-2,\alpha/2} < \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{n_1+n_2-2,\alpha/2}\right) = 1 - \alpha.$$

  This probability equation is seen by examining Figure 8.2.

- After performing algebraic manipulations (similar to those in the last chapter), we obtain

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

  This is a $100(1-\alpha)$ **percent confidence interval** for the difference of two population means $\mu_1 - \mu_2$.

- We see that the interval again has the same form:

$$\underbrace{\text{point estimate}}_{\overline{Y}_{1+}-\overline{Y}_{2+}} \pm \underbrace{\text{quantile}}_{t_{n_1+n_2-2,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

  We interpret the interval in the same way.

  > "We are $100(1-\alpha)$ percent confident that the population mean difference $\mu_1 - \mu_2$ is in this interval."

- **Important:** In two-sample situations, it is usually of interest to compare the population means μ_1 and μ_2:
  - If the confidence interval for $\mu_1 - \mu_2$ includes 0, this does not suggest that the population means μ_1 and μ_2 are different.
  - If the confidence interval for $\mu_1 - \mu_2$ does not include 0, this suggests the population means are different.

Example 8.1. In the vicinity of a nuclear power plant, environmental engineers from the EPA would like to determine if there is a difference between the population mean weight in fish (of the same species) from two locations. Independent samples are taken from each location and the following weights (in ounces) are observed:

Location 1:	21.9	18.5	12.3	16.7	21.0	15.1	18.2	23.0	36.8	26.6
Location 2:	21.0	19.6	14.4	16.9	23.4	14.6	10.4	16.5		

Construct a 90 percent confidence interval for the population mean weight difference $\mu_1 - \mu_2$, where the mean weight μ_1 (μ_2) corresponds to location 1 (2).

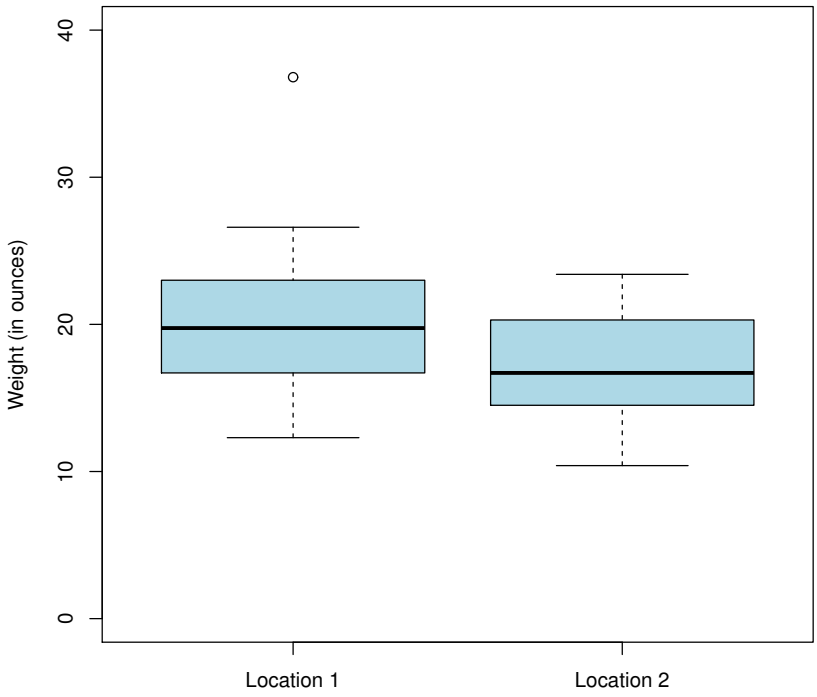

Figure 8.3: Boxplots of fish weight data by location in Example 8.1.

ANALYSIS. In order to visually assess the equal population variance assumption $\sigma_1^2 = \sigma_2^2$, we use **boxplots** to display the data in each sample; see Figure 8.3.

- The equal variance assumption looks reasonable; the spread in each distribution looks roughly the same (save the outlier in Location 1).
- In Section 8.2, we will look at formal statistical inference procedures to compare two population variances. For now, we will rely on this rough assessment (based on sample information only; no inference).

Calculation: We can use R to calculate the confidence interval directly:

```
> t.test(loc.1,loc.2,conf.level=0.90,var.equal=TRUE)$conf.int
[1] -0.9404376  8.7604376
```

A 90 percent confidence interval for the population mean difference $\mu_1 - \mu_2$ is

$$(-0.940, 8.760) \text{ oz.}$$

Interpretation: We are 90 percent confident that the population mean difference $\mu_1 - \mu_2$ is between -0.940 and 8.760 oz. Note that this interval includes "0." Therefore, we do not have sufficient evidence that the population mean fish weights μ_1 and μ_2 are different.

Robustness: Some comments are in order about the robustness properties of the two-sample confidence interval

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

for the difference of two population means $\mu_1 - \mu_2$.

- We should only use this confidence interval if there is **strong evidence** that the population variances σ_1^2 and σ_2^2 are similar.
- In other words, this confidence interval is **not robust** to a violation of the equal population variance assumption.
- Like the one-sample t confidence interval for a single population mean μ, this two-sample t confidence interval is robust to mild departures from normality.
- This means that we can feel comfortable using the interval even if the underlying population distributions are not perfectly normal (Guassian).

Fish weight data: Normal qq plots for the two samples of fish weight data are given in Figure 8.4.

- With such small sample sizes ($n_1 = 10$ and $n_2 = 8$), it is hard to make any conclusive assessments about the normal population assumption.
- This uncertainty manifests itself in the qq plots; note how the "bands of uncertainty" are very wide.
- The very heavy fish (36.8 oz) in the first sample might be regarded as an **outlier**, because it is the only observation that falls outside the bands.

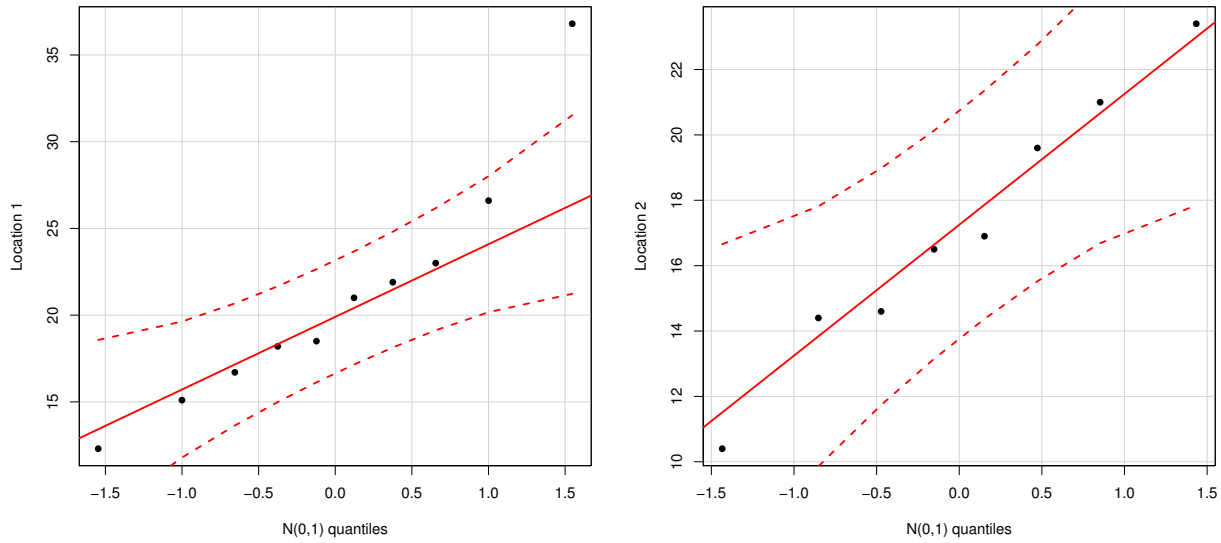

Figure 8.4: Quantile-quantile plots for the fish weight data in Example 8.1.

8.1.2 Independent samples: Unequal population variances

Remark: When $\sigma_1^2 \neq \sigma_2^2$, we can not use

$$t = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

as a basis for inference because the pooled sample variance S_p^2 does not estimate anything meaningful. Constructing a confidence interval for $\mu_1 - \mu_2$ becomes more difficult theoretically. However, we can write an approximate confidence interval.

Formula: An approximate $100(1-\alpha)$ percent confidence interval for the difference of two population means $\mu_1 - \mu_2$ is given by

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where the degrees of freedom ν is calculated as

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}.$$

- This interval is always approximately valid, as long as
  - the two samples are independent
  - the two population distributions are approximately normal (Gaussian).
- This interval remains valid even when $\sigma_1^2 = \sigma_2^2$, but its theoretical properties are not as good as those of the equal population variance interval.
- No one in their right mind would calculate this interval "by hand" (particularly nasty is the formula for ν). R will produce the interval on request.

Example 8.2. You are part of a recycling project that is examining how much paper is being discarded (not recycled) by employees at two large plants. These data are obtained on the amount of white paper thrown out per year by employees (data are in hundreds of pounds). Samples of employees at each plant were randomly selected.

Plant 1:	3.01	2.58	3.04	1.75	2.87	2.57	2.51	2.93	2.85	3.09
	1.43	3.36	3.18	2.74	2.25	1.95	3.68	2.29	1.86	2.63
	2.83	2.04	2.23	1.92	3.02					
Plant 2:	3.99	2.08	3.66	1.53	4.27	4.31	2.62	4.52	3.80	5.30
	3.41	0.82	3.03	1.95	6.45	1.86	1.87	3.98	2.74	4.81

Construct a 95 percent confidence interval for the population mean difference $\mu_1 - \mu_2$, where the mean amount discarded μ_1 (μ_2) corresponds to Plant 1 (2).

Analysis. We use boxplots to display the data; see Figure 8.5. This figure suggests the equal population variance assumption is doubtful. The spread in the two boxplots is markedly difficult (again, this is a rough determination based on the sample information).

Calculation: We can use R to calculate the confidence interval directly:

```
> t.test(plant.1,plant.2,conf.level=0.95,var.equal=FALSE)$conf.int
[1] -1.46179176 -0.06940824
```

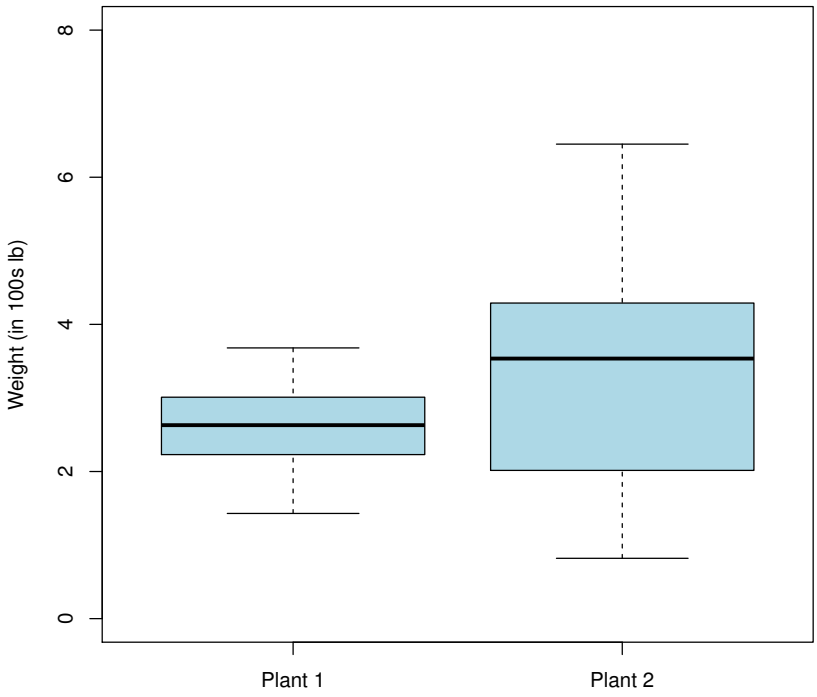

Figure 8.5: Boxplots of discarded white paper amounts (in 100s lb) in Example 8.2.

A 95 percent confidence interval for the population mean difference $\mu_1 - \mu_2$ is

$$(-1.461, -0.069) \text{ 100s lbs.}$$

Interpretation: We are 95 percent confident that the population mean difference $\mu_1 - \mu_2$ is between -146.1 and -6.9 lbs. This interval does not include "0" and contains only negative values. Therefore, we have evidence that the population mean amount of discarded paper is smaller for Plant 1 than it is for Plant 2.

Normality: Normal qq plots for the two samples of white paper data are given in Figure 8.6. There is no cause to question the normality assumption.

Remark: We have presented two confidence intervals for the population mean difference $\mu_1 - \mu_2$. One assumes $\sigma_1^2 = \sigma_2^2$ (equal population variances) and one that assumes $\sigma_1^2 \neq \sigma_2^2$ (unequal population variances). **Advice:** If you are unsure about which interval to use, go

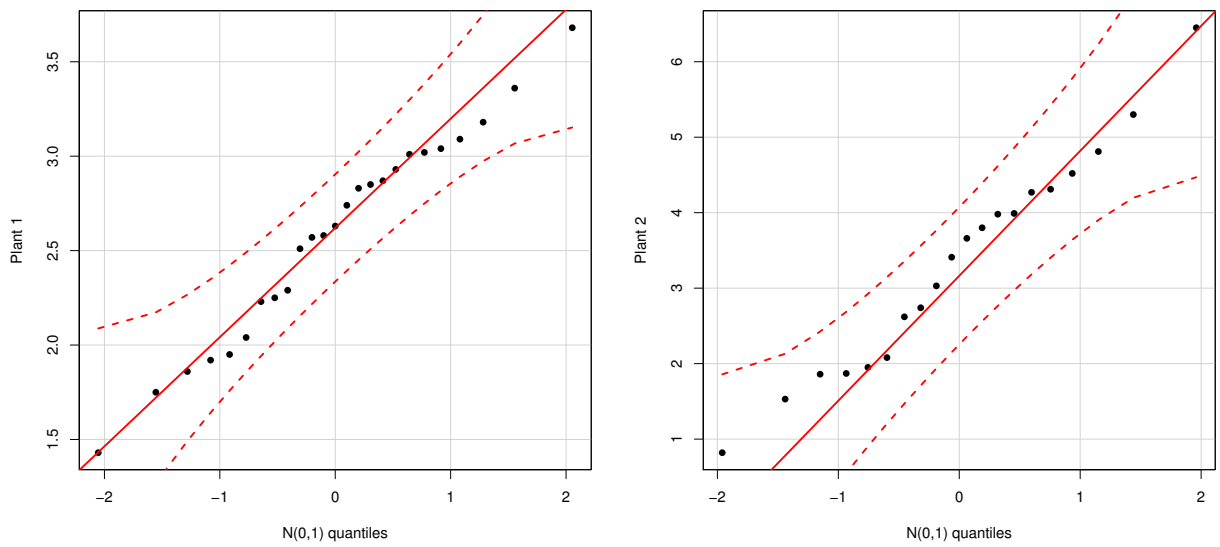

Figure 8.6: Quantile-quantile plots for the white paper data in Example 8.2.

with the unequal variance interval. The penalty for using it when $\sigma_1^2 = \sigma_2^2$ is much smaller than the penalty for using the equal variance interval when $\sigma_1^2 \neq \sigma_2^2$. I usually calculate both intervals (easy to do quickly with R) and then determine whether the intervals lead to drastically different conclusions.

8.1.3 Dependent samples: Matched pairs

Example 8.3. Ergonomics experts hired by a large company designed a study to determine whether more varied work conditions would have any impact on arm movement. The data on the next page were obtained on a random sample of $n = 26$ employees. Each observation is the amount of time, expressed as a percentage of the total time observed, during which arm elevation **was below 30 degrees**. This percentage is a surrogate for the percentage of time spent on repetitive tasks. The two measurements from each employee were obtained 18 months apart. During this 18-month period, work conditions were "changed" by the ergonomics team, and subjects were allowed to engage in a wider variety of work tasks. "Before" and "after" measurements are obtained on each of the 26 employees.

Individual	Before	After	Individual	Before	After
1	81.3	78.9	14	74.9	58.3
2	87.2	91.4	15	75.8	62.5
3	86.1	78.3	16	72.6	70.2
4	82.2	78.3	17	80.8	58.7
5	90.8	84.4	18	66.5	66.6
6	86.9	67.4	19	72.2	60.7
7	96.5	92.8	20	56.5	65.0
8	73.0	69.9	21	82.4	73.7
9	84.2	63.8	22	88.8	80.4
10	74.5	69.7	23	80.0	78.8
11	72.0	68.4	24	91.1	81.8
12	73.8	71.8	25	97.5	91.6
13	74.2	58.3	26	70.0	74.2

Table 8.1: Ergonomics data. Percentage of time arm elevation was less than 30 degrees.

Question: Does the population mean time (during which elevation is below 30 degrees) decrease after the ergonomics team changes the working conditions?

Terminology: A **matched-pairs design** is an experimental design where one obtains a pair of measurements on each individual (e.g., employee, material, machine, etc.):

- one measurement corresponds to "Treatment 1"
- the other measurement corresponds to "Treatment 2"
- Clearly, the two samples are no longer independent. Each individual contributes a response to both samples.
- If possible, it is important to **randomize** the order in which treatments are assigned. This may eliminate "common patterns" that may be seen when always following, say, Treatment 1 with Treatment 2. In practice, the experimenter could flip a fair coin to determine which treatment is applied first.

This type of design removes variation **among** the individuals. This allows you to compare the two treatments (e.g., before/after working environment) under more **homogeneous** conditions where only variation within individuals is present (that is, the variation arising from the difference in the two treatments).

Table 8.2: Ergonomics example. Sources of variation in the two independent sample and matched pairs designs.

Design	Sources of Variation
Two Independent Samples	among employees, within employees
Matched Pairs	within employees

Advantage: When you remove extra variability, this enables you to compare the two experimental conditions (treatments) more precisely. This gives you a better chance of identifying a difference between the treatment means if one really exists.

- In a design with two independent samples, the extra variation among individuals may prevent us from being able to identify this difference!

Implementation: Data from matched pairs experiments are analyzed by examining the difference in responses of the two treatments. Specifically, compute

$$D_j = Y_{1j} - Y_{2j},$$

for each individual $j = 1, 2, ..., n$. After doing this, we have essentially created a "one sample problem," where our data are now

$$D_1, D_2, ..., D_n,$$

the so-called **data differences**. The one sample $100(1-\alpha)$ percent confidence interval

$$\overline{D} \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}},$$

Individual	Before	After	Difference	Individual	Before	After	Difference
1	81.3	78.9	2.4	14	74.9	58.3	16.6
2	87.2	91.4	−4.2	15	75.8	62.5	13.3
3	86.1	78.3	7.8	16	72.6	70.2	2.4
4	82.2	78.3	3.9	17	80.8	58.7	22.1
5	90.8	84.4	6.4	18	66.5	66.6	−0.1
6	86.9	67.4	19.5	19	72.2	60.7	11.5
7	96.5	92.8	3.7	20	56.5	65.0	−8.5
8	73.0	69.9	3.1	21	82.4	73.7	8.7
9	84.2	63.8	20.4	22	88.8	80.4	8.4
10	74.5	69.7	4.8	23	80.0	78.8	1.2
11	72.0	68.4	3.6	24	91.1	81.8	9.3
12	73.8	71.8	2.0	25	97.5	91.6	5.9
13	74.2	58.3	15.9	26	70.0	74.2	−4.2

Table 8.3: Ergonomics data. Percentage of time arm elevation was less than 30 degrees. The data differences $D_j = Y_{1j} - Y_{2j}$ have been added.

where $\overline{D}$ and S_D are the sample mean and sample standard deviation of the differences, respectively, is an interval estimate for

$$\begin{aligned} \mu_D &= \mu_1 - \mu_2 \\ &= \text{population mean difference between the 2 treatments.} \end{aligned}$$

The parameter $\mu_D = \mu_1 - \mu_2$ describes the **population mean difference** for the two treatment groups. If the two population means are then same, then $\mu_D = 0$. Therefore,

- If the confidence interval for μ_D includes 0, this does not suggest that the two population means are different.
- If the confidence interval for μ_D does not include 0, this suggests that the two population means are different.

Analysis: With the ergonomics data, we use R to construct a 95 percent confidence interval for $\mu_D = \mu_1 - \mu_2$:

```
> t.test(diff,conf.level=0.95)$conf.int
[1] 3.636034 9.894735
```

Note that this is a one-sample confidence interval calculated using the data differences.

Interpretation: We are 95 percent confident that the population mean difference $\mu_D = \mu_1 - \mu_2$ is between 3.6 and 9.9 percent. This interval does not include "0" and contains only **positive** values.

- Therefore, we have evidence that the population mean percentage of time that arm elevation is below 30 degrees is larger in the "before" condition than in the "after" condition.
- In other words, there is evidence that the "change" in work conditions implemented by the ergonomics team (in the 18-month interim) did reduce this population mean time.

Assumptions: In matched pairs experiments, the relevant assumptions are

1. The individuals sampled form a random sample.
2. The **data differences** $D_1, D_2, ..., D_n$ are normally distributed.

Ergonomics data: A normal qq plot for the data differences is given in Figure 8.7. A very picky analyst might pick out the mild departure in the upper tail.

- However, remember that one-sample t confidence intervals (for means) are generally robust to these mild departures. Therefore, this slight departure (which I don't think is convincingly real) likely does not affect our conclusion.

