STAT 509-001

STATISTICS FOR ENGINEERS

Fall 2015

Lecture Notes

Dewei Wang

Department of Statistics

University of South Carolina

This set of lecture notes is based on the textbook
Montgomery, D. and Runger, G. (2014). *Applied Statistics and Probability for Engineers*,
Sixth Edition. John Wiley and Sons, Inc.
and the STAT 509 lecture notes by Dr. Joshua Tebbs.

Contents

1 Introduction

1.1 What is Statistics?

- The field of **Statistics** deals with the collection, presentation, analysis, and use of data to make decisions, solve problems, and design products and processes. (Montgomery, D. and Runger G.)
- Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (Davidian, M. and Louis, T. A., 10.1126/science.1218685).

In simple terms, **statistics is the science of data**.

1.2 Where to Use Statistics?

- Statisticians apply statistical thinking and methods to a wide variety of scientific, social, and business endeavors in such areas as astronomy, biology, education, economics, engineering, genetics, marketing, medicine, psychology, public health, sports, among many. **"The best thing about being a statistician is that you get to play in everyone else's backyard."** (John Tukey, Bell Labs, Princeton University)

Here are some examples where statistics could be used:

1. In a reliability (time to event) study, an engineer is interested in quantifying the time until failure for a jet engine fan blade.
2. In an agricultural study in Iowa, researchers want to know which of four fertilizers (which vary in their nitrogen contents) produces the highest corn yield.
3. 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.
4. In a public health study, epidemiologists want to know whether smoking is linked to a particular demographic class in high school students.
5. A food scientist is interested in determining how different feeding schedules (for pigs) could affect the spread of salmonella during the slaughtering process.
6. A research dietician wants to determine if academic achievement is related to body mass index (BMI) among African American students in the fourth grade.

Remark 1. 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 the researcher), to test the validity of important conjectures, and to predict future behavior. Being able to identify and model sources of **variability** is an important part of this process.

1.3 Deterministic and Statistical Models

- A **deterministic model** is one that makes no attempt to explain variability. For example, 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 built from our underlying knowledge of the basic physical mechanism. It is completely determined without any 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 calculate 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 **statistical** (or **stochastic**) **model** might look like

  $$V = IR + \epsilon,$$

  where ϵ is a random term that includes the effects of all unmodeled sources of variability that affect this system.

1.4 Statistical Inference

There are two main types of statistics:

- **Descriptive statistics** describe what is happening now (see Chapter 6 of the textbook).

- **Inferential statistics**, such as estimation and prediction, are based on a sample of the subjects (only a portion of the population) to determine what is probably happening or what might happen in the future.

Example 1.4.1. Let us consider semiconductors. A finished semiconductor is wire-bounded to a frame. Suppose that I am trying to model

$$Y = \text{pull strength (a measure of the amount of force required to break the bond)}$$

of a semiconductor. The population herein could be all the finished semiconductor. A sample of size 25 was collected and from each I measured the pull strength (Y), the wire length (x_1) and the die height (x_2). All 25 observations are plotted in Figure 1.4.1a.

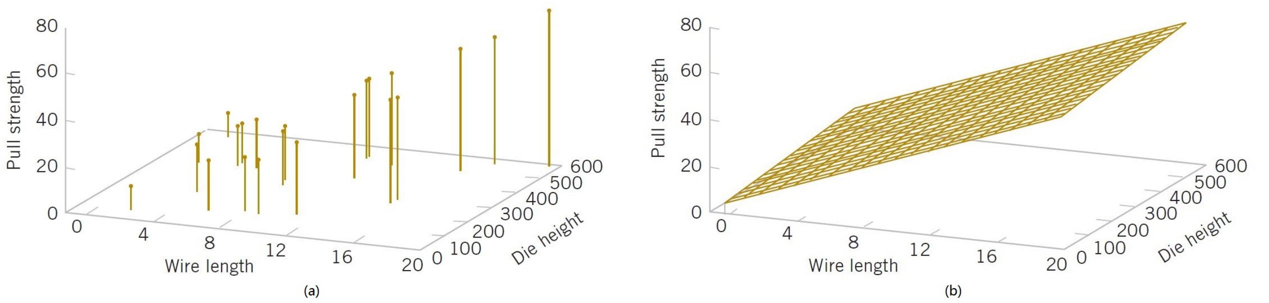


Figure 1.4.1: (a). Three-dimensional plot of the wire bond pull strength data; (b). Plot of predicted values of pull strength from the estimated model.

The goal here is to build a model that can quantify the relationship between pull strength and the variables wire length and die height. A **deterministic model** would be

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

for some unknown function $f : [0,\infty)\times[0,\infty) \to [0,\infty)$. Perhaps a working model could be developed as a **statistical model** of the form:

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

where ϵ is a random term that accounts for not only measurement error but also

(a) all of the other variables not accounted for (e.g., the quality of the wire and/or how all the welding has been done, etc.) and

(b) the error induced by assuming a **linear relationship** between Y and $\{x_1, x_2\}$ when, in fact, it may not be.

In this example, with certain (probabilistic) assumptions on ϵ and a mathematically sensible way to **estimate** the unknown β_0, β_1, and β_2 (i.e., coefficients of the linear function), we can produce point **predictions** of Y for any given $\{x_1, x_2\}$. Using the regression technique (Chapter 12) results in an estimated model (plotted in Figure 1.4.1b)

$$\hat{Y} = 2.26 + 2.74x_1 + 0.0125x_2.$$

It naturally brings up the following questions:

- How accurate are the estimators of the coefficients or the prediction for a given $\{x_1, x_2\}$?
- How significant are the roles of x_1 and x_2?
- How should samples be selected to provide good decisions with acceptable risks?

To answer these questions or to quantify the risks involved in statistical inference, it leads to the study of **probability models**.

2 Probability

If we measure the current in a thin copper wire, we are conducting an experiment. However, day-to-day repetitions of the measurement can differ slightly because of

- changes in ambient temperatures
- slight variations in the gauge
- impurities in the chemical composition of the wire (if selecting different locations)
- current source drifts.

In some cases, the random variations are small enough, relative to our experimental goals, that they can be ignored. However, no matter how carefully our experiment is designed and conducted, the variation is almost always present, and its magnitude can be large enough that the important conclusions from our experiment are not obvious. Hence, how to quantify the variability is a key question, which can be answered by probability.

2.1 Sample Spaces and Events

An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a **random experiment**.
The set of all possible outcomes of a random experiment is called the **sample space** of the experiment. The sample space is denoted as S.
A sample space is **discrete** if it consists of a finite or countable infinite set of outcomes.
A sample space is **continuous** if it contains an interval (either finite or infinite) of real numbers.

Example 2.1.1. Let us find the sample space for each of the following random experiments and identify whether it is discrete or continuous:

- The number of hits (views) is recorded at a high-volume Web site in a day
- The pH reading of a water sample.
- Calls are repeated place to a busy phone line until a connection is achieved.
- A machined part is classified as either above or below the target specification.
- The working time or surviving time of an air conditioner.

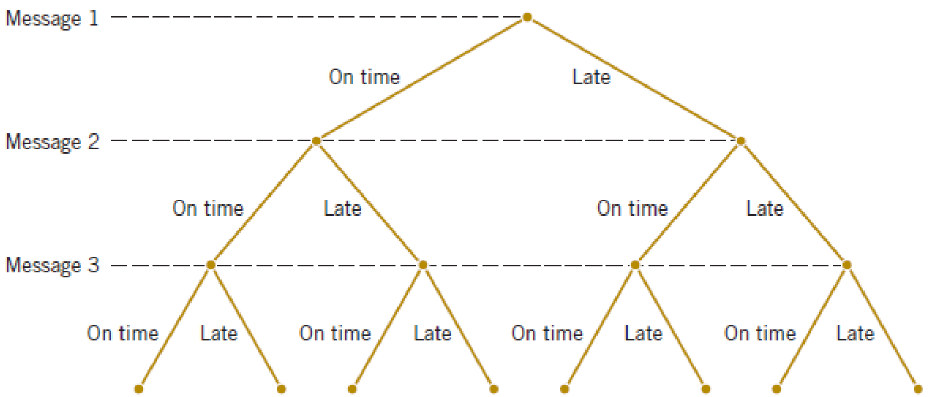


Figure 2.1.1: Tree diagram for three messages.