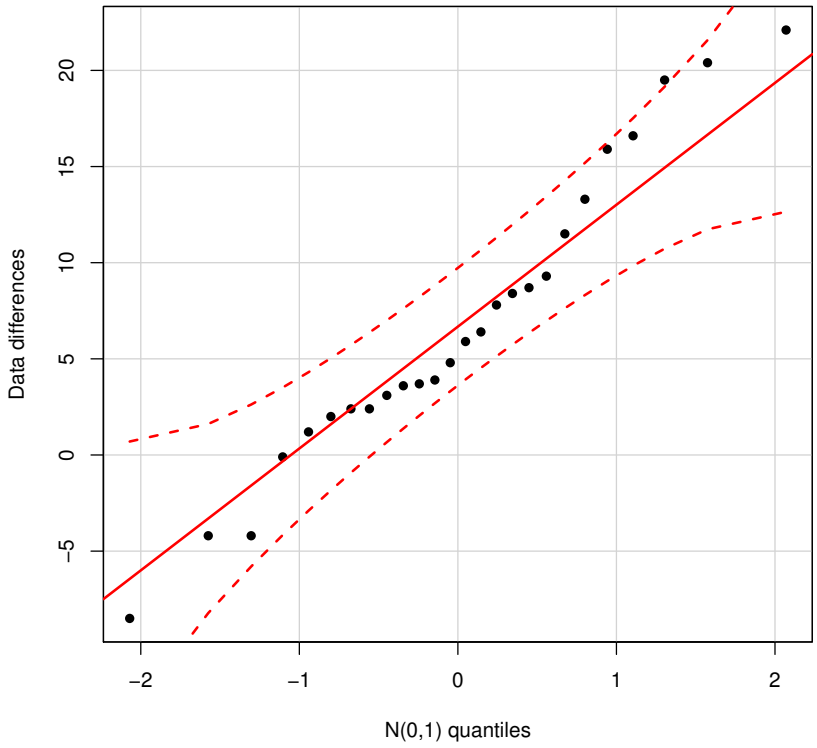

Figure 8.7: Normal qq plot for the ergonomics data in Example 8.3. The observed **data differences** are plotted versus the theoretical quantiles from a normal distribution. The line added passes through the first and third theoretical quartiles.

8.2 Confidence interval for the ratio of two population variances σ_2^2/σ_1^2

Importance: Recall that when we wrote a confidence interval for $\mu_1 - \mu_2$, the difference of the population means (with independent samples), we proposed two intervals:

- one interval that assumed $\sigma_1^2 = \sigma_2^2$
- one interval that assumed $\sigma_1^2 \neq \sigma_2^2$.

We now propose a confidence interval procedure that can be used to determine which assumption is more appropriate.

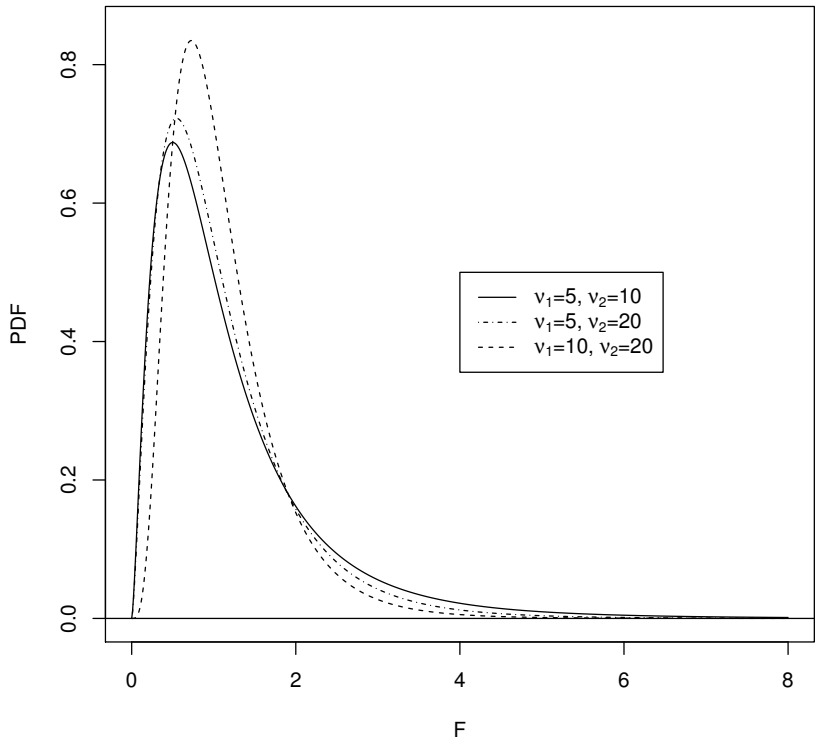

Figure 8.8: F pdfs with different degrees of freedom.

Setting: Suppose that we have two **independent** random samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2). \end{aligned}$$

Goal: Our goal is to construct a $100(1-\alpha)$ percent confidence interval for the **ratio** of population variances σ_2^2/σ_1^2.

Result: Under the setting described above,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1),$$

an F **distribution** with (numerator) $n_1 - 1$ and (denominator) $n_2 - 1$ degrees of freedom.

Facts: The F pdf has the following characteristics:

- continuous, skewed right, and always positive; see Figure 8.8.
- indexed by two **degree of freedom** parameters ν_1 and ν_2; these are usually integers and are related to sample sizes
- the **mean** of an F distribution is close to 1 (regardless of the values of ν_1 and ν_2)
- The F pdf formula is complicated and is unnecessary for our purposes. R will compute F probabilities and quantiles from the F distribution.

F **R CODE:** Suppose that $Q \sim F(\nu_1, \nu_2)$.

$F_Q(q) = P(Q \leq q)$	ϕ_p
`pf(q,`ν_1, ν_2`)`	`qf(p,`ν_1, ν_2`)`

Notation: We introduce new notation that identifies quantiles from an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Define

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= \textbf{upper } \alpha/2 \text{ quantile from } F(n_1-1, n_2-1) \text{ pdf} \\ F_{n_1-1,n_2-1,\alpha/2} &= \textbf{lower } \alpha/2 \text{ quantile from } F(n_1-1, n_2-1) \text{ pdf} \end{aligned}$$

Illustration: If $n_1 = 11$, $n_2 = 11$, and $\alpha = 0.05$ then

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{10,10,0.975} \approx 3.72 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{10,10,0.025} \approx 0.27 \end{aligned}$$

```
> qf(0.975,10,10)
[1] 3.716792
> qf(0.025,10,10)
[1] 0.2690492
```

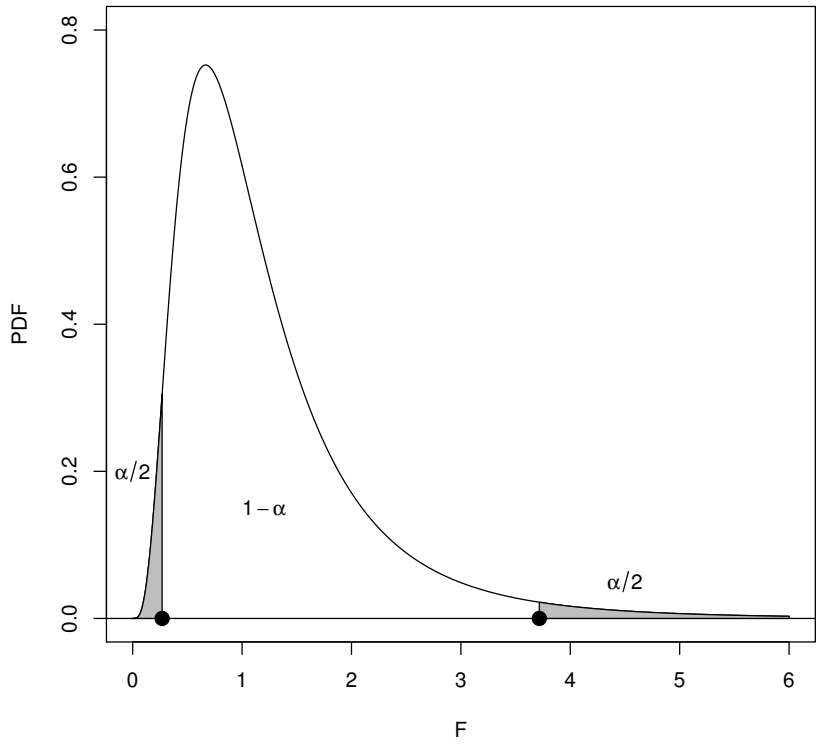

Figure 8.9: An F pdf with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The upper $\alpha/2$ and lower $\alpha/2$ areas are shaded. The associated quantiles, represented in the figure by dark circles, are denoted by $F_{n_1-1,n_2-1,1-\alpha/2}$ (upper) and $F_{n_1-1,n_2-1,\alpha/2}$ (lower), respectively.

Derivation: In general, for any value of α, $0 < \alpha < 1$, we can write

$$\begin{aligned} 1-\alpha &= P\left(F_{n_1-1,n_2-1,\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1,1-\alpha/2}\right) \\ &= P\left(\frac{S_2^2}{S_1^2}\, F_{n_1-1,n_2-1,\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}\, F_{n_1-1,n_2-1,1-\alpha/2}\right). \end{aligned}$$

This argument shows that

$$\left(\frac{S_2^2}{S_1^2}\, F_{n_1-1,n_2-1,\alpha/2},\ \frac{S_2^2}{S_1^2}\, F_{n_1-1,n_2-1,1-\alpha/2}\right)$$

is a $100(1-\alpha)$ **percent confidence interval** for the ratio of the population variances σ_2^2/σ_1^2.

Remarks: We interpret the interval in the same way:

> "We are $100(1-\alpha)$ percent confident that the ratio of the population variances σ_2^2/σ_1^2 is in this interval."

- If the confidence interval for σ_2^2/σ_1^2 includes 1, this does not suggest that the population variances σ_1^2 and σ_2^2 are different.
- If the confidence interval for σ_2^2/σ_1^2 does not include 1, this suggests that the population variances σ_1^2 and σ_2^2 are different.
- Therefore, this interval can be helpful in selecting an appropriate confidence interval for the difference of the population means $\mu_1 - \mu_2$; i.e., between the one that assumes equal population variances and the one that does not.
- Of course, even if inference for population means is not the objective, this interval is still useful in its own right−it allows you to compare the variances of two populations. This is an important problem if one is concerned about variation.

Example 8.4. Two automated filling processes are used in the production of automobile paint. The target weight of each process is 128.0 fluid oz (1 gallon). There is little concern about the process population mean fill amounts (no complaints about under/overfilling on average). However, there is concern that the population variation levels between the two processes are different. To test this claim, industrial engineers took independent random samples of $n_1 = 24$ and $n_2 = 24$ gallons of paint and observed the fill amounts.

	127.75	127.87	127.86	127.92	128.03	127.94	127.91	128.10
Process 1:	128.01	128.11	127.79	127.93	127.89	127.96	127.80	127.94
	128.02	127.82	128.11	127.92	127.74	127.78	127.85	127.96
	127.90	127.90	127.74	127.93	127.62	127.76	127.63	127.93
Process 2:	127.86	127.73	127.82	127.84	128.06	127.88	127.85	127.60
	128.02	128.05	127.95	127.89	127.82	127.92	127.71	127.78

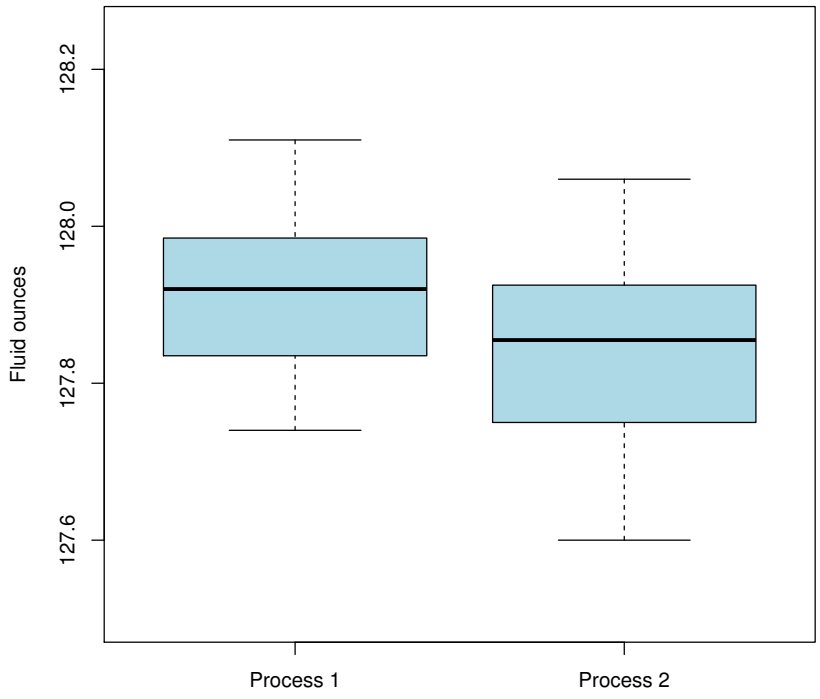

Figure 8.10: Boxplots of paint fill volume data in Example 8.4.

Note: I could not find an internal function in R to calculate the confidence interval for the ratio of two population variances σ_2^2/σ_1^2, so I wrote one:

```
ratio.var.interval = function(data.1,data.2,conf.level=0.95){
    df.1 = length(data.1)-1
    df.2 = length(data.2)-1
    F.lower = qf((1-conf.level)/2,df.1,df.2)
    F.upper = qf((1+conf.level)/2,df.1,df.2)
    s2.1 = var(data.1)
    s2.2 = var(data.2)
    c((s2.2/s2.1)*F.lower,(s2.2/s2.1)*F.upper)
    }
```

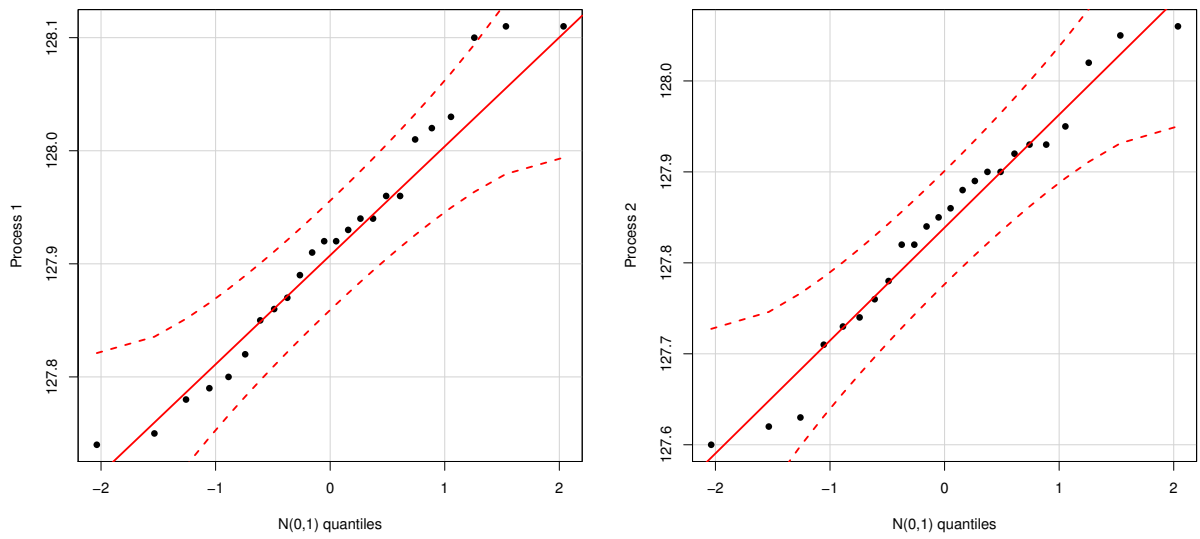

Figure 8.11: Quantile-quantile plots for the paint fill volume data in Example 8.4.

```
> ratio.var.interval(process.1,process.2)
[1] 0.5885236 3.1448830
```

Interpretation: We are 95 percent confident that the ratio of the population variances σ_2^2/σ_1^2 is between 0.589 and 3.145. Because this interval includes "1," we do not have evidence that the population variances σ_1^2 and σ_2^2 are different for the two processes.

Warning: Like the χ^2 interval for single population variance σ^2, the two-sample F interval for the ratio of two population variances σ_2^2/σ_1^2 **is not robust** to normality departures. This is true because the sampling distribution

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

depends critically on the normal distribution assumption for both populations. If either underlying population distribution is non-normal (non-Guassian), then the confidence interval formula for σ_2^2/σ_1^2 is not to be used.

Paint data: Normal qq plots for the two samples of paint fill volume data are given in Figure 8.11. There is no major cause for concern here.

8.3 Confidence interval for the difference of two population proportions $p_1 - p_2$

Interest: We now extend our confidence interval procedure for a single population proportion p to **two populations**. Define

$$\begin{aligned} p_1 &= \text{population proportion in Population 1} \\ p_2 &= \text{population proportion in Population 2.} \end{aligned}$$

For example, we might want to compare the proportion of

- defective circuit boards for two different suppliers
- satisfied customers before and after a product design change (e.g., Facebook, etc.)
- on-time payments for two classes of customers
- HIV positives for individuals in two demographic classes.

Point estimators: We assume that there are two independent random samples of individuals (one sample from each population to be compared). Define

$$\begin{aligned} Y_1 &= \text{number of ``successes'' in Sample 1} \sim b(n_1, p_1) \\ Y_2 &= \text{number of ``successes'' in Sample 2} \sim b(n_2, p_2). \end{aligned}$$

The point estimators for p_1 and p_2 are the **sample proportions**, defined by

$$\begin{aligned} \widehat{p}_1 &= \frac{Y_1}{n_1} \\ \widehat{p}_2 &= \frac{Y_2}{n_2}. \end{aligned}$$

Goal: We would like to construct a $100(1-\alpha)$ percent confidence interval for $p_1 - p_2$, the difference of two population proportions.

Result: We need the following sampling distribution result. When the sample sizes n_1 and n_2 are large,

$$Z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{AN}(0,1).$$

If this sampling distribution holds approximately, then

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

is an approximate $100(1-\alpha)$ **percent confidence interval** for $p_1 - p_2$.

- Note again the form of the interval:

$$\underbrace{\text{point estimate}}_{\widehat{p}_1-\widehat{p}_2} \pm \underbrace{\text{quantile}}_{z_{\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}.$$

  We interpret the interval in the same way:

  > "We are $100(1-\alpha)$ percent confident that the population proportion difference $p_1 - p_2$ is in this interval."

- The value $z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the $\mathcal{N}(0,1)$ distribution.

Note: For the Z sampling distribution to hold approximately (and therefore for the interval above to be useful), we need

- the two random samples to be independent
- the sample sizes n_1 and n_2 to be "large;" common rules of thumb are to require

$$\begin{aligned} n_i\widehat{p}_i &\geq 5 \\ n_i(1-\widehat{p}_i) &\geq 5, \end{aligned}$$

  for each sample $i = 1, 2$. Under these conditions, the Central Limit Theorem should adequately approximate the true sampling distribution of Z, thereby making the confidence interval formula above approximately valid.

Important: In two-sample situations, it is often of interest to see how the population proportions p_1 and p_2 compare.

- If the confidence interval for $p_1 - p_2$ includes 0, this does not suggest that the population proportions p_1 and p_2 are different.
- If the confidence interval for $p_1 - p_2$ does not include 0, this suggests that the population proportions p_1 and p_2 are different.

Example 8.5. A large public health study was conducted to estimate the prevalence and to identify risk factors of hepatitis B virus (HBV) infection among Irish prisoners. Two independent samples of female ($n_1 = 82$) and male ($n_2 = 555$) prisoners were obtained from five prisons in Ireland:

- 18 out of 82 female prisoners were HBV-positive
- 28 out of 555 male prisoners were HBV-positive.

Find a 95 percent confidence interval for $p_1 - p_2$, the difference in the population proportions for the two genders (Female = 1; Male = 2).

ANALYSIS. There is no internal function in R to calculate the confidence interval for the difference of two population proportions (at least not that I could find quickly), so I wrote one:

```
proportion.diff.interval = function(y.1,n.1,y.2,n.2,conf.level=0.95){
    z.upper = qnorm((1+conf.level)/2)
    var.1 = (y.1/n.1)*(1-y.1/n.1)/n.1
    var.2 = (y.2/n.2)*(1-y.2/n.2)/n.2
    se = sqrt(var.1+var.2)
    moe = z.upper*se
    c((y.1/n.1-y.2/n.2)-moe,(y.1/n.1-y.2/n.2)+moe)
    }


> proportion.diff.interval(18,82,28,555)
[1] 0.07764115 0.26048234
```

Interpretation: We are 95 percent confident the difference of the population proportions $p_1 - p_2$ is between 0.078 and 0.260. This interval does not contain "0" and contains only positive values. This suggests that the population proportion of female prisoners who are HBV positive is larger than the corresponding male population proportion. The sample size conditions on the previous page are satisfied.

9 One-Way Analysis of Variance

9.1 Introduction

Recall: In the last chapter, we discussed confidence intervals for the difference of two population means $\mu_1 - \mu_2$. Perhaps more importantly, we also saw that the **design** of the experiment or study completely determined how the analysis should proceed.

- When the two samples are independent, this is called a **(two) independent-sample design**.
- When the two samples are obtained on the same individuals (so that the samples are dependent), this is called a **matched pairs design**.
- Confidence interval procedures for $\mu_1 - \mu_2$ depend on the design of the study.

Terminology: More generally, the purpose of an **experiment** is to investigate differences between or among two or more treatments. In a statistical framework, we do this by comparing the population means of the responses to each treatment.

- In order to detect treatment mean differences, we must try to **control** the effects of error so that any variation we observe can be attributed to the effects of the treatments rather than to structural differences among the individuals.
- For example, in Example 8.2, there may be a **systematic source of variation** arising from the ages of employees in the recycling project (e.g., younger employees may be more inclined to recycle paper instead of discarding it).
- Our two-independent sample design (one sample from Plant 1 and one sample from Plant 2) did not consider this potential **confounding effect**. In other words, even if age of the employee is a significant source of variability, our independent sample analysis does not acknowledge it.

Terminology: Designs involving meaningful grouping of individuals, that is, **blocking**, can help reduce the effects of experimental error by identifying systematic components of variation among individuals.

- The matched pairs design for comparing two treatments is an example of such a design.
- In this situation, the "meaningful grouping of individuals" involves the individuals themselves. Responses to two different treatments on the same individual "blocks out" the variation that would arise had we observed one individual's response to the first treatment and a different individual's response to the second treatment.

Remark: Aside from matched pairs experiments, the analysis of data from experiments involving blocking will not be covered in this course. When there are more than two treatments (populations), we pursue the **one-way classification model**. This is basically an extension of the two independent sample design to two or more populations.

Situation: Consider an **experiment** to compare $t \geq 2$ treatments set up as follows:

- We obtain one random sample of individuals and then randomly assign individuals to treatments (i.e., different experimental conditions). Samples corresponding to the treatment groups are **independent**.
- In an **observational study** (where no treatment is physically applied to individuals), individuals are inherently different to begin with. We therefore simply take random samples from each treatment population (see Example 9.1).
- We do not attempt to group individuals according to some other factor (e.g., location, gender, weight, variety, etc.). This would be an example of blocking.