Example 2.1.2. (Tree diagram) Now let us consider a little bit more complex case. Each message in a digital communication system is classified as to whether it is received on time or late. Describe the sample space of the receive time of three messages.

$$S =$$

An **event** is a subset of the sample space of a random experiment. The following are three basic set operations:

- The **union** of two events is the event that consists of all outcomes that are contained in either of the two events. We denote the union as $E_1 \cup E_2$.
- The **intersection** of two events is the event that consists of all outcomes that are contained in both of the two events. We denote the intersection as $E_1 \cap E_2$.
- The **complement** of an event in a sample space is the set of outcomes in the sample space that are not in the event. We denote the complement of the event E as E'. The notation E^c is also used in other literature to denote the complement.

Example 2.1.3. Consider Example 2.1.2. Denote that E_1 is the event that at least two messages is received late. Then $E_1 = \{100, 010, 001, 000\}$. Let E_2 be the event that the second messages is received later. Then $E_2 = \{101, 100, 001, 000\}$. Now we have

$$\begin{aligned} E_1 \cup E_2 &= \\ E_1 \cap E_2 &= \\ E_1' &= \end{aligned}$$

Example 2.1.4. As in Example 2.1.1, the sample space of the working time of an air conditioner is $S = (0, \infty)$. Let E_1 be the event the working time is no less than 1 and less than 10; i.e., $E_1 = \{x \mid 1 \leq x < 10\} = [1, 10)$, and E_2 be the event the working time is between 5 and 15; i.e., $E_2 = \{x \mid 5 < x < 15\} = (5, 15)$. Then

$$\begin{aligned} E_1 \cup E_2 &= & \qquad\qquad\qquad\qquad E_1 \cap E_2 &= \\ E_1' &= & & \\ E_1' \cap E_2 &= & & \end{aligned}$$

One visualized way to interpret set operations is through **Venn diagrams**. For example

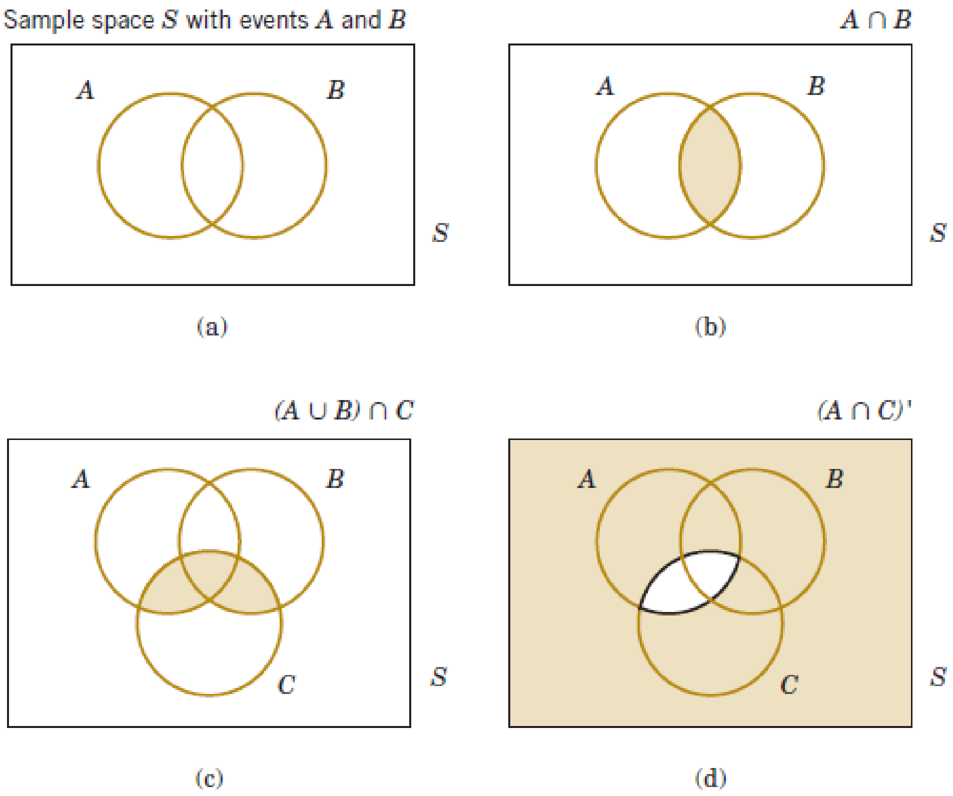
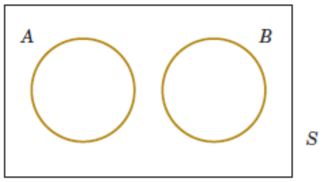