Main point: In a one-way classification, the only way individuals are "classified" is by the treatment group assignment. When individuals are thought to be "basically alike" (other than the possible effect due to treatment), experimental error consists only of the variation among the individuals themselves. There are no other **systematic** sources of variability.

Example 9.1. Mortar mixes are usually classified on the basis of compressive strength and their bonding properties and flexibility. In a building project, engineers wanted to compare specifically the population mean strengths of four types of mortars:

1. ordinary cement mortar (OCM)
2. polymer impregnated mortar (PIM)
3. resin mortar (RM)
4. polymer cement mortar (PCM).

Random samples of specimens of each mortar type were taken; each specimen was subjected to a compression test to measure strength (MPa). Here are the strength measurements taken on different mortar specimens (36 in all).

OCM:	51.45	42.96	41.11	48.06	38.27	38.88	42.74	49.62		
PIM:	64.97	64.21	57.39	52.79	64.87	53.27	51.24	55.87	61.76	67.15
RM:	48.95	62.41	52.11	60.45	58.07	52.16	61.71	61.06	57.63	56.80
PCM:	35.28	38.59	48.64	50.99	51.52	52.85	46.75	48.31		

Side by side boxplots of these data are given in Figure 9.1.

Note: First note that this is an example of an **observational study**. This is not what statisticians would call an experiment, because the "individuals" (here, the mortar specimens) are not treated or influenced by different experimental conditions. We are simply observing individuals from different groups (populations) to begin with.

Note: There is no form of blocking here either. For example, we do not attempt to further classify individual mortar specimens according to different manufacturers or subject individual mortar specimens to different environmental conditions (e.g., high/low temperature, etc.). If the study's purpose was investigate these potential sources of variability, then this would not be a one-way classification.

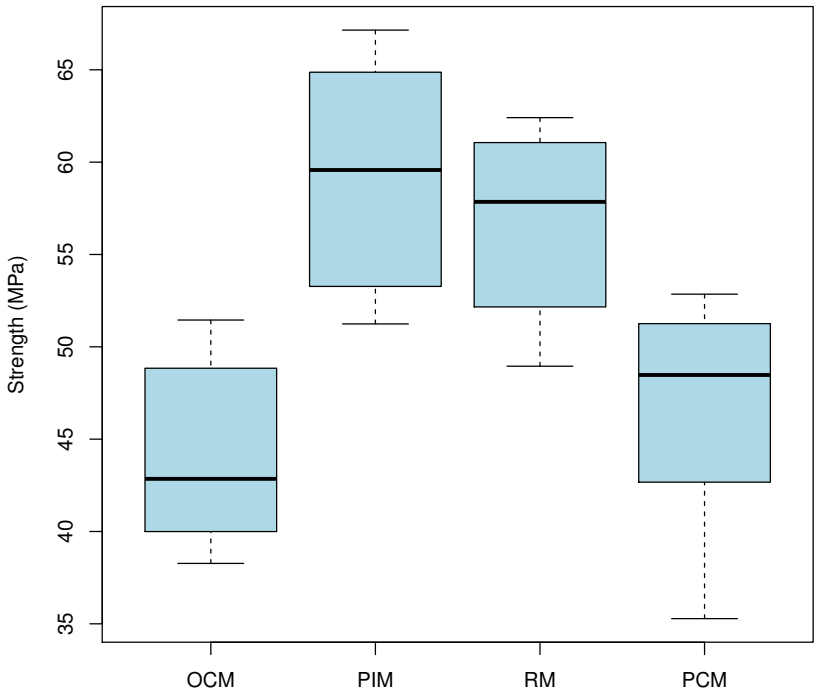

Figure 9.1: Boxplots of strength data (MPa) for four mortar types in Example 9.1.

Query: An initial question that engineers may have is the following:

> "Are the population mean mortar strengths equal among the four types of mortars? Or, are the population means different?"

This initial question can be framed statistically as the following **hypothesis test**:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

versus

$$H_1 : \text{the population means } \mu_i \text{ are not all equal.}$$

Goal: We now develop a **statistical inference** procedure that allows us to test this type of hypothesis in a one-way classification.

9.2 Overall F test

Notation: Let t denote the number of treatments (populations) to be compared. Define

$$Y_{ij} = \text{response on the } j\text{th individual in the } i\text{th treatment group}$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$.

- n_i is the number of **observations** for the ith treatment (population)
  - In Example 9.1, these are $n_1 = 8$, $n_2 = 10$, $n_3 = 10$, and $n_4 = 8$.
- When $n_1 = n_2 = \cdots = n_t = n$, we say the design is **balanced**; otherwise, the design is **unbalanced**.
- Let $N = n_1 + n_2 + \cdots + n_t$ denote the total number of individuals measured. If the design is balanced, then $N = nt$.
- Define the statistics

$$\begin{aligned} \overline{Y}_{i+} &= \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \\ S_i^2 &= \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 \\ \overline{Y}_{++} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}. \end{aligned}$$

  The statistics $\overline{Y}_{i+}$ and S_i^2 denote the **sample mean** and the **sample variance**, respectively, of the ith sample. The **overall sample mean** $\overline{Y}_{++}$ is the sample mean of all the data (aggregated across all t treatment groups).

Terminology: Our goal is to develop a procedure to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$

versus

$$H_1 : \text{the population means } \mu_i \text{ are not all equal.}$$

- The **null hypothesis** H_0 says that there is "no treatment difference," that is, all t population means are the same.
- The **alternative hypothesis** H_1 says that a difference among the t population means exists "somewhere." It does not specify how the means are different.
- When performing a hypothesis test, we basically decide which hypothesis is more supported by the data.

Setting: Suppose that we have t **independent** random samples:

$$\begin{aligned}
\text{Sample 1:} \quad & Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \mathcal{N}(\mu_1, \sigma^2) \\
\text{Sample 2:} \quad & Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \mathcal{N}(\mu_2, \sigma^2) \\
\vdots \qquad & \qquad \vdots \\
\text{Sample } t\text{:} \quad & Y_{t1}, Y_{t2}, ..., Y_{tn_t} \sim \mathcal{N}(\mu_t, \sigma^2).
\end{aligned}$$

Assumptions: Note the **statistical assumptions** we are making:

1. the t random samples are **independent**
2. the t population distributions are **normal** (Gaussian)
3. the t population distributions have the **same variance** σ^2.

Note also that these are the same assumptions we made for the two independent-sample design in Section 8.1 (i.e., the special case when $t = 2$).

Curiosity: If we are trying to learn about how the population means compare, why is the statistical inference procedure designed to do this called "the analysis of variance?"

Answer: We learn about the population means by estimating the common variance σ^2 in two different ways. These two estimators are formed by

- measuring variability of the observations **within** each sample
- measuring variability of the sample means **across** the samples

- **Important:** These two estimates tend to be similar when H_0 is true. The second estimate tends to be larger than the first estimate when H_1 is true.

Within Estimator: Calculate the **residual sum of squares**:

$$\begin{aligned} \text{SS}_{res} &= (n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots + (n_t-1)S_t^2 \\ &= \sum_{i=1}^{t} \underbrace{\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2}_{(n_i-1)S_i^2}. \end{aligned}$$

- The sample variance S_i^2 estimates the population parameter σ^2 (which assumed to be common across all t populations) from **within** the ith sample.

- The weighted average of these estimates

$$\begin{aligned} \text{MS}_{res} &= \left(\frac{n_1-1}{N-t}\right) S_1^2 + \left(\frac{n_2-1}{N-t}\right) S_2^2 + \cdots + \left(\frac{n_t-1}{N-t}\right) S_t^2 \\ &= \frac{\text{SS}_{res}}{N-t} \end{aligned}$$

  is called the **residual mean squares**. It is an unbiased estimator of σ^2 regardless of whether H_0 or H_1 is true.

- The **within estimator** MS_{res} is a generalization of the pooled sample variance estimator S_p^2 we discussed in Section 8.1 with $t = 2$ populations.

Across Estimator: We assume a common sample size $n_1 = n_2 = \cdots = n_t = n$ to simplify notation (i.e., a balanced design).

Recall: From Result 1 in Chapter 6 (pp 75), we know that if a sample arises from a normal population, the sample mean is also normally distributed. Therefore, the sample mean of the ith sample

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

Therefore, **when the null hypothesis** $H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$ **is true**, we have the following sampling distributions for each sample mean:

$$\begin{aligned}
\overline{Y}_{1+} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{Y}_{2+} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \\
&\vdots \\
\overline{Y}_{t+} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),
\end{aligned}$$

where μ is the common population mean under H_0. Now, think of

$$\overline{Y}_{1+}, \overline{Y}_{2+}, ..., \overline{Y}_{t+}$$

as a random sample from the $\mathcal{N}(\mu, \sigma^2/n)$ population distribution. The sample variance of this "random sample" is

$$\frac{1}{t-1}\sum_{i=1}^{t}(\overline{Y}_{i+} - \overline{Y}_{++})^2$$

and is an unbiased estimator of σ^2/n. Therefore,

$$\text{MS}_{trt} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n(\overline{Y}_{i+} - \overline{Y}_{++})^2}_{\text{SS}_{trt}}$$

is an unbiased estimator of σ^2. We call

$$\begin{aligned}
\text{SS}_{trt} &= \text{"treatment sums of squares"} \\
\text{MS}_{trt} &= \text{"treatment mean squares."}
\end{aligned}$$

The **across estimator** MS_{trt} is an unbiased estimator of σ^2 **when** H_0 **is true**.

Remark: Our derivation of the across estimator assumed a balanced design (this was done for simplicity). If we have different sample sizes n_i, we simply adjust MS_{trt} to

$$\text{MS}_{trt} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2}_{\text{SS}_{trt}}.$$

This is still an unbiased estimator for σ^2 **when** H_0 **is true**.

Summary: If you are terrified by the preceding derivation, that is fine. Just know the following:

1. **When H_0 is true** (i.e., the population means are equal), then

$$\begin{aligned} E(\text{MS}_{trt}) &= \sigma^2 \\ E(\text{MS}_{res}) &= \sigma^2. \end{aligned}$$

   These two facts suggest that when H_0 is true,

$$F = \frac{\text{MS}_{trt}}{\text{MS}_{res}} \approx 1.$$

2. **When H_1 is true** (i.e., the population means are different), then

$$\begin{aligned} E(\text{MS}_{trt}) &> \sigma^2 \\ E(\text{MS}_{res}) &= \sigma^2. \end{aligned}$$

   These two facts suggest that when H_1 is true,

$$F = \frac{\text{MS}_{trt}}{\text{MS}_{res}} > 1.$$

Sampling Distribution: When H_0 is true, the statistic

$$F = \frac{\text{MS}_{trt}}{\text{MS}_{res}} \sim F(t-1, N-t).$$

Recall that the mean of an F distribution is around 1. Therefore,

- Values of F in the center of this distribution are consistent with H_0.
- Large values of F (i.e., out in the right tail) are consistent with H_1.
- **Interesting:** Unusually small values of F (i.e., close to zero) are not necessarily consistent with either hypothesis. This is more likely to occur when there is a violation of our statistical assumptions.
  - correlated individuals within/across samples (most likely), unequal population variances, normality departures, etc.

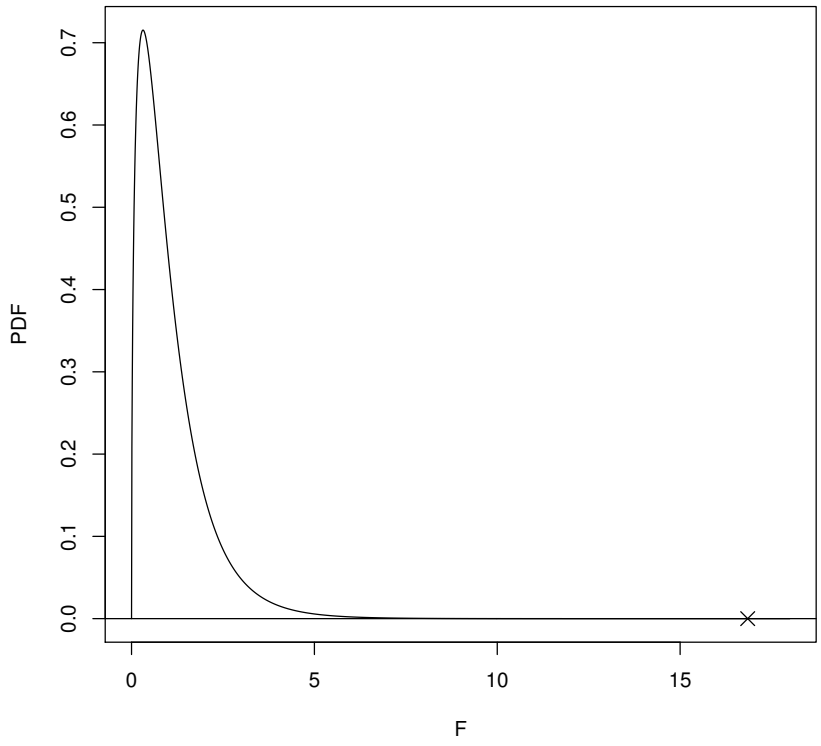

Figure 9.2: $F(3, 32)$ pdf. This is the sampling distribution of F in Example 9.1 when H_0 is true. An "$\times$" at $F = 16.848$ has been added.

Mortar data: We use R to calculate the F statistic in Example 9.1.

```
> anova(lm(strength~mortar.type))
Analysis of Variance Table
Response: strength
            Df  Sum Sq Mean Sq F value    Pr(>F)
mortar.type  3 1520.88  506.96  16.848 9.576e-07 ***
Residuals   32  962.86   30.09
```

Conclusion: There is very strong evidence that at least one of the four mortar strength population means is different. This value of F is not consistent with H_0. It is much more consistent with H_1.

Terminology: It is common to display one-way classification results in an **ANOVA table**. The form of the ANOVA table for the one-way classification is given below:

Source	df	SS	MS	F
Treatments	$t-1$	SS_{trt}	$\text{MS}_{trt} = \frac{\text{SS}_{trt}}{t-1}$	$F = \frac{\text{MS}_{trt}}{\text{MS}_{res}}$
Residuals	$N-t$	SS_{res}	$\text{MS}_{res} = \frac{\text{SS}_{res}}{N-t}$	
Total	$N-1$	SS_{total}		

- In general, it is easy to show that

$$\begin{aligned}\text{SS}_{total} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{++})^2 &= \sum_{i=1}^{t} n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 \\ &= \text{SS}_{trt} + \text{SS}_{res}.\end{aligned}$$

- SS_{total} measures how observations vary about the overall mean, without regard to treatment groups; that is, SS_{total} measures the total variation in all the data.
- SS_{total} can be partitioned into two components:
  - SS_{trt} measures how much of the total variation is due to the treatment groups.
  - SS_{res} measures what is "left over," which we attribute to inherent variation among the individuals.
- Degrees of freedom (df) add down.
- Mean squares (MS) are formed by dividing sums of squares by the corresponding degrees of freedom.
- The ratio of the mean squares (MS) gives the F statistic.

Terminology: The **probability value** (**p-value**) for a hypothesis test measures how much evidence we have against H_0. It is important to remember the following:

$$\textbf{the smaller the p-value} \implies \textbf{the more evidence against } H_0\textbf{.}$$

Mortar data: For the strength/mortar type data in Example 9.1 (from the R output), we see that

$$\text{p-value} \approx 0.0000009576.$$

- This is obviously extremely small which suggests that we have an enormous amount of evidence against H_0.
- In this example, the p-value is calculated as the area to the **right** of $F = 16.848$ on the $F(3, 32)$ probability density function. See Figure 9.2 and it is easy to see why this is so small.
- The p-value is a probability. For the mortar data, the p-value is interpreted as follows:
  - "**If H_0 is true**, the probability we should get a test statistic **equal to or larger** than $F = 16.848$ is 0.0000009576."
- Because this event (i.e., getting an $F \geq 16.848$) is extremely unlikely, this suggests strongly that H_0 is not true.
- In other words, this very small p-value (which comes from a very large F statistic) is more consistent with H_1.

P-value Rules: Probability values are used in more general hypothesis test settings in statistics (not just in one-way classification).
Q: How small does a p-value have to get before we "reject H_0 in favor of H_1?"
A: Unfortunately, there is no right answer to this question. What is commonly done is the following.

- First choose a **significance level** α that is small. This represents the probability that we will reject a true H_0, that is,

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ true}).$$

- Common values of α chosen beforehand are $\alpha = 0.10$, $\alpha = 0.05$ (the most common), and $\alpha = 0.01$.

- The smaller the α is chosen to be, the more evidence one requires to reject H_0. This is a true statement because of the following well-known decision rule:

$$\textbf{p-value} < \alpha \Longrightarrow \textbf{reject } H_0.$$

- Therefore, the value of α chosen by the experimenter (you!) determines how small the p-value must get before H_0 is ultimately rejected.

- For the strength/mortar type data, there is no ambiguity in our decision. For other situations (e.g., p-value = 0.063), the decision may not be as clear cut.

Assumptions/Robustness: There are three main assumptions when performing an analysis of variance:

1. **Independent random samples.**

   - This assumption holding is largely up to the experimenter/investigator; i.e., drawing random samples from the different populations independently (in the case of an observational study) or using randomization to assign individuals to treatments (in an experiment).

2. **Normality.** Each of the t population distributions is normal (Gaussian).

   - This assumption can be assessed empirically using qq plots for each sample separately. Of course, if the sample sizes are small (as in the mortar strength study), these plots may not be all that useful.

   - Thankfully, as with other statistical inference procedures involving means, a one-way ANOVA analysis is robust to normality departures.

3. **Equal population variances.** This is the most important assumption.

- A one-way ANOVA analysis is not robust to departures from this assumption, and it is very critical.
- Therefore, if you suspect the population variances may be markedly different, then you should not use a one-way ANOVA analysis.
- There is a statistical inference procedure that is designed to test the equality of the population variances; i.e., to test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$$

versus

$$H_1 : \text{the population variances } \sigma_i^2 \text{ are not all equal.}$$

The test is called **Bartlett's test**. However, I almost never use this test because it depends critically on the normality assumption. A **nonparametric** version of this test (i.e., one that does not assume normality) is available; it is called **Levene's test**.

9.3 Multiple comparisons/Follow-up analysis

Recall: In a one-way classification, the overall F test is used to test:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$

versus

$$H_1 : \text{the population means } \mu_i \text{ are not all equal.}$$

Note: If we do "reject H_0" in favor of H_1, we conclude that at least one population mean is different. However, we do not know which one(s) or how many. In this light, the decision to reject H_0 is not all that informative or useful.

Follow-up analysis: If H_0 is rejected, the obvious game becomes determining which population mean(s) is(are) different and how they are different. To do this, we will construct **Tukey pairwise confidence intervals** for all population treatment mean differences $\mu_i - \mu_{i'}$,

$1 \leq i < i' \leq t$. If there are t treatments, then there are

$$\binom{t}{2} = \frac{t(t-1)}{2}$$

pairwise confidence intervals to construct. For example, in the mortar strength study (Example 9.1), there are $t = 4$ populations and therefore 6 pairwise intervals:

$$\mu_1 - \mu_2, \qquad \mu_1 - \mu_3, \qquad \mu_1 - \mu_4, \qquad \mu_2 - \mu_3, \qquad \mu_2 - \mu_4, \qquad \mu_3 - \mu_4,$$

where

$$\begin{aligned}
\mu_1 &= \text{population mean strength for mortar type OCM} \\
\mu_2 &= \text{population mean strength for mortar type PIM} \\
\mu_3 &= \text{population mean strength for mortar type RM} \\
\mu_4 &= \text{population mean strength for mortar type PCM.}
\end{aligned}$$

Problem: If we construct multiple confidence intervals (here, 6 of them), and if we construct each one using a $100(1-\alpha)$ percent confidence level, then the overall confidence level in the 6 intervals together will be less than $100(1-\alpha)$ percent. In statistics, this is known as the **multiple comparisons problem**.

- There is a well-known inequality in probability called **Bonferroni's Inequality**, which states that if we have events $A_1, A_2, ..., A_J$, the probability that each event occurs

$$P\left(\bigcap_{j=1}^{J} A_j\right) \geq \sum_{j=1}^{J} P(A_j) - (J-1).$$

- To see how this inequality can be used in our current discussion, define the event

$$A_j = \{j\text{th confidence interval includes its population mean difference}\},$$

for $j = 1, 2, ..., J$. The event

$$\bigcap_{j=1}^{J} A_j = \{\text{each of the } J \text{ intervals includes its population mean difference}\}.$$

- In this light, consider the following table, which contains a lower bound on how small this probability can be (for different values of t and J). This table assumes that each pairwise interval has been constructed at the nominal $1-\alpha = 0.95$ level.

# of treatments t	# of intervals $J = \binom{t}{2}$	Lower bound
3	3	$3(0.95) - 2 = 0.85$
4	6	$6(0.95) - 5 = 0.70$
5	10	$10(0.95) - 9 = 0.50$
6	15	$15(0.95) - 14 = 0.25$
$\vdots$	$\vdots$	$\vdots$
10	45	$45(0.95) - 44 = -1.25$!!

  Therefore, with $t = 4$ treatments (populations), the probability that each of the 6 95 percent intervals will contain its population mean difference can be as low as 0.7! For larger experiments with more treatments, this probability is even lower!! Clearly, we have to do something to address this.

Goal: Construct confidence intervals for all pairwise intervals $\mu_i - \mu_{i'}$, $1 \leq i < i' \leq t$, and have our **family-wise confidence level** still be at $100(1-\alpha)$ percent. By "family-wise," we mean that our level of confidence applies to the collection of all $\binom{t}{2}$ intervals (not to the intervals individually).

Solution: Increase the confidence level associated with each individual interval. **Tukey's method** is designed to do this. The intervals are of the form:

$$(\overline{Y}_{i+} - \overline{Y}_{i'+}) \pm q_{t,N-t,\alpha}\sqrt{\text{MS}_{res}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)},$$

where $q_{t,N-t,\alpha}$ is the **Tukey quantile** that guarantees a **family-wise confidence level** of $100(1-\alpha)$ percent.

Mortar data: We use R to construct the Tukey confidence intervals. The family-wise confidence level is 95 percent:

```
> TukeyHSD(aov(lm(strength~mortar.type)),conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(strength ~ mortar.type))

$mortar.type
            diff       lwr       upr     p adj
PCM-OCM  2.48000 -4.950955  9.910955 0.8026758
PIM-OCM 15.21575  8.166127 22.265373 0.0000097
RM-OCM  12.99875  5.949127 20.048373 0.0001138
PIM-PCM 12.73575  5.686127 19.785373 0.0001522
RM-PCM  10.51875  3.469127 17.568373 0.0016850
RM-PIM  -2.21700 -8.863448  4.429448 0.8029266
```

Note: In the R output, the columns labeled `lwr` and `upr` give, respectively, the lower and upper limits of the pairwise confidence intervals.

- `PCM-OCM`: We are (at least) 95 percent confident that the difference in the population mean strengths for the PCM and OCM mortars is between -4.95 and 9.91 MPa.
  - This confidence interval includes "0," so we cannot conclude these two population means are different.
  - An equivalent finding is that the **adjusted p-value** for these two mortar types, given in the `p adj` column, is large (0.803).
- `PIM-OCM`: We are (at least) 95 percent confident that the difference in the population mean strengths for the PIM and OCM mortars is between 8.17 and 22.27 MPa.
  - This confidence interval does not include "0" and contains only positive values. This suggests that the population mean strength of the PIM mortar is greater than the population mean strength of the OCM mortar.

  - An equivalent finding is that the **adjusted p-value** for these two mortar types, given in the `p adj` column, is very small (<0.001).

- Interpretations for the remaining 4 confidence intervals are written similarly.

- The main point is this:

  - If a pairwise confidence interval (for two population means) includes "0," then these population means are not declared to be different.

  - If a pairwise interval does not include "0," then the population means are declared to be different.

  - The conclusions we make for **all possible pairwise comparisons** are at the $100(1-\alpha)$ percent confidence level.

Mortar data: The following pairs of population means are declared to be different:

`PIM-OCM` `RM-OCM` `PIM-PCM` `RM-PCM`.

The following pairs of population means are declared to be not different:

`PCM-OCM` `RM-PIM`.

We can therefore conclude:

- The `PIM` and `RM` population mean strengths are larger than the `OCM` and `PCM` population mean strengths.

- The `PIM` and `RM` population mean strengths are not different.

- The `OCM` and `PCM` population mean strengths are not different.

Furthermore, we have an overall (family-wise) confidence level of 95 percent that all of our conclusions are correct. Had we not used an adjusted analysis based on Tukey's method (e.g., just calculate all unadjusted pairwise intervals), our overall confidence level would have been much lower (as low as 70 percent).

10 Simple Linear Regression

10.1 Introduction

Importance: A problem arising in engineering, economics, medicine, and other areas, is that of investigating the relationship between two or more variables. In such settings, the goal is to model a random variable Y (often continuous) as a function of one or more independent variables, say, $x_1, x_2, ..., x_k$. Mathematically, we can express this model as

$$Y = g(x_1, x_2, ..., x_k) + \epsilon,$$

where $g : \mathbb{R}^k \to \mathbb{R}$ is a function (whose form may or may not be specified). This is called a **regression model**.

- The presence of the (random) error ϵ conveys the fact that the relationship between the dependent variable Y and the independent variables $x_1, x_2, ..., x_k$ through g is not deterministic. Instead, the term ϵ "absorbs" all variation in Y that is not explained by $g(x_1, x_2, ..., x_k)$.

Terminology: In this course, we will consider models of the form

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}_{g(x_1, x_2, ..., x_k)} + \epsilon,$$

that is, g is a linear function of $\beta_0, \beta_1, \beta_2, ..., \beta_k$. We call this a **linear regression model**.

- The **response variable** Y is random (but we do get to observe its value).
- The **independent variables** $x_1, x_2, ..., x_k$ are fixed (and observed).
- The **regression parameters** $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are unknown. These are to be estimated on the basis of the observed data.
- The **error term** ϵ is random (and not observed).

Terminology: More precisely, we call a regression model a **linear regression model** if the regression parameters enter the g function in a linear fashion. For example, each of the models is a linear regression model:

$$\begin{aligned} Y &= \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}_{g(x_1, x_2)} + \epsilon. \end{aligned}$$

The term "linear" does not refer to the shape of the regression function g. It refers to how the regression parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$ enter the g function.

Important: Regression models (linear or otherwise) are models for a **population** of individuals. From a statistical inference standpoint, our goal is the same as in previous chapters. We will use sample information to estimate the population parameters in the model. We say that we are "estimating" or "fitting the model" with the observed data.

10.2 Simple linear regression model

Terminology: A **simple linear regression model** includes only one independent variable x and is of the form

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

The population regression function $g(x) = \beta_0 + \beta_1 x$ is a straight line with intercept β_0 and slope β_1. These parameters describe the population of individuals for which this model is assumed.

Note: If $E(\epsilon) = 0$, then

$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 x + \epsilon) \\ &= \beta_0 + \beta_1 x + E(\epsilon) \\ &= \beta_0 + \beta_1 x. \end{aligned}$$

Therefore, we have the following interpretations for the population regression parameters β_0 and β_1:

- β_0 quantifies the population mean of Y when $x = 0$.
- β_1 quantifies the population-level change in $E(Y)$ brought about by a one-unit change in x.

Example 10.1. As part of a waste removal project, a new compression machine for processing sewage sludge is being studied. Engineers are interested in the following variables:

$$\begin{aligned} Y &= \text{moisture control of compressed pellets (measured as a percent)} \\ x &= \text{machine filtration rate (kg-DS/m/hr).} \end{aligned}$$

Engineers collect observations of (x, Y) from a random sample of $n = 20$ sewage specimens; the data are given below.

Obs	x	Y	Obs	x	Y
1	125.3	77.9	11	159.5	79.9
2	98.2	76.8	12	145.8	79.0
3	201.4	81.5	13	75.1	76.7
4	147.3	79.8	14	151.4	78.2
5	145.9	78.2	15	144.2	79.5
6	124.7	78.3	16	125.0	78.1
7	112.2	77.5	17	198.8	81.5
8	120.2	77.0	18	132.5	77.0
9	161.2	80.1	19	159.6	79.0
10	178.9	80.2	20	110.7	78.6

Table 10.1: Sewage data. Moisture (Y, measured as a percentage) and machine filtration rate (x, measured in kg-DS/m/hr). There are $n = 20$ observations.

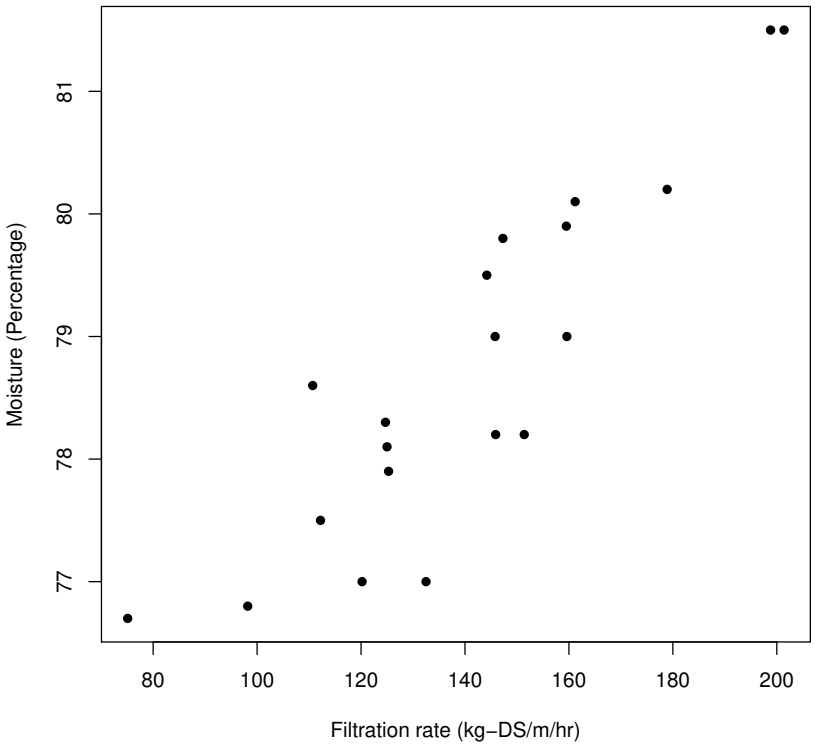

Figure 10.1: Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr).

Figure 10.1 displays the sample data in a **scatterplot**. This sample information suggests the variables Y and x are **linearly related**, although there is a large amount of variation that is unexplained.

- This unexplained variability could arise from other independent variables (e.g., applied temperature, pressure, sludge mass, etc.) that also influence the moisture percentage Y but are not present in the model.
- It could also arise from measurement error or just random variation in the sludge compression process.

Inference: What does the sample information suggest about the population? Do we have evidence that Y and x are linearly related in the population?

10.3 Least squares estimation

Terminology: When we say, "fit a regression model," we mean that we are estimating the population regression parameters in the model with the observed sample information (data). Suppose we have a random sample of observations (x_i, Y_i), $i = 1, 2, ..., n$, and postulate the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, 2, ..., n$. Our first goal is to estimate β_0 and β_1. Formal assumptions for the error terms ϵ_i will be given later.

Terminology: The most common method of estimating the population parameters β_0 and β_1 is least squares. The **method of least squares** says to choose the values of β_0 and β_1 that minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Denote the least squares estimators by b_0 and b_1, respectively, that is, the values of β_0 and β_1 that minimize $Q(\beta_0, \beta_1)$. A two-variable calculus minimization argument can be used to find expressions for b_0 and b_1. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\begin{aligned} \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \stackrel{\text{set}}{=} 0 \\ \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i \stackrel{\text{set}}{=} 0. \end{aligned}$$

Solving for β_0 and β_1 gives the **least squares estimators**

$$\begin{aligned} b_0 &= \overline{Y} - b_1 \overline{x} \\ b_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\text{SS}_{xy}}{\text{SS}_{xx}}. \end{aligned}$$

The estimated model is written as follows:

$$\widehat{Y} = b_0 + b_1 x.$$

Example 10.1 (continued). We use R to calculate the equation of the least squares regression line for the sewage data in Example 10.1. Here is the output:

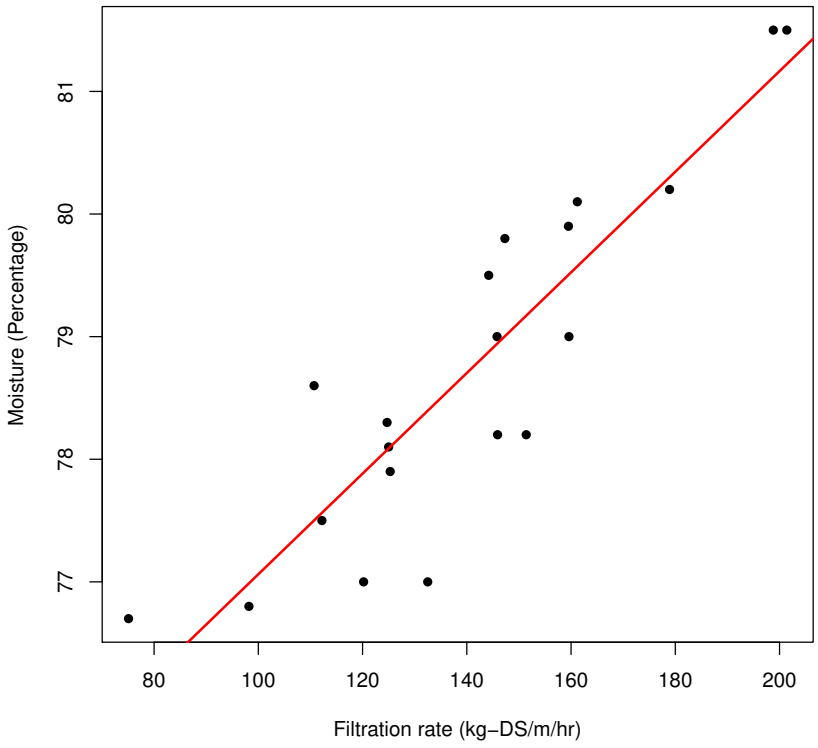

Figure 10.2: Scatterplot of pellet moisture Y (measured as a percentage) as a function of filtration rate x (measured in kg-DS/m/hr). The least squares line has been added.

```
> fit = lm(moisture~filtration.rate)
> fit
lm(formula = moisture ~ filtration.rate)
Coefficients:
    (Intercept)  filtration.rate
       72.95855          0.04103
```

The least squares estimates (to 3 dp) for the sewage data are

$$b_0 = 72.959$$
$$b_1 = 0.041.$$

The estimated model is

$$\widehat{Y} = 72.959 + 0.041x,$$

or, in other words,

$$\widehat{\texttt{Moisture}} = 72.959 + 0.041 \ \texttt{Filtration rate}.$$

Note: The estimated model is also called the **prediction equation**. This is because we can now predict the value of Y (moisture percentage) for a given value of x (filtration rate). For example, when the filtration rate is $x = 150$ kg-DS/m/hr, we would predict the moisture percentage to be

$$\widehat{Y}(150) = 72.959 + 0.041(150) \approx 79.11.$$

Of course, this prediction comes directly from the sample of observations used to fit the regression model. Therefore, we will eventually want to quantify the **uncertainty** in this prediction; e.g., how variable is this prediction?

10.4 Model assumptions and sampling distributions

Interest: We investigate the properties of the least squares estimators b_0 and b_1 as estimators of the population-level regression parameters β_0 and β_1 in the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$. To do this, we need statistical assumptions on the errors ϵ_i.

Assumptions: We will assume throughout that

- $E(\epsilon_i) = 0$, for $i = 1, 2, ..., n$
- $\text{var}(\epsilon_i) = \sigma^2$, for $i = 1, 2, ..., n$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are normally distributed.

Results: Under the assumptions stated on the previous page, we can derive the following results for the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

- **Result 1:**

  $$Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

  In other words, the response variable Y is normally distributed with mean $\beta_0 + \beta_1 x$ and variance σ^2. Note that the population mean of Y depends on x. The population variance of Y does not depend on x.

- **Result 2.** The least squares estimators b_0 and b_1 are **unbiased estimators** of β_0 and β_1, respectively, that is,

  $$\begin{aligned} E(b_0) &= \beta_0 \\ E(b_1) &= \beta_1. \end{aligned}$$

- **Result 3.** The least squares estimators b_0 and b_1 have normal sampling distributions; specifically,

  $$b_0 \sim \mathcal{N}(\beta_0, c_{00}\sigma^2) \quad \text{and} \quad b_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2),$$

  where

  $$c_{00} = \frac{1}{n} + \frac{\overline{x}^2}{\text{SS}_{xx}} \quad \text{and} \quad c_{11} = \frac{1}{\text{SS}_{xx}}.$$

  These distributions are needed to write confidence intervals and perform hypothesis tests for β_0 and β_1 (i.e., to perform statistical inference for the population).

10.5 Estimating the error variance

Goal: In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we now turn our attention to estimating σ^2, the **error variance**.

Recall: As we did in estimating β_0 and β_1 (the population level regression parameters), we will use the observed data (x_i, Y_i), $i = 1, 2, ..., n$, to estimate the error variance σ^2. The error variance is also a population level parameter and quantifies how variable the population is for a given model.

Terminology: Define the ith **fitted value** by

$$\widehat{Y}_i = b_0 + b_1 x_i,$$

where b_0 and b_1 are the least squares estimators. Each observation has its own fitted value. Define the ith **residual** by

$$e_i = Y_i - \widehat{Y}_i.$$

Each observation has its own residual.

Sewage data: I calculated the fitted values and residuals for each observation:

Obs	x	Y	$\widehat{Y} = b_0 + b_1 x$	$e = Y - \widehat{Y}$	Obs	x	Y	$\widehat{Y} = b_0 + b_1 x$	$e = Y - \widehat{Y}$
1	125.3	77.9	78.100	−0.200	11	159.5	79.9	79.503	0.397
2	98.2	76.8	76.988	−0.188	12	145.8	79.0	78.941	0.059
3	201.4	81.5	81.223	0.277	13	75.1	76.7	76.040	0.660
4	147.3	79.8	79.003	0.797	14	151.4	78.2	79.171	−0.971
5	145.9	78.2	78.945	−0.745	15	144.2	79.5	78.876	0.624
6	124.7	78.3	78.075	0.225	16	125.0	78.1	78.088	0.012
7	112.2	77.5	77.563	−0.062	17	198.8	81.5	81.116	0.384
8	120.2	77.0	77.891	−0.891	18	132.5	77.0	78.396	−1.396
9	161.2	80.1	79.573	0.527	19	159.6	79.0	79.508	−0.508
10	178.9	80.2	80.299	−0.099	20	110.7	78.6	77.501	1.099

Table 10.2: Sewage data. Fitted values and residuals from the least squares fit.

Note that

- If an observation's Y value is above the least squares regression line, then $Y_i > \widehat{Y}_i$ and its residual e_i is positive.
- If an observation's Y value is below the least squares regression line, then $Y_i < \widehat{Y}_i$ and its residual e_i is negative.

- If an observation's Y value is on the least squares regression line, then $Y_i = \widehat{Y}_i$ and its residual e_i is zero.

Interesting fact: In our simple linear regression model,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i) = 0.$$

That is, the residuals sum to zero. For the sewage data in Example 10.1,

```
# Calculate fitted values and residuals
fitted.values = predict(fit)
residuals = moisture-fitted.values
# Show residuals sum to 0
> sum(residuals)
[1] 2.273737e-13
```

Terminology: Define the **residual sum of squares** by

$$\begin{aligned} \text{SS}_{res} &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2. \end{aligned}$$

In the simple linear regression model, the **residual mean squares**

$$\text{MS}_{res} = \frac{\text{SS}_{res}}{n-2}$$

is an unbiased estimator of σ^2, that is,

$$E(\text{MS}_{res}) = \sigma^2.$$

The quantity

$$\widehat{\sigma} = \sqrt{\text{MS}_{res}} = \sqrt{\frac{\text{SS}_{res}}{n-2}}$$

estimates σ and is called the **residual standard error**.

Sewage data: To illustrate for the sewage data in Example 10.1,

```
# Calculate MSres
MSres = sum(residuals^2)/(length(moisture)-2)
> MSres
[1] 0.4426659
# Calculate residual standard error
resid.std.error = sqrt(MSres)
> resid.std.error
[1] 0.6653314
```

In Chapter 11, we will see that R calculates these values automatically (as part of a regression analysis of variance).

10.6 Statistical inference for β_0 and β_1

Interest: In the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

we now discuss the formal question:

> "What does the sample information from an estimated regression model suggest about the population?"

In other words, we pursue **statistical inference** for the population level regression parameters β_0 and β_1.

- In practice, inference for the slope parameter β_1 is of primary interest because of its connection to the independent variable x in the model. For example, if $\beta_1 = 0$, then Y and x are not linearly related in the population.
- Statistical inference for β_0 is less meaningful, unless one is explicitly interested in the mean of Y when $x = 0$. We will not pursue this.

Confidence interval: Under our regression model assumptions, the following sampling distribution arises:

$$t = \frac{b_1 - \beta_1}{\sqrt{\frac{\text{MS}_{res}}{\text{SS}_{xx}}}} \sim t(n-2).$$

This result can be used to derive a $100(1-\alpha)$ **percent confidence interval** for β_1, which is given by

$$b_1 \pm t_{n-2,\alpha/2}\sqrt{\frac{\text{MS}_{res}}{\text{SS}_{xx}}}.$$

- The value $t_{n-2,\alpha/2}$ is the upper $\alpha/2$ quantile from the $t(n-2)$ distribution.
- Note the form of the interval:

$$\underbrace{\text{point estimate}}_{b_1} \pm \underbrace{\text{quantile}}_{t_{n-2,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\text{MS}_{res}}{\text{SS}_{xx}}}}.$$

  We interpret the interval in the same way:

  > "We are $100(1-\alpha)$ percent confident that the population parameter β_1 is in this interval."

- Of particular interest is the value $\beta_1 = 0$:
  - If the confidence interval for β_1 contains "0," this suggests that Y and x are not linearly related in the population.
  - If the confidence interval for β_1 does not contain "0," this suggests that Y and x are linearly related in the population.

Sewage data: We can use the `confint` function in R to calculate a 95 percent confidence interval for β_1:

```
> confint(fit,level=0.95)
                      2.5 %      97.5 %
(Intercept)     71.49309400 74.42399995
filtration.rate  0.03087207  0.05119547
```

Interpretation: We are 95 percent confident that the population parameter β_1 is between 0.0309 and 0.0511. This means

- for every one unit increase in the machine filtration rate x, we are 95 percent confident that the population mean absorption $E(Y)$ will increase between 0.0309 and 0.0511 percent.

Note that this interval does not contain "0" and includes only positive values. There is strong evidence that the absorption rate Y is positively linearly related to machine filtration rate x in the population. The confidence interval gives information about how strong this relationship is.

Hypothesis test: Under our regression model assumptions, if we wanted to formally test

$$H_0 : \beta_1 = 0$$

versus

$$H_1 : \beta_1 \neq 0,$$

we would use

$$t = \frac{b_1}{\sqrt{\frac{\text{MS}_{res}}{\text{SS}_{xx}}}}$$

as a test statistic and reject H_0 if the corresponding p-value was small.