Figure 2.1.2: Venn diagrams.

Two events, denoted as A and B, such that $A \cap B = \emptyset$, i.e.,



are said to be **mutually exclusive**.

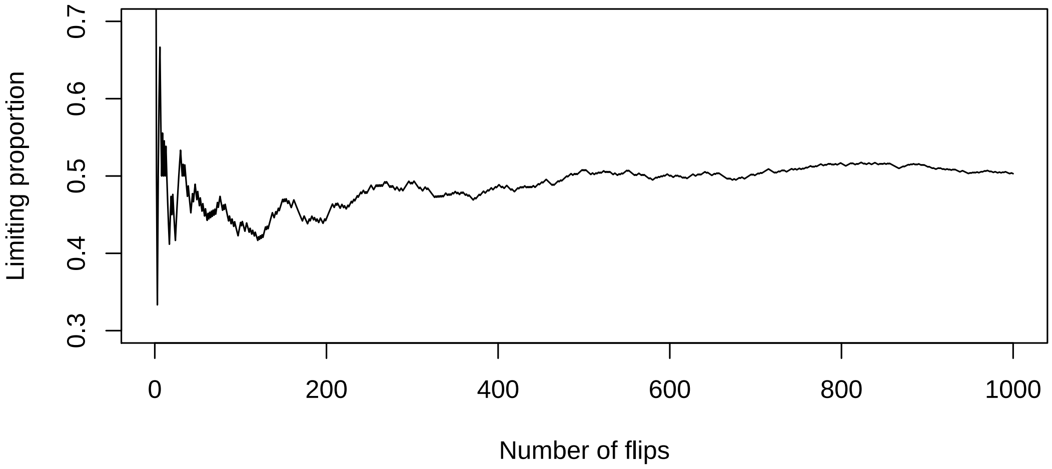
2.2 Axioms of Probability and Addition Rule

Probability is used to quantify the likelihood, or chance, that an outcome of a random experiment will occur. The probability of an event E is denoted by $P(E)$.

"My chance of getting an A in this course is 80%" could be a statement that quantifies your feeling about the possibility of getting A. The likelihood of an outcome is quantified by assigning a number from the interval [0, 1] to the outcome (or a percentage from 0 to 100%). Higher numbers indicate that the outcome is more likely than lower numbers. A 0 indicates an outcome will not occur. A probability of 1 indicates that an outcome will occur with certainty. The probability of an outcome can be interpreted as our subjective probability, or **degree of belief**, that the outcome will occur. Different individuals will no doubt assign different probabilities to the same outcomes.

Another interpretation of probability is based on the conceptual model of repeated replications of the random experiment. The probability of an outcome is interpreted as the limiting value of the proportion of times the outcome occurs in n repetitions of the random experiment as n increases beyond all bounds. For example, we want to quantify the probability of the event that flipping a fair coin gets a head. One way is to flip a fair coin n times, and record how many times you get a head. Then

$$P(\text{flipping a fair coin gets a head}) = \lim_{n\to\infty} \frac{\text{number of heads out of } n \text{ flips}}{n} = \frac{1}{2}.$$



This type of experiment is said of **equally likely outcomes**.

Equally Likely Outcomes: Whenever a sample space consists of N possible outcomes that are equally likely, the probability of each outcome is $1/N$.

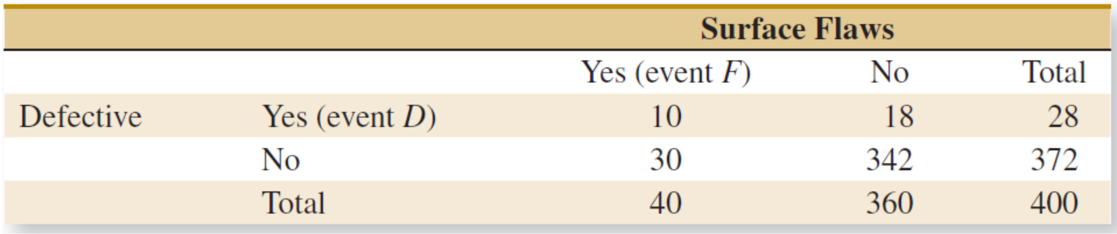
For example, we want to detect the rate of defectiveness of products form a same product line. The number of products could be million. It is time-consuming and expensive to exam every product. People usually *randomly* select a certain number of product and count how many of them are defective. We call the selected items as a **random samples**.

To select **ramdonly** implies that at each step of the sample, the remained items are equally likely to be selected. .

It means that, suppose there are N items. When drawing the first sample, each item has the chance of $1/N$ being selected. To select the second sample, each of the $N-1$ remained items will be selected with probability $1/(N-1)$, so and so on.

Another interpretation of probability is through **relative frequency**.

Example 2.2.1. The following table provides an example of 400 parts classified by surface flaws and as (functionally) defective.



		Surface Flaws		
		Yes (event F)	No	Total
Defective	Yes (event D)	10	18	28
	No	30	342	372
	Total	40	360	400

Then

$$P(\text{defective}) = P(D) =$$
$$P(\text{surface flaws}) = P(F) =$$
$$P(\text{surface flaws and also defective}) = P(D \cap F) =$$
$$P(\text{surface flaws but not defective}) = P(D' \cap F) =$$

For a discrete sample space, $P(E)$ equals the sum of the probabilities of the outcomes in E.

Example 2.2.2. A random experiment can result in one of the outcomes $\{a, b, c, d\}$ with probabilities $0.1, 0.3, 0.5$, and 0.1, respectively. Let A denote the event $\{a, b\}$, B the event $\{b, c, d\}$ and C the event $\{d\}$. Then

$P(A) =$ $\quad$ $P(B) =$ $\quad$ $P(C) =$

$P(A') =$ $\quad$ $P(B') =$ $\quad$ $P(C') =$

$P(A \cap B) =$

$P(A \cup B) =$

$P(A \cap C) =$

Axioms of Probability: Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties: if S is the sample space and E is any event in a random experiment,

1. $P(S) = 1$
2. $0 \leq P(E) \leq 1$
3. For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$ (mutually exclusive),

$$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

These axioms imply the following results. The derivations are left as exercises at the end of this section. Now,

$$P(\emptyset) = \underline{\hspace{4cm}}$$

and for any event E,

$$P(E') = \underline{\hspace{4cm}}$$

Furthermore, if the event E_1 is contained in the event E_2,

$$P(E_1)\underline{\hspace{2cm}}P(E_2).$$

Addition rule:

$$P(A \cup B) = \underline{\hspace{12cm}}$$

A collection of events, $E_1, E_2, \ldots, E_k$, is said to be mutually exclusive if for all pairs,

$$E_i \cap E_j = \emptyset.$$

For a collection of mutually exclusive events,

$$P(E_1 \cup E_2 \cup \cdots \cup E_k) = \underline{\hspace{12cm}}$$

Example 2.2.3. Let $S = [0, \infty)$ be the sample space of working time of an air conditioner. Define the events $E_1 = (2, 10)$, $E_2 = (5, 20)$, $E_3 = (5, 10)$, $E_4 = (0, 2]$. Suppose $P(E_1) = .4$, $P(E_2) = 0.7$, $P(E_3) = 0.2$, $P(E_4) = .05$. Then

$$P(E_5 = (2, 20)) = \underline{\hspace{8cm}}$$
$$P(E_6 = (0, 20)) = \underline{\hspace{8cm}}$$