Sewage data: We use the `summary` function in R to perform this hypothesis test:

```
> summary(fit)
lm(formula = moisture ~ filtration.rate)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     72.958547   0.697528 104.596  < 2e-16 ***
filtration.rate  0.041034   0.004837   8.484 1.05e-07 ***
```

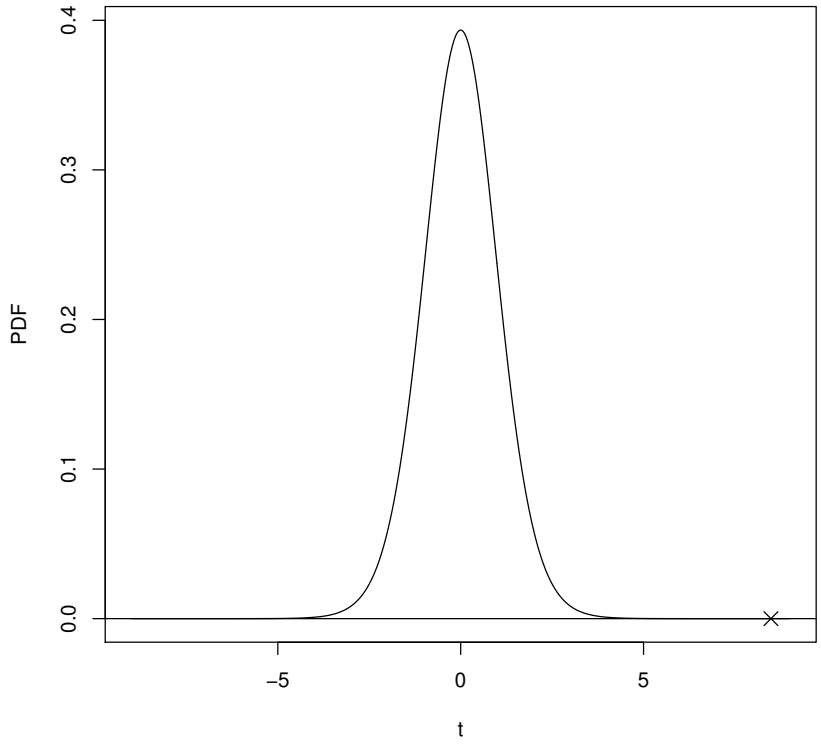

Figure 10.3: Sewage data: $t(18)$ pdf. This is the sampling distribution of t when $H_0 : \beta_1 = 0$ is true. An "$\times$" at $t = 8.484$ has been added.

Interpretation: For the sewage data,

$$t = \frac{b_1}{\sqrt{\frac{\text{MS}_{res}}{\text{SS}_{xx}}}} = \frac{0.041034}{0.004837} = 8.484.$$

Figure 10.3 shows that $t = 8.484$ is not an expected outcome from the $t(18)$ distribution, the sampling distribution of

$$t = \frac{b_1}{\sqrt{\frac{\text{MS}_{res}}{\text{SS}_{xx}}}}$$

when $H_0 : \beta_1 = 0$ is true. The p-value for the test is

$$\text{p-value} = 0.000000105.$$

This is strong evidence against H_0. There is strong evidence that the absorption percentage Y is positively linearly related to machine filtration rate x in the population.

10.7 Confidence and prediction intervals for a given $x = x_0$

Interest: Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

We are often interested in learning about the response Y at a certain setting of the independent variable, say $x = x_0$. For the sewage data, for example, suppose we are interested in the moisture percentage Y when the filtration rate is $x = 150$ kg-DS/m/hr. Two potential goals arise:

- We might be interested in **estimating the population mean** of Y when $x = x_0$. This mean response is denoted by $E(Y|x_0)$. This is the mean of the following probability distribution:
$$\mathcal{N}(\beta_0 + \beta_1 x_0, \sigma^2).$$
- We might be interested in **predicting a new response** Y when $x = x_0$. This predicted response is denoted by $Y^*(x_0)$. This is a new value from the following probability distribution:
$$\mathcal{N}(\beta_0 + \beta_1 x_0, \sigma^2).$$

Difference: In the first problem, we are **estimating** the mean of a distribution. In the second problem, we are **predicting** the value of a new response from this distribution. The second problem is more difficult than the first.

Goals: We would like to create $100(1-\alpha)$ percent intervals for the population mean $E(Y|x_0)$ and for the new response $Y^*(x_0)$. The former is called a **confidence interval**. The latter is called a **prediction interval**.

Point Estimator/Predictor: To construct either interval, we start with the same quantity:

$$\widehat{Y}(x_0) = b_0 + b_1 x_0,$$

where b_0 and b_1 are the least squares estimates from the fit of the model.

- In the confidence interval for $E(Y|x_0)$, we call $\widehat{Y}(x_0)$ a **point estimator**.
- In the prediction interval for $Y(x_0)$, we call $\widehat{Y}(x_0)$ a **point predictor**.

The primary difference in the intervals arises in assessing the variability of $\widehat{Y}(x_0)$.

Confidence interval: A $100(1-\alpha)$ percent confidence interval for the population mean $E(Y|x_0)$ is given by

$$\widehat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{\text{MS}_{res}\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\text{SS}_{xx}}\right]}.$$

Prediction interval: A $100(1-\alpha)$ percent prediction interval for the new response $Y^*(x_0)$ is given by

$$\widehat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{\text{MS}_{res}\left[1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\text{SS}_{xx}}\right]}.$$

- **Comparison:** The two intervals have the same form and are nearly identical.
  - The extra "1" in the prediction interval's standard error arises from the additional uncertainty associated with predicting a new response from the $\mathcal{N}(\beta_0+\beta_1 x_0, \sigma^2)$ distribution.
  - Therefore, at the same value of x_0, a $100(1-\alpha)$ percent prediction interval for $Y^*(x_0)$ will necessarily be **wider** than the corresponding $100(1-\alpha)$ percent confidence interval for $E(Y|x_0)$.
- **Interval length:** The length of both intervals depends on the value of x_0.
  - The standard error in either interval will be smallest when $x_0 = \overline{x}$ and will get larger the farther x_0 is from $\overline{x}$ in either direction.
  - This implies that the precision with which we estimate $E(Y|x_0)$ or predict $Y^*(x_0)$ decreases the farther we get away from $\overline{x}$.
  - This makes intuitive sense, namely, we would expect to have the most "confidence" in our fitted model near the "center" of the observed data.

- **Warning:** It is sometimes desired to estimate $E(Y|x_0)$ or predict $Y^*(x_0)$ for values of x_0 outside the range of x values used in the study. This is called **extrapolation** and can be very dangerous.
  - In order for our inferences to be valid, we must believe that the model holds for x values outside the range where we have observed data.
  - In some situations, this may be reasonable. In others, we may have no theoretical basis for making such a claim without data to support it.

Example 10.1 (continued). In our sewage example, suppose that we are interested in estimating $E(Y|x_0)$ and predicting a new $Y^*(x_0)$ when the filtration rate is $x_0 = 150$ kg-DS/m/hr.

- $E(Y|x_0)$ denotes the population mean moisture percentage when the machine filtration rate is $x_0 = 150$ kg-DS/m/hr.
- $Y^*(x_0)$ denotes the moisture percentage Y for an individual sludge specimen when the filtration rate is $x_0 = 150$ kg-DS/m/hr.
- R automates the calculation of confidence and prediction intervals, as seen below.

```
> predict(fit,data.frame(filtration.rate=150),level=0.95,interval="confidence")
       fit      lwr      upr
  79.11361 78.78765 79.43958
> predict(fit,data.frame(filtration.rate=150),level=0.95,interval="prediction")
       fit     lwr      upr
  79.11361 77.6783 80.54893
```

- Note that the point estimate (point prediction) is easily calculated:

$$\widehat{Y}(x_0 = 150) = 72.959 + 0.041(150) \approx 79.11361.$$

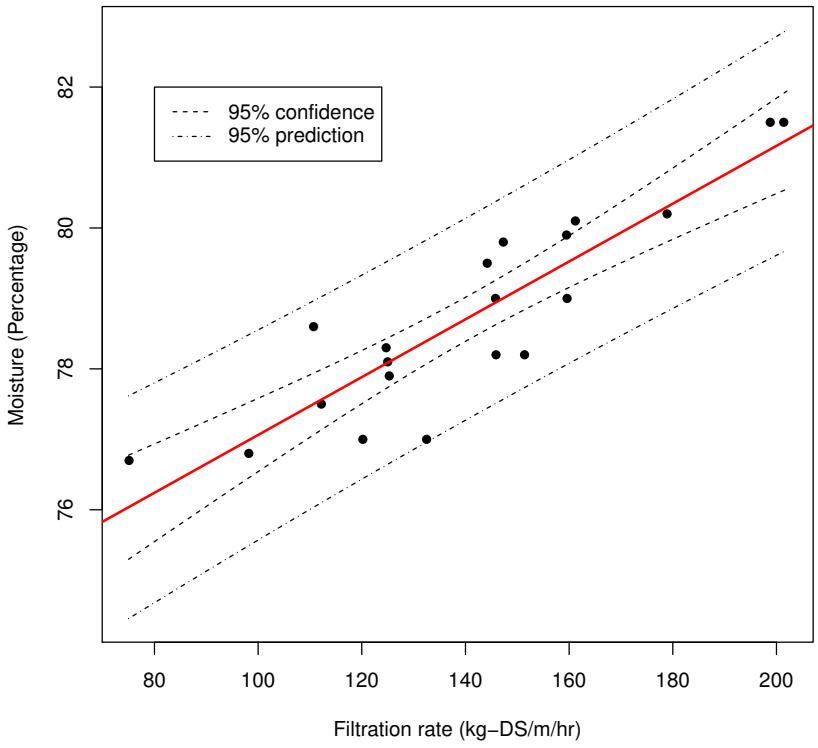

Figure 10.4: Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr). The least squares regression line has been added. Ninety-five percent confidence/prediction bands have been added.

- A 95 percent **confidence interval** for $E(Y|x_0 = 150)$ is $(78.79, 79.44)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95 percent confident that the population mean moisture percentage is between 78.79 and 79.44 percent.

- A 95 percent **prediction interval** for $Y^*(x_0 = 150)$ is $(77.68, 80.55)$. When the filtration rate is $x_0 = 150$ kg-DS/m/hr, we are 95 percent confident that the moisture percentage for a single specimen will be between 77.68 and 80.55 percent.

- Figure 10.4 shows 95 percent confidence bands for $E(Y|x_0)$ and 95 percent prediction bands for $Y^*(x_0)$. These are not simultaneous bands (i.e., these are not bands for the entire population regression function).

11 Multiple Linear Regression

11.1 Introduction

Preview: We have considered the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We now extend this basic model to include multiple independent variables $x_1, x_2, ..., x_k$. This is more realistic because Y often depends on multiple variables (not just one). Specifically, we consider models of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

We call this a **multiple linear regression model**.

- There are now $p = k + 1$ regression parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$.
  - In simple linear regression, $k = 1$ and $p = 2$.
- The regression parameters describe the **population** for which this model is applicable. They are unknown and are to be estimated with the observed data; i.e., based on a sample from the population.
- We continue to assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- We also assume that the independent variables $x_1, x_2, ..., x_k$ are fixed and are measured without error.

Example 11.1. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study from the LaTrobe Valley of Victoria, Australia, specimens of cheddar cheese were analyzed for their chemical composition and were subjected to taste tests. For each specimen, the taste Y was obtained by combining the scores from

Specimen	TASTE	ACETIC	H2S	LACTIC	Specimen	TASTE	ACETIC	H2S	LACTIC
1	12.3	4.543	3.135	0.86	16	40.9	6.365	9.588	1.74
2	20.9	5.159	5.043	1.53	17	15.9	4.787	3.912	1.16
3	39.0	5.366	5.438	1.57	18	6.4	5.412	4.700	1.49
4	47.9	5.759	7.496	1.81	19	18.0	5.247	6.174	1.63
5	5.6	4.663	3.807	0.99	20	38.9	5.438	9.064	1.99
6	25.9	5.697	7.601	1.09	21	14.0	4.564	4.949	1.15
7	37.3	5.892	8.726	1.29	22	15.2	5.298	5.220	1.33
8	21.9	6.078	7.966	1.78	23	32.0	5.455	9.242	1.44
9	18.1	4.898	3.850	1.29	24	56.7	5.855	10.20	2.01
10	21.0	5.242	4.174	1.58	25	16.8	5.366	3.664	1.31
11	34.9	5.740	6.142	1.68	26	11.6	6.043	3.219	1.46
12	57.2	6.446	7.908	1.90	27	26.5	6.458	6.962	1.72
13	0.7	4.477	2.996	1.06	28	0.7	5.328	3.912	1.25
14	25.9	5.236	4.942	1.30	29	13.4	5.802	6.685	1.08
15	54.9	6.151	6.752	1.52	30	5.5	6.176	4.787	1.25

Table 11.1: Cheese data. ACETIC, H2S, and LACTIC are independent variables. The response variable is TASTE.

several tasters. Data were collected on the following variables:

$$\begin{aligned} Y &= \text{taste score (TASTE)} \\ x_1 &= \text{concentration of acetic acid (ACETIC)} \\ x_2 &= \text{concentration of hydrogen sulfide (H2S)} \\ x_3 &= \text{concentration of lactic acid (LACTIC).} \end{aligned}$$

The variables ACETIC and H2S were measured on the log scale. The variable LACTIC has not been transformed. Table 11.1 contains concentrations of the chemicals in a random sample of $n = 30$ specimens of cheddar cheese and the corresponding taste scores. Researchers postulate that each of the three variables x_1, x_2, and x_3 is important in describing TASTE and consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

to model this relationship.

11.2 Least squares estimation

Data: Suppose we have a random sample of n individuals from a population. In a multiple linear regression application, we can envision the observed data as follows:

Individual	Y	x_1	x_2	$\cdots$	x_k
1	Y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	Y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Each of the n individuals contributes a response Y and a value of each of the independent variables. The value

$$x_{ij} = \text{ measurement on the } j\text{th independent variable for the } i\text{th individual,}$$

for $i = 1, 2, ..., n$ and $j = 1, 2, ..., k$. For the n individuals, we write

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$.

Matrix representation: To estimate the population parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$, we again use least squares. In doing so, it is advantageous to express multiple linear regression models in terms of matrices and vectors. This streamlines notation and makes the presentation easier. Define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

With these definitions, the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, can be expressed equivalently as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In this representation,

- $\mathbf{Y}$ is an $n \times 1$ (random) vector of responses
- $\mathbf{X}$ is an $n \times p$ (fixed) matrix of independent variable measurements ($p = k + 1$)
- $\boldsymbol{\beta}$ is a $p \times 1$ (fixed) vector of unknown population regression parameters
- $\boldsymbol{\epsilon}$ is an $n \times 1$ (random) vector of unobserved errors.

Illustration: Here are $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ for the cheese data in Example 11.1. Recall there are $n = 30$ individuals and $k = 3$ independent variables. The data are in Table 11.1.

$$\mathbf{Y} = \begin{pmatrix} 12.3 \\ 20.9 \\ \vdots \\ 5.5 \end{pmatrix}_{30\times 1} \quad \mathbf{X} = \begin{pmatrix} 1 & 4.543 & 3.135 & 0.86 \\ 1 & 5.159 & 5.043 & 1.53 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.176 & 4.787 & 1.25 \end{pmatrix}_{30\times 4} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}_{4\times 1} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{30} \end{pmatrix}_{30\times 1}.$$

Least Squares: The notion of least squares is the same in multiple linear regression as it was in simple linear regression. Specifically, we want to find the values of $\beta_0, \beta_1, \beta_2, ..., \beta_k$ that minimize

$$Q(\beta_0, \beta_1, \beta_2, ..., \beta_k) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})]^2.$$

First recognize that

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

is the inner (dot) product of the ith row of $\mathbf{X}$ and $\boldsymbol{\beta}$. Therefore,

$$Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

is the ith entry in the difference vector $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. The objective function Q is

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

the inner (dot) product of $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ with itself; i.e., the squared length of $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$.

Solution: We want to find the value of $\boldsymbol{\beta}$ that minimizes $Q(\boldsymbol{\beta})$. Because $Q(\boldsymbol{\beta})$ is a scalar function of the $p = k+1$ elements of $\boldsymbol{\beta}$, it is possible to use calculus to determine the values of the p elements that minimize it. Formally, we can take p partial derivatives, one with respect to each of $\beta_0, \beta_1, \beta_2, ..., \beta_k$, and set these equal to zero. Using the calculus of matrices, we can write this resulting system of p equations (and p unknowns) as follows:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

These are called the **normal equations**. Provided that $\mathbf{X}'\mathbf{X}$ is full rank, the (unique) solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}.$$

This is the **least squares estimator** of $\boldsymbol{\beta}$.

Technical note: For the least squares estimator

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

to be unique, we need $\mathbf{X}$ to be of **full column rank**; i.e., $r(\mathbf{X}) = p = k+1$. This will occur when there are no linear dependencies among the columns of $\mathbf{X}$. If $r(\mathbf{X}) < p$, then $\mathbf{X}'\mathbf{X}$ does not have a unique inverse, and the normal equations can not be solved uniquely. Statistical software such as R will alert you when $\mathbf{X}'\mathbf{X}$ is not full rank.

Cheese data: We now use R to calculate the least squares estimate $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ for the cheese data in Example 11.1:

```
> fit = lm(taste~acetic+h2s+lactic)
> fit
Coefficients:
(Intercept)       acetic          h2s       lactic
    -28.877        0.328        3.912       19.670
```

This output gives the value of the least squares estimate

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} -28.877 \\ 0.328 \\ 3.912 \\ 19.670 \end{pmatrix}.$$

Therefore, the estimated regression model based on the data is

$$\widehat{Y} = -28.877 + 0.328x_1 + 3.912x_2 + 19.670x_3,$$

or, in other words,

$$\widehat{\texttt{TASTE}} = -28.877 + 0.328\ \texttt{ACETIC} + 3.912\ \texttt{H2S} + 19.670\ \texttt{LACTIC}.$$

11.3 Estimating the error variance

Goal: In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we now turn our attention to estimating σ^2, the **error variance**.

Terminology: The **residual sum of squares** is given by

$$\text{SS}_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2,$$

just as in simple linear regression. In matrix notation, we can write this as

$$\begin{aligned} \text{SS}_{res} &= (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) \\ &= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{e}'\mathbf{e}. \end{aligned}$$

- The $n \times 1$ vector $\widehat{\mathbf{Y}} = \mathbf{Xb}$ contains the least squares **fitted values**.
- The $n \times 1$ vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$ contains the least squares **residuals**.
- R calculates these upon request; e.g., `fitted.values = predict(fit)`.

Result: The **residual mean squares**

$$\text{MS}_{res} = \frac{\text{SS}_{res}}{n-p}$$

is an unbiased estimator of σ^2, that is,

$$E(\text{MS}_{res}) = \sigma^2.$$

The quantity

$$\widehat{\sigma} = \sqrt{\text{MS}_{res}} = \sqrt{\frac{\text{SS}_{res}}{n-p}}$$

estimates σ and is called the **residual standard error**. This result is analogous to the simple linear regression result (see pp 161). The only difference is in the divisor in MS_{res}.

11.4 Analysis of variance for linear regression

Identity: The following algebraic identity arises for a linear regression model fit (simple or multiple):

$$\begin{aligned}\text{SS}_{total} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \\ &= \text{SS}_{reg} + \text{SS}_{res}.\end{aligned}$$

This information is used to produce an **analysis of variance** (**ANOVA**) table.

Table 11.2: Analysis of variance table for linear regression.

Source	df	SS	MS	F
Regression	$p-1$	SS_{reg}	$\text{MS}_{reg} = \frac{\text{SS}_{reg}}{p-1}$	$F = \frac{\text{MS}_{reg}}{\text{MS}_{res}}$
Residual	$n-p$	SS_{res}	$\text{MS}_{res} = \frac{\text{SS}_{res}}{n-p}$	
Total	$n-1$	SS_{total}		

Notes:

- This table summarizes how the variability in the response data is **partitioned**.

- SS_{total} is the **total sum of squares**. It measures the total variation in the response data.
- SS_{reg} is the **regression sum of squares**. It measures the variation in the response data explained by the estimated regression model.
- SS_{res} is the **residual sum of squares**. It measures the variation in the response data not explained by the estimated regression model.

- The **degrees of freedom** (df) add down.
  - The degrees of freedom for SS_{total} is the divisor in the sample variance
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\text{SS}_{total}}{n-1}.$$
  - The degrees of freedom for SS_{reg} is $p-1$, the number of independent variables in the model fit (recall $p = k+1 \Longrightarrow k = p-1$).
  - The degrees of freedom for SS_{res} is the divisor needed to create an unbiased estimator of σ^2. Recall that
$$\text{MS}_{res} = \frac{\text{SS}_{res}}{n-p}$$
is an unbiased estimator of σ^2.

- **Mean squares** (MS) are the sums of squares divided by their degrees of freedom.
- The F statistic is formed by taking the ratio of MS_{reg} and MS_{res}. More on this in a moment.

Cheese data: I used SAS to calculate the ANOVA table for the cheese data:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Regression	3	4994.508	1664.836	16.22	<.0001
Residual	26	2668.378	102.629		
Corrected Total	29	7662.886			

Remark: The reason I used SAS here is that R does something different in displaying the analysis of variance (it breaks down the regression sum of squares further). More on this in a moment. You can get R to produce this type of ANOVA table, but it takes extra work and is not worth it.

Overall F test: In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

the F statistic in the ANOVA table can be used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus

$$H_1 : \text{at least one of the } \beta_j\text{'s is nonzero.}$$

In other words, F tests if **at least one** of the independent variables $x_1, x_2, ..., x_k$ is important in describing the response Y in the population (H_0 says no; H_1 says yes). If H_0 is rejected, we do not know which one or how many of the β_j's are nonzero; only that at least one is.

Sampling distribution: When H_0 is true, both MS_{reg} and MS_{res} are unbiased estimators of σ^2. Therefore, when H_0 is true,

$$F = \frac{\text{MS}_{reg}}{\text{MS}_{res}} \approx 1.$$

The sampling distribution of F when H_0 is true is

$$F = \frac{\text{MS}_{reg}}{\text{MS}_{res}} \sim F(p-1, n-p).$$

Recall that the mean of an F distribution is around 1. Therefore,

- Values of F in the center of this distribution are consistent with H_0.
- Large values of F (i.e., out in the right tail) are consistent with H_1.
- Unusually small values of F (i.e., close to zero) might indicate there is a violation of our statistical assumptions or we have fit the incorrect model.