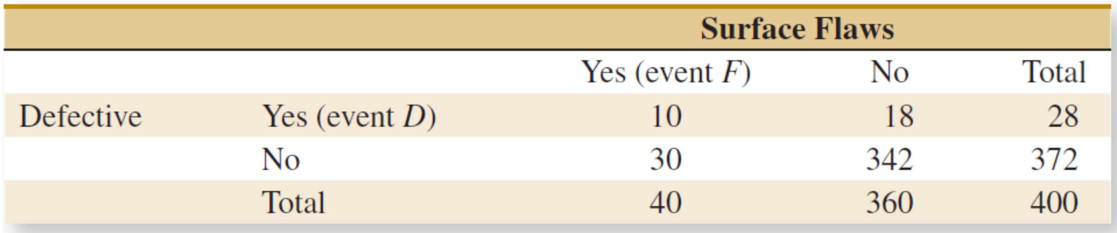
2.3 Conditional Probability and Multiplication Rule

Sometimes probabilities need to be reevaluated as additional information becomes available. A useful way to incorporate additional information into a probability model is to assume that the outcome that will be generated is a member of a given event. This event, say A, defines the conditions that the outcome is known to satisfy. Then probabilities can be revised to include this knowledge. The probability of an event B under the knowledge that the outcome will be in event A is denoted as

____

and this is called the conditional probability of B given A.

Example 2.3.1. Let consider Example 2.2.1.



		Surface Flaws		
		Yes (event F)	No	Total
Defective	Yes (event D)	10	18	28
	No	30	342	372
	Total	40	360	400

Of the parts with surface flaws (40 parts), the number of defective ones is 10. Therefore,

$P(D \mid F) =$ ____

and of the parts without surface flaws (360 parts), the number of defective ones is 18. Therefore,

$P(D \mid F') =$ ____

Practical Interpretation: The probability of being defective is five times greater for parts with surface flaws. This calculation illustrates how probabilities are adjusted for additional information. The result also suggests that there may be a link between surface flaws and functionally defective parts, which should be investigated.

The **conditional probability** of an event B given an event A, denoted as $P(B \mid A)$, is

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

Recalculate the probabilities in last example, we have

$P(D \mid F) =$ ____

$P(D \mid F') =$ ____

Multiplication Rule:

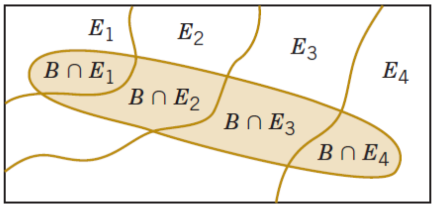
$$P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B).$$

Total Probability Rule (Multiple Events): A collection of sets $E_1, E_2, \ldots, E_k$ is said to be exhaustive if and only if

$$E_1 \cup E_2 \cup \cdots \cup E_k = S.$$

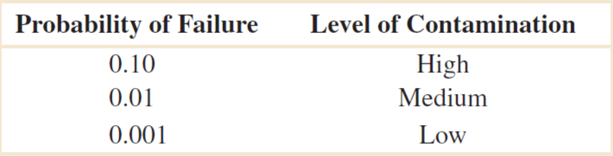
Assume $E_1, E_2, \ldots, E_k$ are k mutually exclusive and exhaustive sets, then for any event B, we have

$$\begin{aligned} P(B) =& P(B \cap E_1) + P(B \cap E_2) + \cdots + P(B \cap E_k) \\ =& P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_k)P(E_k). \end{aligned}$$



$B = (B \cap E_1) \cup (B \cap E_2) \cup (B \cap E_3) \cup (B \cap E_4)$

Example 2.3.2. Assume the following probabilities for product failure subject to levels of contamination in manufacturing:



Probability of Failure	Level of Contamination
0.10	High
0.01	Medium
0.001	Low

In a particular production run, 20% of the chips are subjected to high levels of contamination, 30% to medium levels of contamination, and 50% to low levels of contamination. What is the probability of the event F that a product using one of these chips fails?

Let

- H denote the event that a chip is exposed to high levels of contamination
- M denote the event that a chip is exposed to medium levels of contamination
- L denote the event that a chip is exposed to low levels of contamination

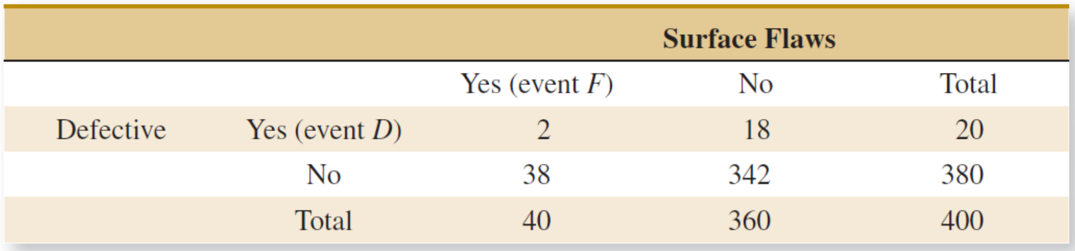
Then

$$\begin{aligned} P(F) &= \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad} \\ &= \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad} \end{aligned}$$

2.4 Independence

In some cases, the conditional probability of $P(B \mid A)$ might equal $P(B)$; i.e., the outcome of the experiment is in event A does not affect the probability that the outcome is in event B.

Example 2.4.1. As in Example 2.2.1, surface flaws related to functionally defective parts since $P(D \mid F) = 0.25$ and $P(D) = 0.07$. Suppose now the situation is different as the following Table.



		Surface Flaws		
		Yes (event *F*)	No	Total
Defective	Yes (event *D*)	2	18	20
	No	38	342	380
	Total	40	360	400

Then,

$$P(D \mid F) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \quad \text{and} \quad P(D) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}.$$

That is, the probability that the part is defective does not depend on whether it has surface flaws. Also,

$$P(F \mid D) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \quad \text{and} \quad P(F) = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

so the probability of a surface flaw does not depend on whether the part is defective. Furthermore, the definition of conditional probability implies that $P(F \cap D) = P(D \mid F)P(F)$, but in the special case of this problem,

$$P(F \cap D) = P(D)P(F).$$

Two events are **independent** if any one of the following equivalent statements is true:

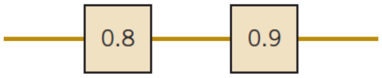
1. $P(A \mid B) = P(A)$
2. $P(B \mid A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

Noting that when A and B are independent events,

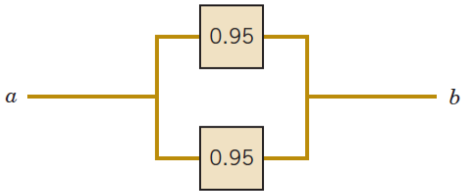
$$\begin{aligned} P(A' \cap B') &= \underline{\hspace{8cm}} \\ &= \underline{\hspace{8cm}} \\ &= \underline{\hspace{8cm}} \end{aligned}$$

Question: If A and B are mutually exclusive, and $P(A) > 0$, $P(B) > 0$, Are A and B independent?

Example 2.4.2. (**Series Circuit**) The following circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that devices fail independently. What is the probability that the circuit operates?



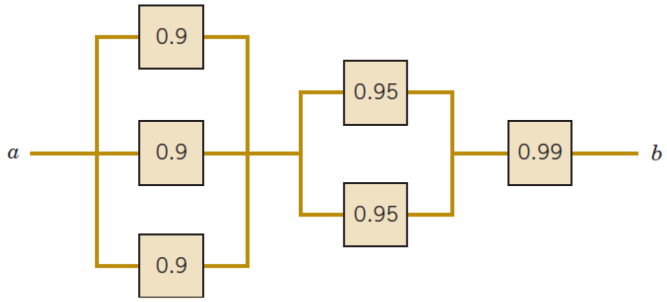
Example 2.4.3. (**Parallel Circuit**) The following circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that devices fail independently. What is the probability that the circuit operates?