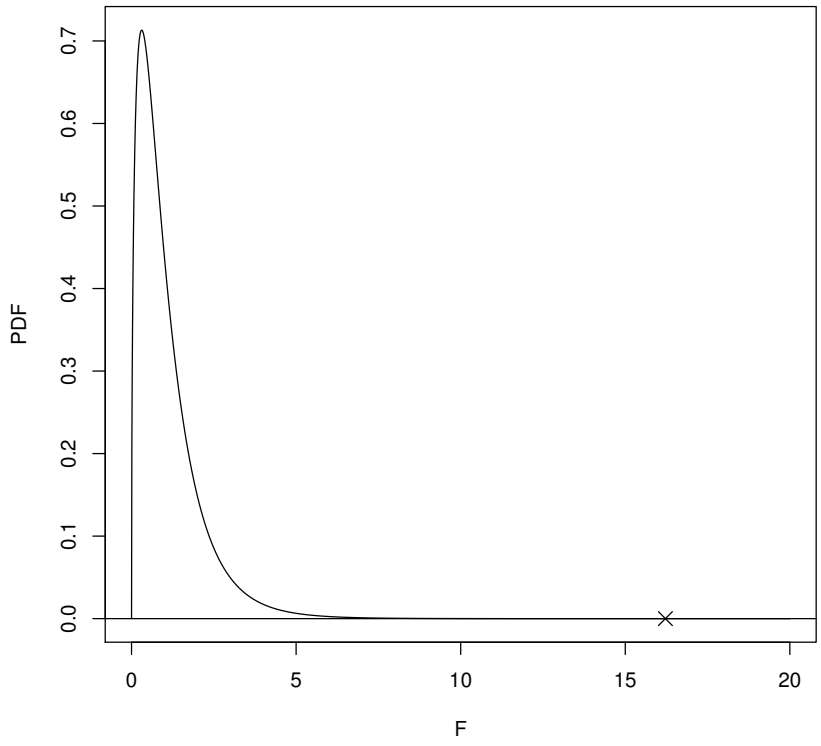

Figure 11.1: Cheese data: $F(3, 26)$ pdf. This is the sampling distribution of F when H_0 is true. An "$\times$" at $F = 16.22$ has been added.

Cheese data: For the cheese data in Example 11.1, the F statistic is used to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

versus

$$H_1 : \text{at least one of the } \beta_j \text{ is nonzero.}$$

Interpretation: Based on the F statistic ($F = 16.22$), and the corresponding probability value (p-value < 0.0001), we have strong evidence to reject H_0. See also Figure 11.1. We conclude that at least one of the independent variables (`ACETIC`, `H2S`, `LACTIC`) is important in describing `TASTE` in the population.

Remark: In the next section, we learn how to investigate the population-level effects of each variable separately.

Terminology: In the regression analysis of variance,

$$\text{SS}_{total} = \text{SS}_{reg} + \text{SS}_{res}.$$

Therefore, the proportion of the total variation in the response data explained by the estimated regression model is

$$R^2 = \frac{\text{SS}_{reg}}{\text{SS}_{total}}.$$

This statistic is called the **coefficient of determination**. Clearly,

$$0 \leq R^2 \leq 1.$$

In general, the larger the R^2, the better the estimated regression model explains the variability in the response data.

Cheese data: For the cheese data in Example 11.1, recall the ANOVA table presented earlier:

```
Analysis of Variance

                          Sum of       Mean
Source             DF    Squares     Square   F Value   Pr > F
Regression          3   4994.508   1664.836     16.22   <.0001
Residual           26   2668.378    102.629
Corrected Total    29   7662.886
```

Therefore, the coefficient of determination

$$R^2 = \frac{\text{SS}_{reg}}{\text{SS}_{total}} = \frac{4994.508}{7662.886} \approx 0.652.$$

Interpretation: About 65.2 percent of the variability in the `TASTE` data is explained by the linear regression model that includes `ACETIC`, `H2S`, and `LACTIC`. The remaining 34.8 percent of the variability in the taste data is explained by other sources.

Warning: It is important to understand what R^2 measures and what it does not. Its value is computed under the assumption that the regression model is **correct** and assesses how much of the variation in the response is attributed to that relationship.

- If R^2 is small, it may be that there is just a lot of random inherent variation in the data. Although the estimated regression model is reasonable, it can explain only so much of the overall variation.
- Alternatively, R^2 may be large (e.g., close to 1) but for an estimated model that is not appropriate for the data. A better model may exist.

Question: How does R display an analysis of variance table? For the cheese data in Example 11.1, R provides

```
> fit = lm(taste~acetic+h2s+lactic)
> anova(fit)
          Df  Sum Sq Mean Sq F value    Pr(>F)
acetic     1 2314.14 2314.14 22.5484 6.528e-05 ***
h2s        1 2147.11 2147.11 20.9209 0.0001035 ***
lactic     1  533.26  533.26  5.1959 0.0310870 *
Residuals 26 2668.38  102.63
```

Note: The convention used by R is to partition the regression sum of squares

$$\text{SS}_{reg} = 4994.508$$

into sums of squares for each of the three independent variables `ACETIC`, `H2S`, and `LACTIC`, as they are added to the model sequentially. These are called **sequential sums of squares**. Note that, after rounding,

$$\begin{aligned} \text{SS}_{reg} = 4994.51 &= 2314.14 + 2147.11 + 533.26 \\ &= \text{SS}(\texttt{ACETIC}) + \text{SS}(\texttt{H2S}) + \text{SS}(\texttt{LACTIC}). \end{aligned}$$

- SS(`ACETIC`) is the sum of squares added when compared to a model that includes only an intercept term.
- SS(`H2S`) is the sum of squares added when compared to a model that includes an intercept term and `ACETIC`.

- SS(LACTIC) is the sum of squares added when compared to a model that includes an intercept term, ACETIC, and H2S.

In other words, we can use the sequential sums of squares to assess the impact of adding independent variables ACETIC, H2S, and LACTIC to the model in sequence. The p-values provided by R help you assess the statistical significance of each independent variable as you add them. Small p-values suggest statistical significance.

Interesting: If you change the order of the independent variables in the lm function, then you will get a different sequential sum of squares partition. For example,

```
> fit.2 = lm(taste~h2s+lactic+acetic)
> anova(fit.2)
          Df Sum Sq Mean Sq F value    Pr(>F)
h2s        1 4376.8  4376.8 42.6468 6.356e-07 ***
lactic     1  617.1   617.1  6.0131   0.02123 *
acetic     1    0.6     0.6  0.0054   0.94193
Residuals 26 2668.4   102.6
```

This table suggests that ACETIC does not add significantly to a regression model that already includes H2S and LACTIC (p-value = 0.941). Note that the previous sequential sum of squares partition (on the previous page) does not enable us to see this.

11.5 Inference for individual regression parameters

Goal: In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we are interested in writing **confidence intervals** for individual regression parameters β_j.

- This can help us assess the importance of using the independent variable x_j in a model that includes the other independent variables.
- That is, inference regarding the population parameter β_j is always **conditional** on the other variables being included in the model.

Confidence intervals: Under our linear regression model assumptions, a $100(1-\alpha)$ percent confidence interval for β_j, for $j = 0, 1, 2, ..., k$, is given by

$$b_j \pm t_{n-p,\alpha/2}\sqrt{\text{MS}_{res}c_{jj}},$$

where b_j is the least squares estimate of β_j, MS_{res} is our estimate of the error variance σ^2, and $c_{jj} = (\mathbf{X}'\mathbf{X})^{-1}_{jj}$ is the corresponding diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix.

- The value $t_{n-p,\alpha/2}$ is the upper $\alpha/2$ quantile from the $t(n-p)$ distribution.
- Note the familiar form of the interval:

$$\underbrace{\text{point estimate}}_{b_j} \pm \underbrace{\text{quantile}}_{t_{n-p,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\text{MS}_{res}c_{jj}}}.$$

  We interpret the interval in the same way:

  > "We are $100(1-\alpha)$ percent confident that the population parameter β_j is in this interval."

- Of particular interest is the value $\beta_j = 0$:
  - If the confidence interval for β_j contains "0," this suggests (at the population level) that the independent variable x_j does not significantly add to a model that contains the other independent variables.
  - If the confidence interval for β_j does not contain "0," this suggests (at the population level) that the independent variable x_j does significantly add to a model that contains the other independent variables.

Cheese data: We can use the `confint` function in R to calculate confidence intervals for the population regression parameters:

```
> fit = lm(taste~acetic+h2s+lactic)
> confint(fit,level=0.95)
                 2.5 %    97.5 %
(Intercept) -69.443161 11.689630
acetic       -8.839009  9.495026
h2s           1.345693  6.477870
lactic        1.932318 37.407035
```

Interpretation: I will ignore the intercept confidence interval, which describes $E(Y)$ when $x_1 = x_2 = x_3 = 0$, a nonsensical quantity. Here is how you interpret the other confidence intervals:

- We are 95 percent confident that β_1 (the population parameter for `ACETIC`) is between -8.84 and 9.50.
  - This interval includes "0." Therefore, `ACETIC` does not significantly add to a model that includes `H2S` and `LACTIC`.
  - This reaffirms what we saw in the sequential SS when `ACETIC` was added last.
- We are 95 percent confident that β_2 (the population parameter for `H2S`) is between 1.35 and 6.48.
  - This interval does not include "0." Therefore, `H2S` does significantly add to a model that includes `ACETIC` and `LACTIC`.
- We are 95 percent confident that β_3 (the population parameter for `LACTIC`) is between 1.93 and 37.41.
  - This interval does not include "0." Therefore, `LACTIC` does significantly add to a model that includes `ACETIC` and `H2S`.

11.6 Confidence and prediction intervals for a given $\mathbf{x} = \mathbf{x}_0$

Goals: We would like to create $100(1-\alpha)$ percent intervals for the mean $E(Y|\mathbf{x}_0)$ and for the new value $Y^*(\mathbf{x}_0)$. As in simple linear regression, the former is called a **confidence interval** (because it is for a mean response) and the latter is called a **prediction interval** (because it is for a new random variable).

Cheese data: Suppose we are interested estimating $E(Y|\mathbf{x}_0)$ and predicting a new $Y^*(\mathbf{x}_0)$ when `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, so that

$$\mathbf{x}_0 = \begin{pmatrix} 5.5 \\ 6.0 \\ 1.4 \end{pmatrix}.$$

We use R to compute the following:

```
> predict(fit,data.frame(acetic=5.5,h2s=6.0,lactic=1.4),level=0.95,interval="confidence")
     fit      lwr     upr
23.93552 20.04506 27.82597
> predict(fit,data.frame(acetic=5.5,h2s=6.0,lactic=1.4),level=0.95,interval="prediction")
     fit      lwr      upr
23.93552 2.751379 45.11966
```

- Note that the point estimate/prediction is

$$\begin{aligned} \widehat{Y}(\mathbf{x}_0) &= b_0 + b_1 x_{10} + b_2 x_{20} + b_3 x_{30} \\ &= -28.877 + 0.328(5.5) + 3.912(6.0) + 19.670(1.4) \approx 23.936. \end{aligned}$$

- A 95 percent **confidence interval** for $E(Y|\mathbf{x}_0)$ is $(20.05, 27.83)$. When `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, we are 95 percent confident that the population mean taste rating is between 20.05 and 27.83.

- A 95 percent **prediction interval** for $Y^*(\mathbf{x}_0)$, when $\mathbf{x} = \mathbf{x}_0$, is $(2.75, 45.12)$. When `ACETIC` $= 5.5$, `H2S` $= 6.0$, and `LACTIC` $= 1.4$, we are 95 percent confident that the taste rating for a new specimen will be between 2.75 and 45.12.

11.7 Model diagnostics (residual analysis)

Importance: We now discuss diagnostic techniques for linear regression (simple and multiple). The term "diagnostics" refers to the process of "checking the model assumptions." This is an important exercise because if the model assumptions are violated, then our analysis and all subsequent interpretations could be compromised.

Recall: We first recall the model assumptions on the error terms in the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. Specifically, we have made the following assumptions:

- $E(\epsilon_i) = 0$, for $i = 1, 2, ..., n$
- $\text{var}(\epsilon_i) = \sigma^2$, for $i = 1, 2, ..., n$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are normally distributed.

Residuals: In checking our model assumptions, we first have to deal with the obvious problem; namely, the error terms ϵ_i in the model are never observed. However, after fitting the model, we can calculate the residuals

$$e_i = Y_i - \widehat{Y}_i,$$

where the ith fitted value

$$\widehat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}.$$

We can think of the residuals $e_1, e_2, ..., e_n$ as "proxies" for the error terms $\epsilon_1, \epsilon_2, ..., \epsilon_n$. Therefore, we can use the residuals to check our model assumptions instead.

Normality: To check the normality assumption for the errors in linear regression, we can examine the qq-plot of the residuals.

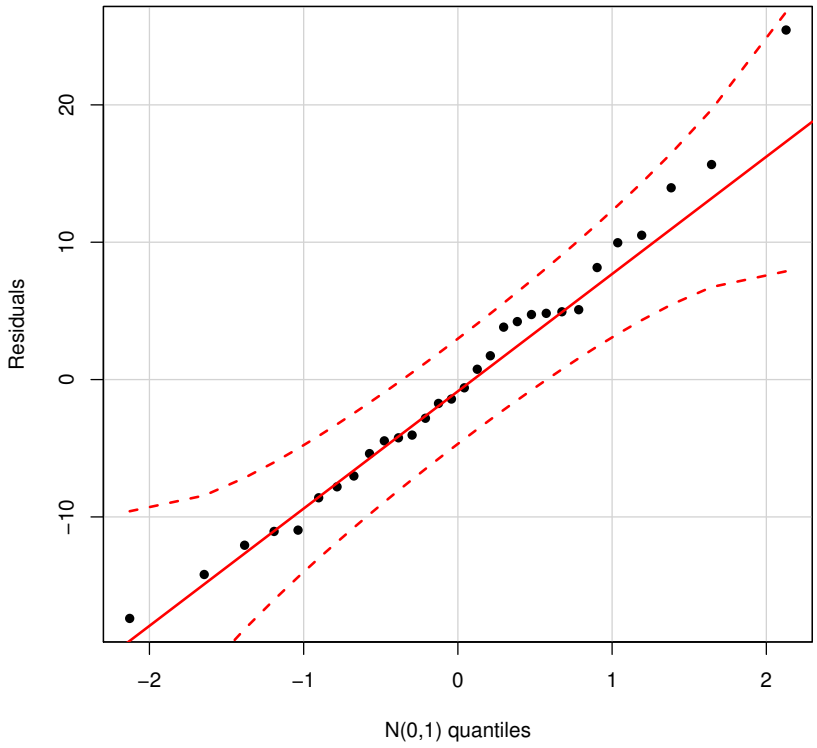

Figure 11.2: Cheese data. Normal qq-plot of the least squares residuals.

- Recall that if the plotted points follow a straight line (approximately), this supports the normality assumption.
- Substantial deviation from linearity is not consistent with the normality assumption.
- The plot in Figure 11.2 supports the normality assumption for the errors in the multiple linear regression model for the cheese data.

Importance: If the normality assumption is violated in a linear regression analysis, this could affect population level inferences for regression parameters β_j and confidence/prediction intervals. Mild departures are generally not a problem unless the sample size is very small. Substantial departures from normality should raise concern.

Terminology: A **residual plot** is a scatterplot of the residuals e_i (on the vertical axis) versus the predicted values $\widehat{Y}_i$ (on the horizontal axis). A residual plot can be very useful in detecting the following violations:

- misspecifying the true regression function
  - i.e., a violation of the $E(\epsilon_i) = 0$ assumption
- non-constant variance (heteroscedasticity)
  - i.e., a violation of the $\text{var}(\epsilon_i) = \sigma^2$ assumption
- correlated observations over time; i.e., a violation of the assumption that the ϵ_i's are independent random variables.

Important: Mathematical arguments show that if all of the linear regression model assumptions hold, then the residuals and fitted values are **independent**.

- Therefore, if the residual plot appears to be random in appearance with no noticeable patterns (i.e., the plot looks like a random scatter of points), this suggests there are no model inadequacies.
- On the other hand, if there are structural (non-random) patterns in the residual plot, this suggests that the model is inadequate in some way.
- Furthermore, the residual plot often reveals what type of model violation is occurring.

Cheese data: The residual plot in Figure 11.3 does not suggest any obvious model inadequacies. It looks completely random in appearance.

Note: We now look at two new regression examples. We use these examples to illustrate model violations that are commonly seen in practice. We also discuss remedies to handle these violations.

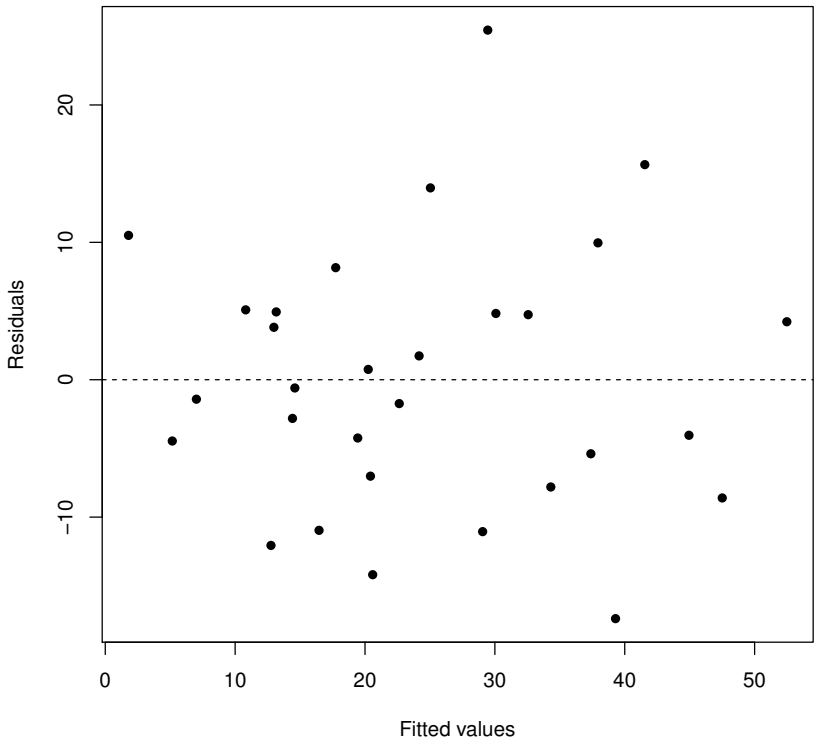

Figure 11.3: Cheese data. Residual plot for the multiple linear regression model fit. A horizontal line at zero has been added.

Example 11.2. An electric company is interested in describing the relationship between the following two variables:

$$\begin{aligned} Y &= \text{peak hour electricity demand (measured in kWh)} \\ x &= \text{total monthly energy usage (measured in kWh).} \end{aligned}$$

This is important for planning purposes because the generating system must be large enough to meet the maximum demand imposed by customers. Engineers consider the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

to describe the relationship. A random sample of $n = 53$ residual customers is obtained to estimate the model; see Figure 11.4.

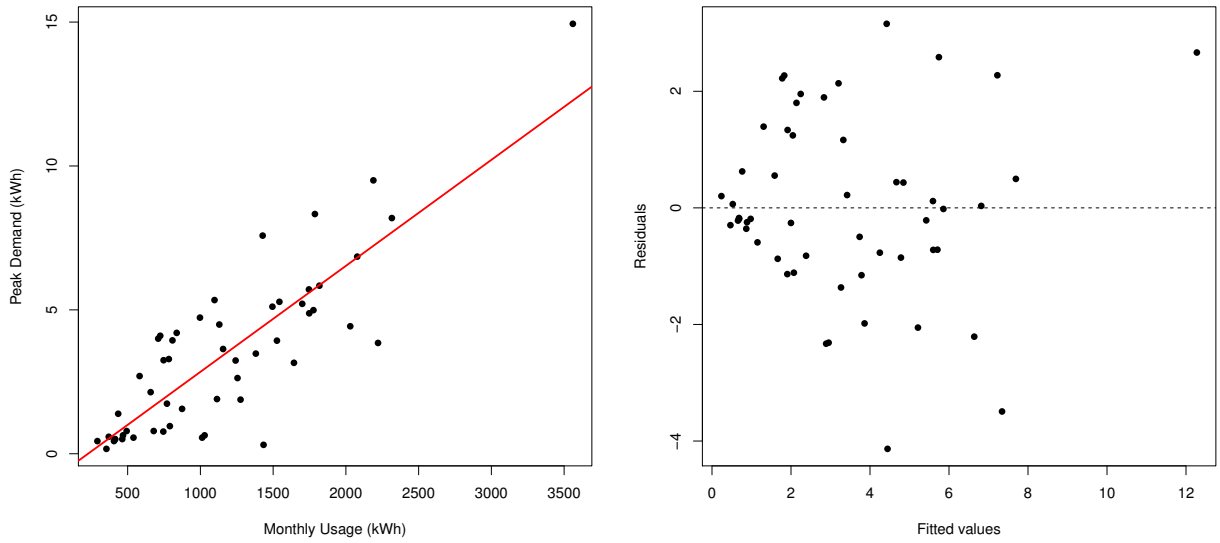

Figure 11.4: Electricity data. Left: Scatterplot of peak demand (Y, measured in kWh) versus monthly usage (x, measured in kWh) with estimated simple linear regression line superimposed. Right: Residual plot for the simple linear regression model fit.

Problem: There is a clear problem with **non-constant variance** here. Note how the residual plot "fans out" like the bell of a trumpet. This violation may have been missed by looking at the scatterplot alone, but the residual plot highlights it.

Remedy: When faced with a non-constant variance violation in regression, a common remedy is to **transform** the response variable Y. Common transformations are the logarithmic ($\ln Y$) and square-root ($\sqrt{Y}$) transformations.

- A more advanced remedy is to use a model fitting technique known as **weighted least squares**; this involves weighting certain observations more/less depending on their level of variability. We will not pursue this.

- The advantage of using a transformation is that you can still use least squares without weighting. However, all inferences will pertain to the population model with the **transformed response**, not the response Y itself. This can sometimes complicate how the results are interpreted.

Analysis: We apply a square-root transformation $W = \sqrt{Y}$ and consider the model

$$W = \beta_0 + \beta_1 x + \epsilon,$$

where

$$\begin{aligned} W &= \text{peak hour electricity demand (measured in } \sqrt{\text{kWh}}) \\ x &= \text{total monthly energy usage (measured in kWh).} \end{aligned}$$

Fitting this model in R gives the least squares estimates

```
> fit.2 = lm(sqrt(peak.demand)~monthly.usage)
> fit.2
Coefficients:
  (Intercept)  monthly.usage
    0.5808309      0.0009529
```

Therefore, the estimated model on the transformed scale is

$$\widehat{W} = 0.581 + 0.000953x,$$

or, in other words,

$$\widehat{\sqrt{\texttt{Peak demand}}} = 0.581 + 0.000953 \ \texttt{Monthly usage}.$$

Discussion: First note that applying the transformation did help to reduce the non-constant variance problem considerably; see Figure 11.5. The noticeable "fanning out" shape that we saw in the residual plot previously (i.e., based on the untransformed response Y) is now largely absent. Let's proceed with inference for β_1 to determine if the linear relationship is significant for the population:

```
> confint(fit.2,level=0.95)
                     2.5 %      97.5 %
(Intercept)   0.3208043932 0.840857384
monthly.usage 0.0007563267 0.001149532
```

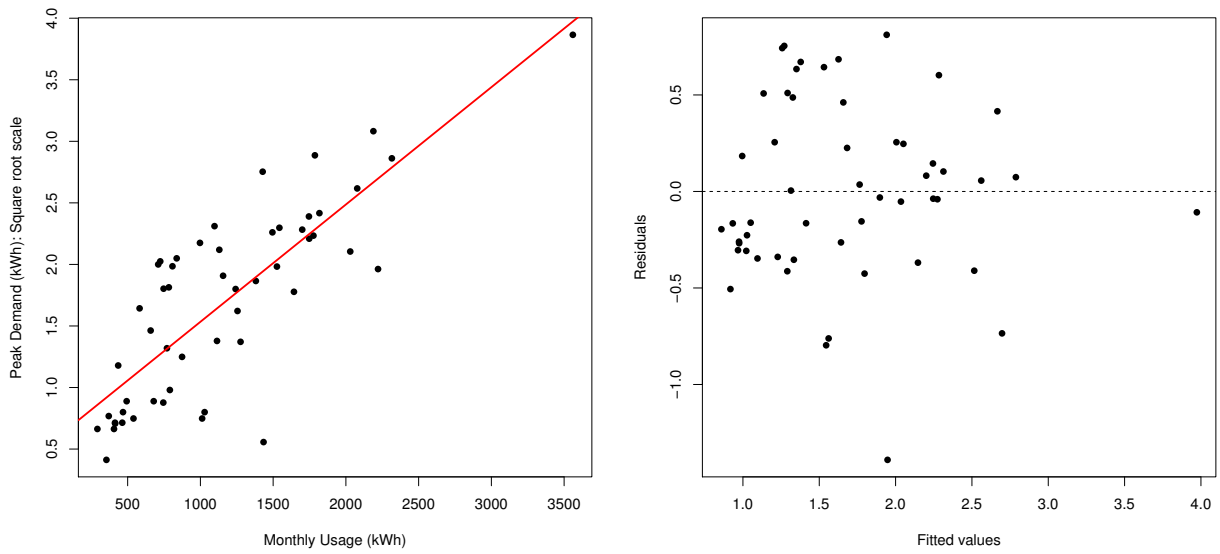

Figure 11.5: Electricity data. Left: Scatterplot of the square root of peak demand ($\sqrt{Y}$) versus monthly usage (x, measured in kWh) with estimated simple linear regression line superimposed. Right: Residual plot for the simple linear regression model fit with transformed response.

Interpretation: We are 95 percent confident that the population regression parameter β_1 (in the transformed model) is between 0.000756 and 0.001150.

- Note that this interval does not include "0" and includes only positive values. This suggests that peak demand (on the square root scale) and monthly usage are positively related in the population.
- Specifically, for every one-unit increase in x (monthly usage measured in kWh), we are 95 percent confident that the mean peak demand will increase between 0.000756 and 0.001150 $\sqrt{\text{kWh}}$.
- I examined the qq plot for normality (using the residuals from the transformed model fit). This plot (not shown) did reveal some potential mild departures on the high side, but nothing that was serious.

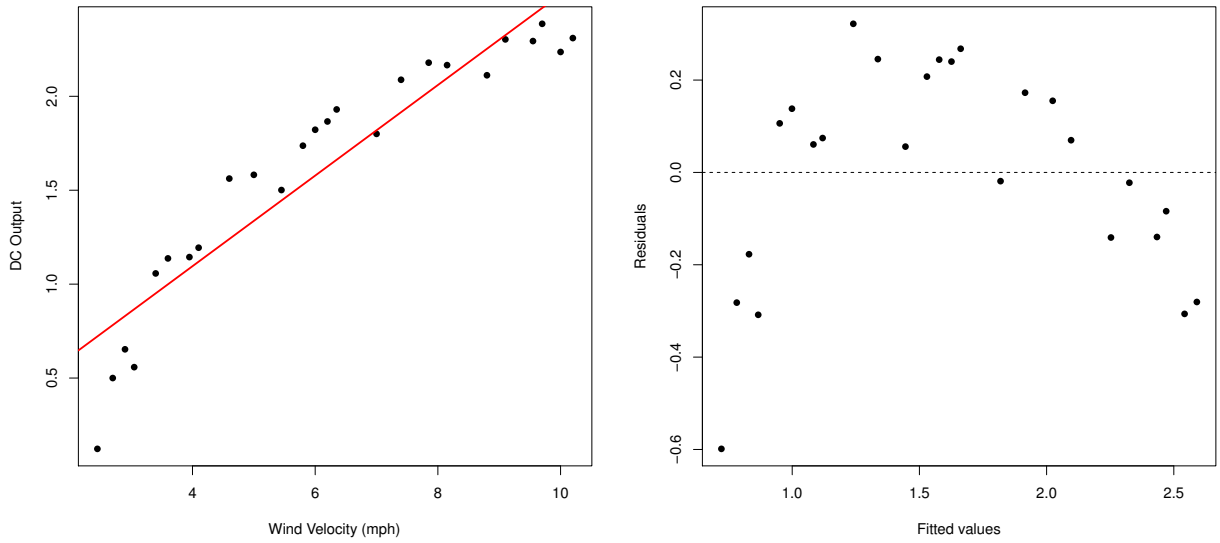

Figure 11.6: Windmill data. Left: Scatterplot of DC output Y versus wind velocity (x, measured in mph) with least squares simple linear regression line superimposed. Right: Residual plot for the simple linear regression model fit.

Example 11.3. An engineer is investigating the use of a windmill to generate electricity. He has collected data on

$$\begin{aligned} Y &= \text{direct current (DC) output} \\ x &= \text{wind velocity (measured in mph).} \end{aligned}$$

Data for $n = 25$ observation pairs are shown in Figure 11.6. The engineer initially assumes a simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

to describe the relationship and fits this model.

Problem: There is a clear **quadratic relationship** between DC output and wind velocity. The residual plot in Figure 11.6 from the simple linear regression model fit shows a pronounced quadratic pattern. It is easy to see why this is happening—a simple linear regression model is inappropriate here (it does not explain quadratic relationships).

Remark: I used R to calculate the coefficient of determination from the simple linear regression model fit. It is

$$R^2 \approx 0.875.$$

A novice data analyst (especially one that doesn't even bother to graph the data) might think that because this is "pretty large," the model we have fit is a "good model." However, it is easy to see from Figure 11.6 that a simple linear regression model is not a good model for the data. Even though 0.875 is in fact "pretty large," its value refers specifically to a model that is inappropriate.

Remedy: Fit a multiple linear regression model with two independent variables: wind velocity x and its square x^2. The model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

is called a **quadratic regression model**. It is straightforward to fit a quadratic regression model in R. We simply regress Y on both x and x^2.

```
> wind.velocity.sq = wind.velocity^2
> fit.2 = lm(DC.output~wind.velocity+wind.velocity.sq)
> fit.2
Coefficients:
     (Intercept)     wind.velocity  wind.velocity.sq
        -1.15590           0.72294          -0.03812
```

The estimated quadratic regression model is

$$\widehat{Y} = -1.15590 + 0.72294x - 0.03812x^2$$

or, in other words,

$$\widehat{\texttt{DC output}} = -1.15590 + 0.72294\ \texttt{Wind.velocity} - 0.03812\ (\texttt{Wind.velocity})^2.$$

Analysis: Note that the residual plot from the **quadratic model** fit, shown in Figure 11.7, now looks much more random. The quadratic trend has disappeared (because the model now incorporates it).

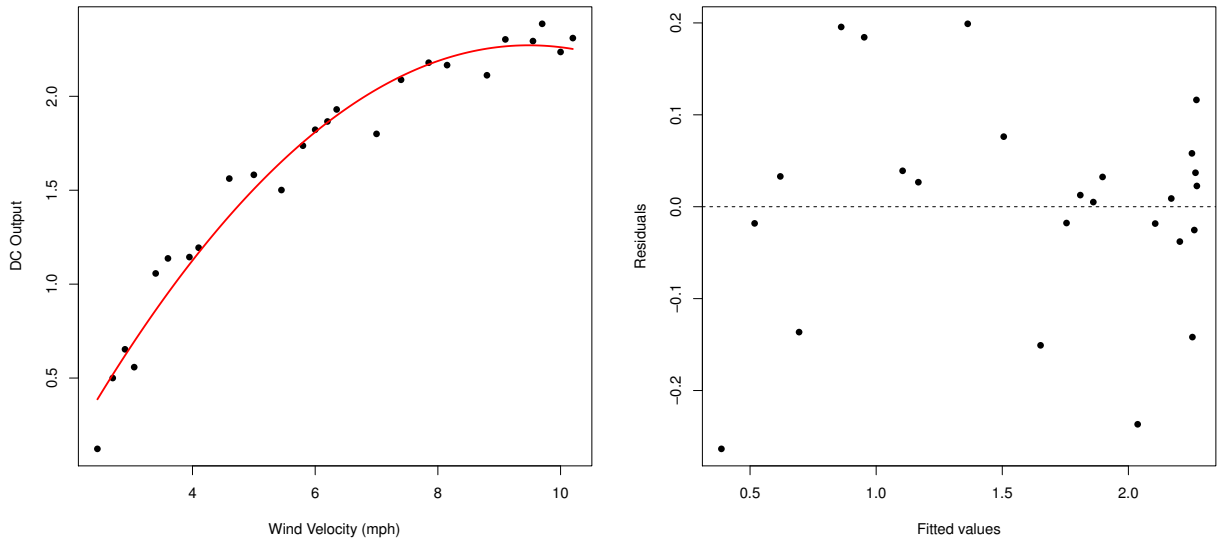

Figure 11.7: Windmill data. Scatterplot of DC output Y versus wind velocity (x, measured in mph) with least squares quadratic regression curve superimposed. Right: Residual plot for the quadratic regression model fit.

Confidence interval: To see if the quadratic effect between DC output and wind velocity is significant, we can write a confidence interval for β_2, the population parameter in the quadratic regression model that describes the quadratic effect.

```
> confint(fit.2,level=0.95)
                          2.5 %      97.5 %
(Intercept)       -1.51810023 -0.79369625
wind.velocity      0.59554751  0.85032429
wind.velocity.sq -0.04806859 -0.02817318
```

Interpretation: We are 95 percent confident that the population regression parameter β_2 (in the quadratic model) is between -0.0481 and -0.0282. Note that this interval does not include "0" and includes only negative values. This suggests that quadratic effect between DC output and wind velocity is significant in the population.

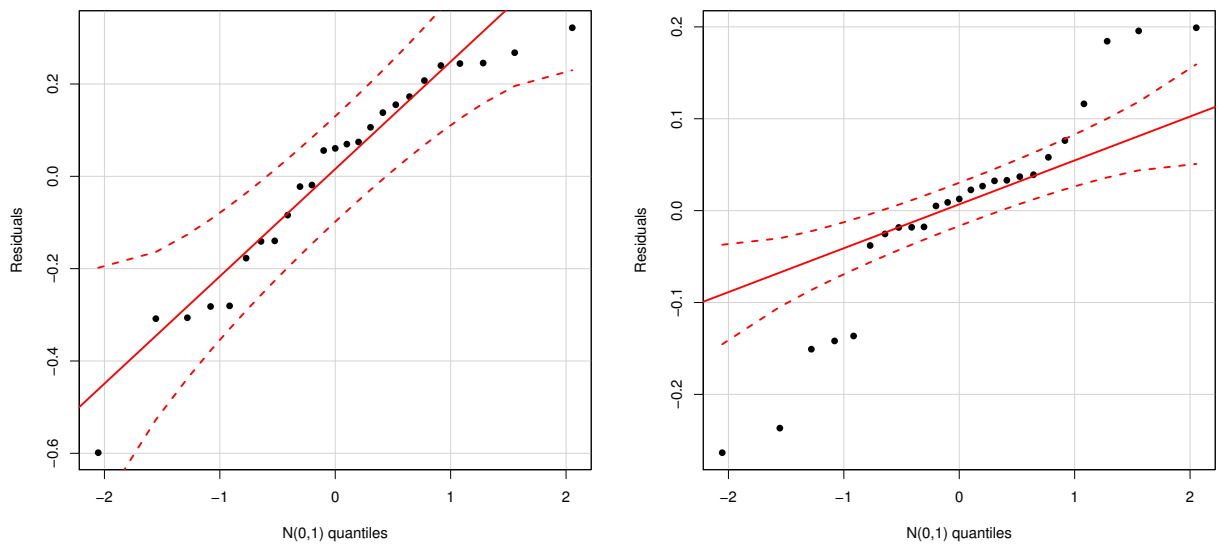

Figure 11.8: Windmill data. QQ plots of the residuals. Left: Simple linear regression fit. Right: Quadratic regression fit.

Remark: I used R to calculate the coefficient of determination from the quadratic regression model fit. It is

$$R^2 \approx 0.968.$$

This means that about 96.8 percent of the variability in the DC output data is explained by the estimated model that includes both wind velocity and (wind velocity)^2. The remaining 3.2 percent is not explained by the estimated model. This is an improvement over the largely meaningless $R^2 \approx 0.875$ calculated from the simple linear regression.

New Problem: Figure 11.8 shows the normal qq plots for the simple linear regression model fit (left) and the quadratic model fit (right).

- I say "new" problem, because now it looks like the normality assumption (for the quadratic model) is violated. Interestingly, this was not a problem with the simple linear regression model.
- It appears that fitting the quadratic regression model fixed one problem (i.e., selecting a better regression function) but created another (normality violation).

12 Factorial Experiments

12.1 Introduction

Importance: In engineering experiments, particularly those carried out in manufacturing settings, there are often several variables of interest and the goal is to understand the effects of these variables on a continuous response Y (e.g., yield, lifetime, fill weights, etc.). A **factorial treatment structure** is an efficient way of defining treatments in these types of experiments.

- One example of a factorial treatment structure uses k **factors**, where each factor has two **levels**. This is called a 2^k factorial experiment.
- Factorial experiments are common in the early stages of experimental work. For this reason, they are also called **factor screening experiments**.

Example 12.1. A nickel-titanium alloy is used to make components for jet turbine aircraft engines. Cracking is a potentially serious problem in the final part, as it can lead to nonrecoverable failure. A test is run at the parts producer to determine the effect of $k = 4$ factors on cracks: pouring temperature (A), titanium content (B), heat treatment method (C), and amount of grain refiner used (D).

- Factor A has 2 levels: "low" temperature and "high" temperature
- Factor B has 2 levels: "low" content and "high" content
- Factor C has 2 levels: Method 1 and Method 2
- Factor D has 2 levels: "low" amount and "high" amount.

The **response variable** in the experiment is

$$Y = \text{length of largest crack (in mm) induced in a piece of sample material.}$$

Note: In this example, there are 4 factors, each with 2 levels. Thus, there are

$$2 \times 2 \times 2 \times 2 = 2^4 = 16$$

different **treatment combinations**. These are listed here:

$$\begin{array}{cccc} a_1b_1c_1d_1 & a_1b_2c_1d_1 & a_2b_1c_1d_1 & a_2b_2c_1d_1 \\ a_1b_1c_1d_2 & a_1b_2c_1d_2 & a_2b_1c_1d_2 & a_2b_2c_1d_2 \\ a_1b_1c_2d_1 & a_1b_2c_2d_1 & a_2b_1c_2d_1 & a_2b_2c_2d_1 \\ a_1b_1c_2d_2 & a_1b_2c_2d_2 & a_2b_1c_2d_2 & a_2b_2c_2d_2 \end{array}$$

For example, the treatment combination $a_1b_1c_1d_1$ holds each factor at its "low" level, the treatment combination $a_1b_1c_2d_2$ holds Factors A and B at their "low" level and Factors C and D at their "high" level, and so on.

Terminology: In a 2^k factorial experiment, one **replicate** of the experiment uses 2^k runs, one at each of the 2^k treatment combinations.

- Therefore, in Example 12.1, one replicate of the experiment would require 16 runs (one at each treatment combination listed above).
- Two replicates would require 32 runs, three replicates would require 48 runs, and so on.

Terminology: There are different types of effects of interest in factorial experiments: **main effects** and **interaction effects**. For example, in a 2^4 factorial experiment,

- there is **1** "effect" that does not depend on any of the factors
- there are **4** main effects: A, B, C, and D
- there are **6** two-way interaction effects: AB, AC, AD, BC, BD, and CD
- there are **4** three-way interaction effects: ABC, ABD, ACD, and BCD
- there is **1** four-way interaction effect: ABCD.

Observation: Note that $1 + 4 + 6 + 4 + 1 = 16$. In other words, with 16 observations (from one 2^4 replicate), we can estimate the 4 main effects and we can estimate all of the 11 interaction effects. We will have 1 observation left to estimate the overall mean of Y, that is, the "effect" that depends on none of the 4 factors.

Generalization: In a 2^k factorial experiment, there is/are

- $\binom{k}{0} = 1$ overall mean (the mean of Y ignoring all factors)
- $\binom{k}{1} = k$ main effects
- $\binom{k}{2} = \frac{k(k-1)}{2}$ two-way interaction effects
- $\binom{k}{3}$ three-way interaction effects, and so on.

Note that

$$\binom{k}{0} + \binom{k}{1} + \binom{k}{2} + \cdots + \binom{k}{k} = \sum_{j=0}^{k} \binom{k}{j} = 2^k$$

and additionally that $\binom{k}{0}, \binom{k}{1}, ..., \binom{k}{k}$ are the entries in the $(k+1)$st row of Pascal's Triangle. Observe also that 2^k grows quickly in size as k increases. For example, if there are $k = 10$ factors (A, B, C, D, E, F, G, H, I, and J, say), then performing just one replicate of the experiment would require $2^{10} = 1024$ runs! In real life, rarely would this type of experiment be possible.

12.2 Example: A 2^2 experiment with replication

Remark: We first consider 2^k factorial experiments where $k = 2$, that is, there are only two factors, denoted by A and B. This is called a 2^2 experiment. We illustrate with an agricultural example.

Example 12.2. Predicting corn yield prior to harvest is useful for making feed supply and marketing decisions. Corn must have an adequate amount of nitrogen (Factor A) and phosphorus (Factor B) for profitable production and also for environmental concerns.

Table 12.1: Corn yield data (bushels/plot).

Treatment combination	Yield (Y)	Treatment sample mean
a_1b_1	35, 26, 25, 33, 31	30
a_1b_2	39, 33, 41, 31, 36	36
a_2b_1	37, 27, 35, 27, 34	32
a_2b_2	49, 39, 39, 47, 46	44

Experimental design: In a $2 \times 2 = 2^2$ factorial experiment, two levels of nitrogen ($a_1 = 10$ and $a_2 = 15$) and two levels of phosphorus were used ($b_1 = 2$ and $b_2 = 4$). Applications of nitrogen and phosphorus were measured in pounds per plot. Twenty small (quarter acre) plots were available for experimentation, and the four **treatment combinations** a_1b_1, a_1b_2, a_2b_1, and a_2b_2 were randomly assigned to plots. Note that there are 5 **replications**.

Response: The response variable is

$$Y = \text{ yield per plot (measured in \# bushels).}$$

Side-by-side boxplots of the data are shown in Figure 12.1.

Naive analysis: One silly way to analyze these data would be to simply regard each of the combinations a_1b_1, a_1b_2, a_2b_1, and a_2b_2 as a "treatment" and perform a one-way ANOVA with $t = 4$ treatment groups like we did in Chapter 9. This would produce the following ANOVA table:

```
> anova(lm(yield~treatment))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
treatment  3    575  191.67  9.5833 0.0007362 ***
Residuals 16    320   20.00
```

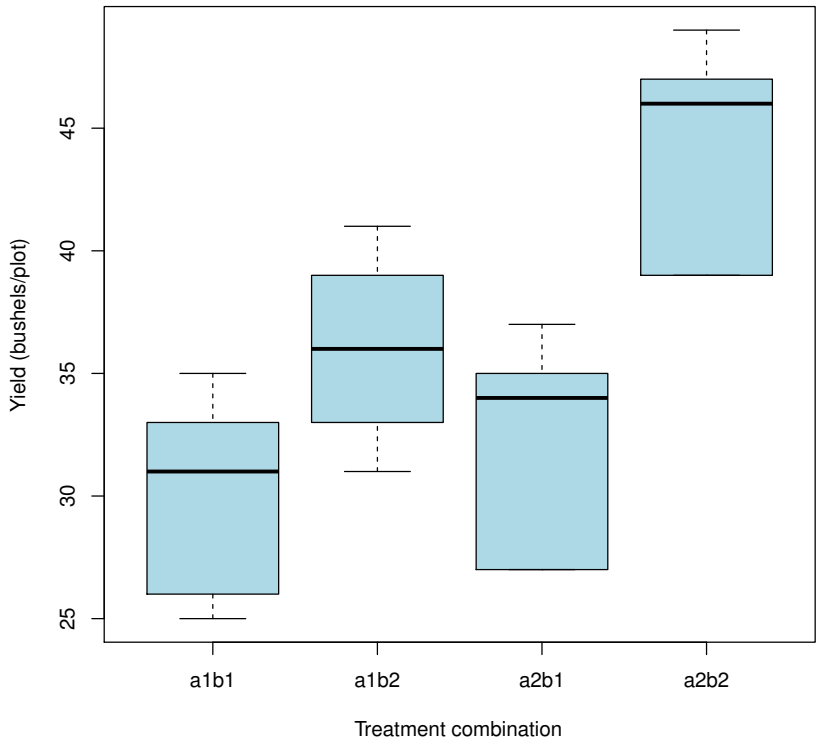

Figure 12.1: Boxplots of corn yields (bushels/plot) for four treatment groups.

Uninteresting conclusion: The value $F = 9.5833$ is not what we would expect from an $F(3, 16)$ distribution, the distribution of F when

$$H_0 : \mu_{11} = \mu_{12} = \mu_{21} = \mu_{22}$$

is true (p-value ≈ 0.0007). Therefore, we conclude that at least one of the factorial treatment population means is different.

Remark: As we have discussed before in one-way analyses, the overall F test provides very little information. However, with a factorial treatment structure, it is possible to explore the data further. We can target the (main) effects due to nitrogen (Factor A) and due to phosphorus (Factor B) individually. We can also determine if the two factors nitrogen and phosphorus interact.

Partition: Let us first recall the treatment sum of squares from the one-way ANOVA:

$$\text{SS}_{trt} = 575.$$

The way we learn more about specific effects is to partition SS_{trt} into the following pieces: SS_A, SS_B, and SS_{AB}. By "partition," I mean that we will write

$$\text{SS}_{trt} = \text{SS}_A + \text{SS}_B + \text{SS}_{AB}.$$

In words,

- SS_A is the sum of squares due to the main effect of A (nitrogen)
- SS_B is the sum of squares due to the main effect of B (phosphorus)
- SS_{AB} is the sum of squares due to the interaction effect of A and B (nitrogen and phosphorus).

Two-way analysis: We can use R to write this partition in a richer ANOVA table:

```
> fit = lm(yield~nitrogen+phosphorus+nitrogen*phosphorus)
> anova(fit)
Analysis of Variance Table
                    Df Sum Sq Mean Sq F value    Pr(>F)
nitrogen             1    125     125    6.25 0.0236742 *
phosphorus           1    405     405   20.25 0.0003635 ***
nitrogen:phosphorus  1     45      45    2.25 0.1530877
Residuals           16    320      20
```

Interpretation: The F statistics, say F_A, F_B, and F_{AB}, can be used to determine if the respective effects are significant in the population. Small p-values (e.g., p-value < 0.05) indicate that the effect is significant. Effects with large p-values can be treated as not significant.

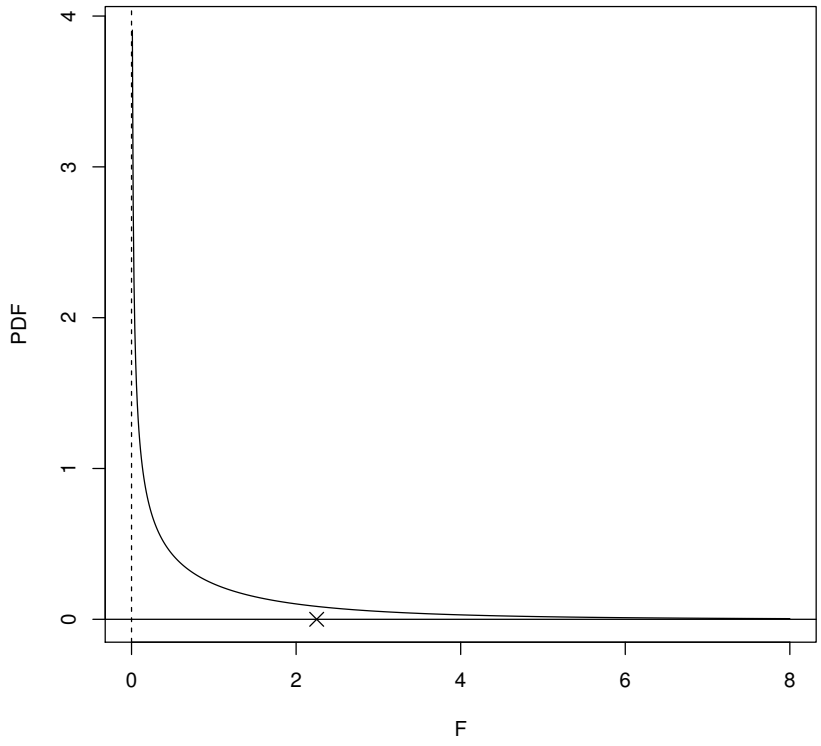

Figure 12.2: Corn yield data. $F(1, 16)$ pdf. An "$\times$" at $F_{AB} = 2.25$ has been added.

Analysis: When analyzing data from a 2^2 factorial experiment, the first task is to determine if the interaction effect is significant in the population. For the corn yield data in Example 12.2, we see that

$$F_{AB} = 2.25 \text{ (p-value} \approx 0.153).$$

- This value of F_{AB} is not all that unreasonable when compared to the $F(1, 16)$ distribution, the sampling distribution of F_{AB} when nitrogen and phosphorus do not interact (i.e., when the population-level interaction effect is zero).
- In other words, we do not have substantial (population-level) evidence that nitrogen and phosphorus interact.

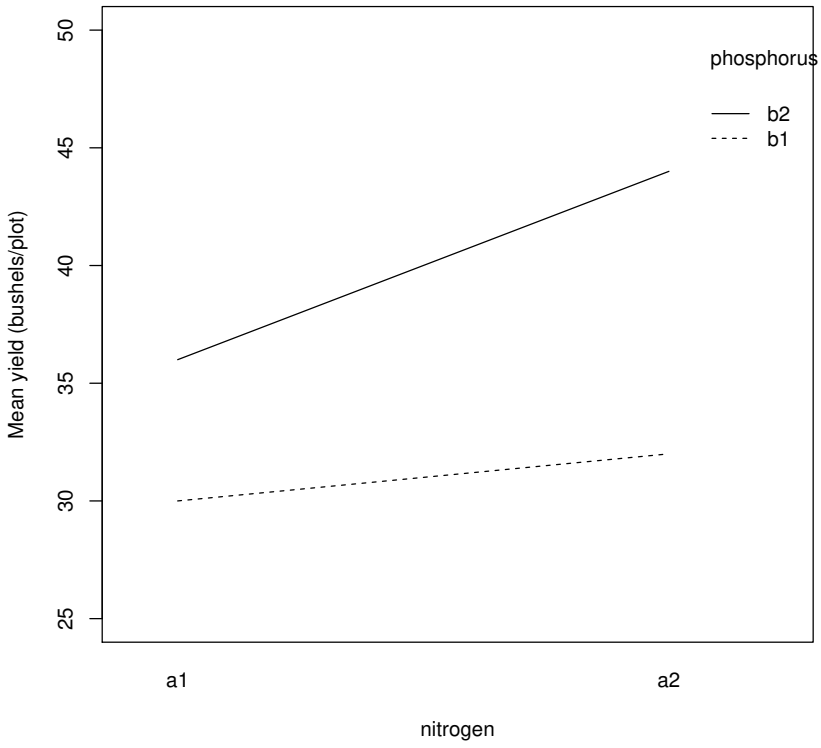

Figure 12.3: Interaction plot for nitrogen and phosphorus in Example 12.2.

Note: An **interaction plot** is a graphical display that can help us assess (visually) whether two factors interact. In this plot, the levels of Factor A are marked on the horizontal axis. The sample means of the treatments are plotted against the levels of A, and the points corresponding to the same level of Factor B are joined by straight lines.

- If Factors A and B do not interact **in the population**, the interaction plot should display parallel lines.
  - That is, the effect of one factor on the response Y stays constant across the levels of the other factor. This is essentially what it means to have no interaction.
- If the interaction plot displays a significant departure from parallelism (including an overwhelming case where the lines cross), then this is visual evidence of interaction.

  - That is, the effect of one factor on the response Y depends on the levels of the other factor.

- The F test that uses F_{AB} provides numerical evidence of interaction. The interaction plot provides visual evidence.

Conclusion: We have already used the interaction test statistic F_{AB} to conclude that the interaction effect of nitrogen and phosphorus is not significant in the population. Although the interaction plot in Figure 12.3 is not parallel (remember, it is constructed from the sample data), the departure from parallelism is not statistically significant.

Strategy for analyzing 2^2 factorial experiments:

1. Start by looking at whether the interaction effect is significant. This can be done by using an interaction plot and an F test that uses F_{AB}.

2. **If the interaction is significant**, then formal analysis of main effects is not all that meaningful because their interpretations depend on the interaction.

- In this situation, the best approach is to just ignore the factorial treatment structure and redo the entire analysis as a one-way ANOVA with four treatments.

- Tukey pairwise confidence intervals can help you formulate an ordering among the 4 treatment population means.

3. **If the interaction is not significant**, I prefer to re-estimate the model without the interaction term and then examine the main effects. This can be done numerically by examining the sizes of F_A and F_B, respectively.

- Confidence intervals for differences $\mu_{A1} - \mu_{A2}$ and $\mu_{B1} - \mu_{B2}$ can be used to quantify the size of these effects in the population.

4. Check model assumptions!

Corn yield data: Because the nitrogen/phosphorus interaction is not significant, I redid the ANOVA leaving out the interaction term:

```
> fit.2 = lm(yield~nitrogen+phosphorus)
> anova(fit.2)
Analysis of Variance Table
           Df Sum Sq Mean Sq F value   Pr(>F)
nitrogen    1    125  125.00  5.8219 0.027403 *
phosphorus  1    405  405.00 18.8630 0.000442 ***
Residuals  17    365   21.47
```

Comparing this to the ANOVA table that includes interaction (pp 202). Note that the interaction sum of squares $\text{SS}_{AB} = 45$ from that ANOVA table has now been "absorbed" into the residual sum of squares in the no-interaction analysis. Furthermore,

- the main effect of nitrogen (Factor A) is significant in describing yield in the population ($F_A = 5.82$, p-value $= 0.027$).
- the main effect of phosphorus (Factor B) is significant in describing yield in the population ($F_B = 18.86$, p-value $= 0.0004$).

Confidence intervals: A 95 percent confidence interval for $\mu_{A1} - \mu_{A2}$, the difference in the population means for the two levels of nitrogen (Factor A) is

$$(\overline{Y}_{A1} - \overline{Y}_{A2}) \pm t_{17,0.025}\sqrt{\text{MS}_{res}\left(\frac{1}{10} + \frac{1}{10}\right)}.$$

A 95 percent confidence interval for $\mu_{B1} - \mu_{B2}$, the difference in population means for the two levels of phosphorus (Factor B) is

$$(\overline{Y}_{B1} - \overline{Y}_{B2}) \pm t_{17,0.025}\sqrt{\text{MS}_{res}\left(\frac{1}{10} + \frac{1}{10}\right)}.$$

The R code online can be used to calculate these intervals:

$$\text{95\% CI for } \mu_{A1} - \mu_{A2}: \quad (-9.37, -0.62) \text{ bushels/acre}$$

$$\text{95\% CI for } \mu_{B1} - \mu_{B2}: \quad (-13.37, -4.63) \text{ bushels/acre}$$

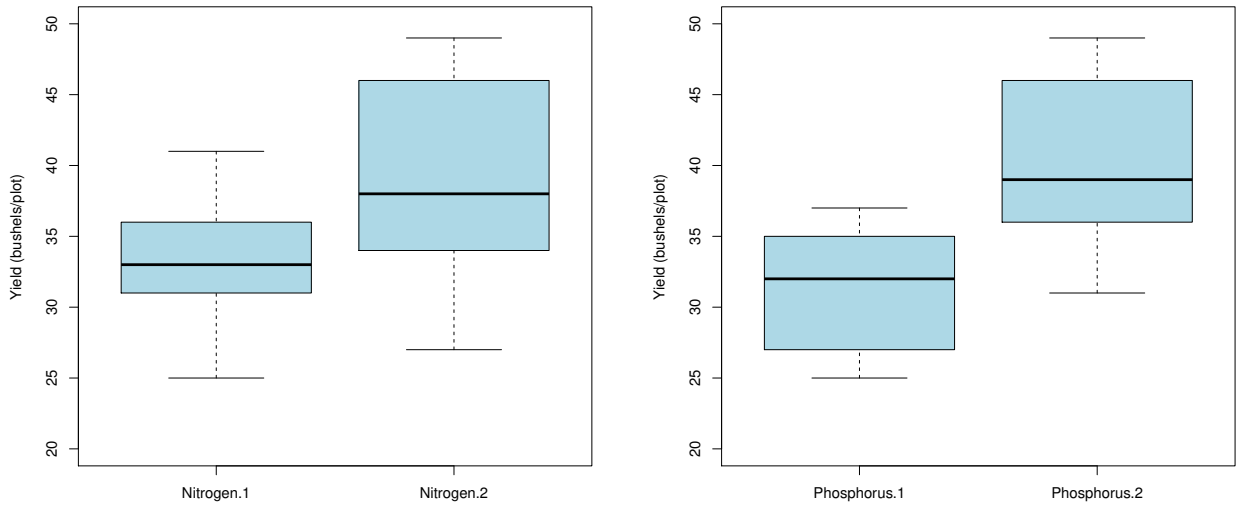

Figure 12.4: Left: Side by side boxplots of corn yields for nitrogen (Factor A). Right: Side by side boxplots of corn yields for phosphorus (Factor B).

Interpretation:

- We are 95 percent confident that the difference in the population mean yields (for low nitrogen/high nitrogen) is between -9.37 and -0.62 bushels per acre.
  - Note that this interval does not include "0," and includes only negative values.
  - This suggests that the population mean yield at the high level of nitrogen is larger than the population mean yield at the low level of nitrogen.
- We are 95 percent confident that the difference in the population mean yields (for low phosphorus/high phosphorus) is between -13.37 and -4.63 bushels per acre.
  - Note that this interval does not include "0," and includes only negative values.
  - This suggests that the population mean yield at the high level of phosphorus is larger than the population mean yield at the low level of phosphorus.
- Normal qq plots (not shown) for the data look fine.

12.3 Example: A 2^4 experiment without replication

Example 12.3. A chemical product is produced in a pressure vessel. A factorial experiment is carried out to study the factors thought to influence the filtration rate of this product. The four factors are temperature (A), pressure (B), concentration of formaldehyde (C) and stirring rate (D). Each factor is present at two levels (e.g., "low" and "high"). A 2^4 experiment is performed with one replication; the data are shown below.

	Factor					Filtration rate
Run	A	B	C	D	Run label	(Y, gal/hr)
1	−	−	−	−	$a_1b_1c_1d_1$	45
2	+	−	−	−	$a_2b_1c_1d_1$	71
3	−	+	−	−	$a_1b_2c_1d_1$	48
4	+	+	−	−	$a_2b_2c_1d_1$	65
5	−	−	+	−	$a_1b_1c_2d_1$	68
6	+	−	+	−	$a_2b_1c_2d_1$	60
7	−	+	+	−	$a_1b_2c_2d_1$	80
8	+	+	+	−	$a_2b_2c_2d_1$	65
9	−	−	−	+	$a_1b_1c_1d_2$	43
10	+	−	−	+	$a_2b_1c_1d_2$	100
11	−	+	−	+	$a_1b_2c_1d_2$	45
12	+	+	−	+	$a_2b_2c_1d_2$	104
13	−	−	+	+	$a_1b_1c_2d_2$	75
14	+	−	+	+	$a_2b_1c_2d_2$	86
15	−	+	+	+	$a_1b_2c_2d_2$	70
16	+	+	+	+	$a_2b_2c_2d_2$	96

Notation: When discussing factorial experiments, it is common to use the symbol "−" to denote the low level of a factor and the symbol "+" to denote the high level. For example, the first row of the table above indicates that each factor (A, B, C, and D) is run at its "low" level. The response Y for this run is 45 gal/hr.

Note: In this experiment, there are $k = 4$ factors, so there are 16 effects to estimate:

- the 1 overall mean
- the 4 main effects: A, B, C, and D
- the 6 two-way interactions: AB, AC, AD, BC, BD, and CD
- the 4 three-way interactions: ABC, ABD, ACD, BCD
- the 1 four-way interaction: ABCD.

In this 2^4 experiment, we have 16 values of Y and 16 effects to estimate. This leaves us with no observations (and therefore no degrees of freedom) to perform statistical tests. This is an obvious problem! We have no way to judge which main effects are significant, and we cannot learn about how these factors interact.

Terminology: A single replicate of a 2^k factorial experiment is called an **unreplicated factorial**. With only one replicate, there is no internal "error estimate," so we cannot perform statistical tests to judge significance. What do we do?

- One approach is to assume that certain higher-order interactions are "negligible" and then combine their sum of squares to estimate the error.
- This is an appeal to the **sparsity of effects principle**, which states that most systems are dominated by some of the main effects and low-order interactions and that most high-order interactions are negligible.
- To learn about which effects may be negligible, we can fit the full ANOVA model and obtain the sum of squares (SS) for each effect (see next page).
- Effects with "large" SS can be retained. Effects with "small" SS can be discarded. A smaller model with only the "large" effects can then be fit. This smaller model will have an error estimate formed by taking all of the effects with "small" SS and combining them together.

Analysis: Here is the R output from fitting the full model:

```
> fit = lm(filtration~A*B*C*D)
> anova(fit)
Analysis of Variance Table
          Df  Sum Sq Mean Sq F value Pr(>F)
A          1 1870.56 1870.56
B          1   39.06   39.06
C          1  390.06  390.06
D          1  855.56  855.56
A:B        1    0.06    0.06
A:C        1 1314.06 1314.06
B:C        1   22.56   22.56
A:D        1 1105.56 1105.56
B:D        1    0.56    0.56
C:D        1    5.06    5.06
A:B:C      1   14.06   14.06
A:B:D      1   68.06   68.06
A:C:D      1   10.56   10.56
B:C:D      1   27.56   27.56
A:B:C:D    1    7.56    7.56
Residuals  0    0.00
Warning message:
In anova.lm(fit) :
  ANOVA F-tests on an essentially perfect fit are unreliable
```

Note: From this table, it is easy to see that the effects

$$\text{A, C, D, AC, AD}$$

are far more relevant than the others. For example, the smallest SS in this set is 390.06 (Factor C) which is over 5 times larger than the largest remaining SS (68.06). As a next step, we therefore consider fitting a smaller model with these 5 effects only. This will "free up" 10 degrees of freedom that can be used to estimate the error variance.

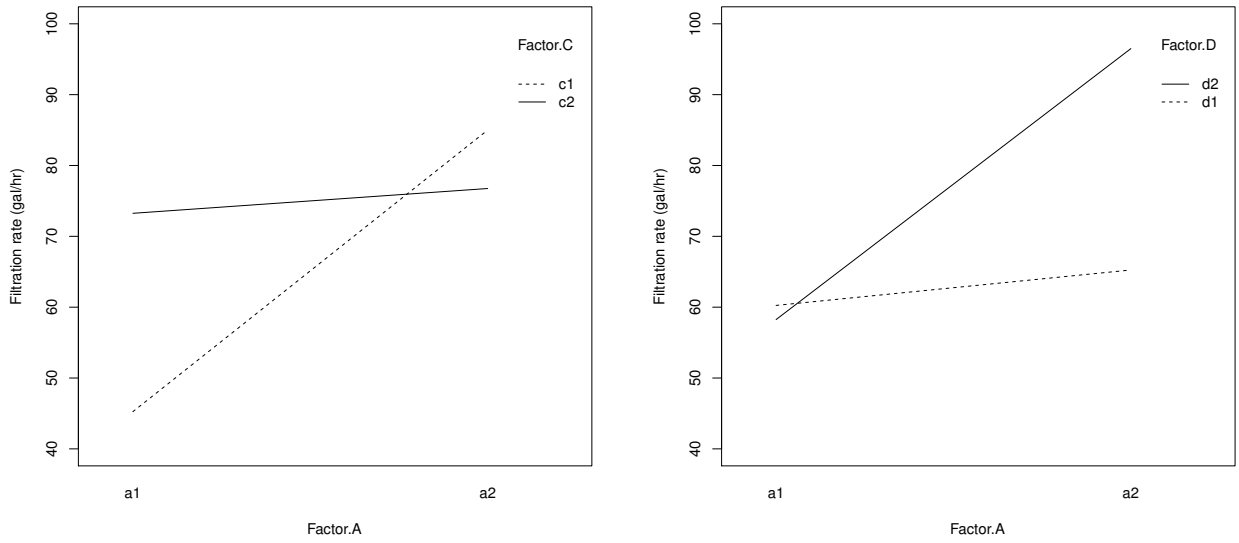

Figure 12.5: Interaction plots in Example 12.3. Left: AC. Right: AD.

Analysis: Here is the R output summarizing the fit of the smaller model that includes only A, C, D, AC, and AD:

```
> # Fit smaller model
> fit = lm(filtration~A+C+D+A*C+A*D)
> anova(fit)
Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
A          1 1870.56 1870.56  95.865 1.928e-06 ***
C          1  390.06  390.06  19.990  0.001195 **
D          1  855.56  855.56  43.847 5.915e-05 ***
A:C        1 1314.06 1314.06  67.345 9.414e-06 ***
A:D        1 1105.56 1105.56  56.659 1.999e-05 ***
Residuals 10  195.13   19.51
```

Note: Each of these five effects is strongly significant in the population. The AC (temperature/concentration of formaldehyde) and AD (temperature/stirring rate) interaction plots in Figure 12.5 each show strong interaction.

Regression analysis: In Example 12.3, even though there are no numerical values attached to the levels of temperature (Factor A), concentration of formaldehyde (Factor C), and stirring rate (Factor D), we can still use regression to estimate a population-level model. Specifically, we can introduce the following variables with arbitrary numerical codings assigned:

$$\begin{aligned} x_1 &= \text{temperature } (-1 = \text{low}; 1 = \text{high}) \\ x_2 &= \text{concentration of formaldehyde } (-1 = \text{low}; 1 = \text{high}) \\ x_3 &= \text{stirring rate } (-1 = \text{low}; 1 = \text{high}). \end{aligned}$$

With these values, we can fit the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon.$$

Doing so in R gives

```
> fit = lm(filtration~temp+conc+stir+temp:conc+temp:stir)
> fit
Coefficients:
(Intercept)         temp         conc         stir    temp:conc    temp:stir
     70.062       10.812        4.938        7.313       -9.062        8.312
```

Therefore, the estimated regression model for the filtration rate data is

$$\widehat{Y} = 70.062 + 10.812x_1 + 4.938x_2 + 7.313x_3 - 9.062x_1x_2 + 8.312x_1x_3$$

or, in other words,

$$\widehat{\texttt{FILT}} = 70.062 + 10.812\ \texttt{TEMP} + 4.938\ \texttt{CONC} + 7.313\ \texttt{STIR} - 9.062\ \texttt{TEMP*CONC} + 8.312\ \texttt{TEMP*STIR}$$

Provided that our model assumptions hold, this fitted regression model can be used to estimate the mean filtration rate (or predict a future filtration rate) at specific combinations of temperature (± 1), concentration (± 1), and stirring rate (± 1).