If the events $E_1, E_2, \ldots, E_k$ are independent, then

$$P(E_1 \cap E_2 \cap \cdots \cap E_k) = P(E_1)P(E_2) \cdots P(E_k).$$

Example 2.4.4. (**Advanced Circuit**) The following circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that devices fail independently. What is the probability that the circuit operates?



3 Random Variables and Probability Distributions

A **random variable** is a function that assigns a real number to each outcome in the sample space of a random experiment.
A **discrete random variable** is a random variable with a finite (or countably infinite) range.
A **continuous random variable** is a random variable with an interval (either finite or infinite) of real numbers for its range.
Notation: A random variable is denoted by an uppercase letter such as X and Y. After experiment is conducted, the measured value of the random variable is denoted by a lowercase letter such as x and y.

For example, let X be a random variable denoting the outcome of flipping a coin. The sample space of this random experiment is $\{head, tail\}$. We can let $X = 1$ if it is a head; $X = 0$ otherwise. When you are actually conduct this experiment, you may observe a head. Then the notation for describing this observation is $x = 1$.

- Examples of discrete random variables: result of flipping a coin, number of scratches on a surface, proportion of defective parts among 1000 tested, number of transmitted bits received in error.

- Examples of continuous random variables: electrical current, length, pressure, temperature, time, voltage, weight.

3.1 General Discrete Distributions

3.1.1 Probability Mass Function

The **probability distribution** of a random variable X is a description of the probabilities associated with the possible values of X. For a discrete random variable, the distribution is often specified by just a list of the possible values along with the probability of each. In some cases, it is convenient to express the probability in terms of a formula.

For a discrete random variable X with possible values $x_1, x_2, \dots, x_k$, a **probability mass function (pmf)** is a function such that

(1) $f(x_i) \geq 0$

(2) $\sum_{i=1}^{k} f(x_i) = 1$

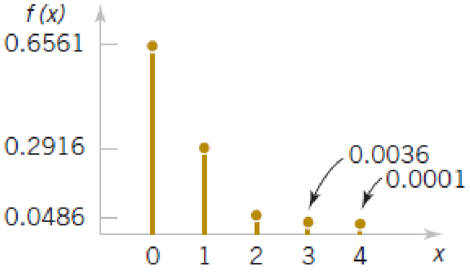
(3) $f(x_i) = P(X = x_i)$

Example 3.1.1. (**Digital Channel**) There is a chance that a bit transmitted through a digital transmission channel is received in error. Let X equal the number of bits in error in the next four bits transmitted. The possible values for X are $\{0, 1, 2, 3, 4\}$. Based on a model for the errors that is presented in the following section, probabilities for these values will be determined.

$$\begin{aligned} &P(X = 0) = 0.6561, \ \ P(X = 1) = 0.2916, \\ &P(X = 2) = 0.0486, \ \ P(X = 3) = 0.0036, \\ &P(X = 4) = 0.0001. \end{aligned}$$

The probability distribution of X is specified by the possible values along with the probability of each. The pmf of X is then

A graphical description of the probability distribution of X is shown as



Once a probability mass function of X is presented, one should be able to calculate all types of events in the sample space; i.e., $P(X \leq a), P(X < a), P(X \geq a), P(X > a), P(a < X < b), P(a \leq X < b), P(a < X \leq b), P(a \leq X \leq b)$. For example, in the example above,

$P(X < 1) =$____

$P(X \leq 1) =$____

$P(X \leq 3) - P(X \leq 2) =$____

$P(1 \leq X < 3) =$____

$P(1 < X < 3) =$____

$P(1 < X \leq 3) =$____

$P(1 \leq X \leq 3) =$____

3.1.2 Cumulative Distribution Function

The **cumulative distribution function (cdf)** of a discrete random variable X, denoted as $F(X)$, is

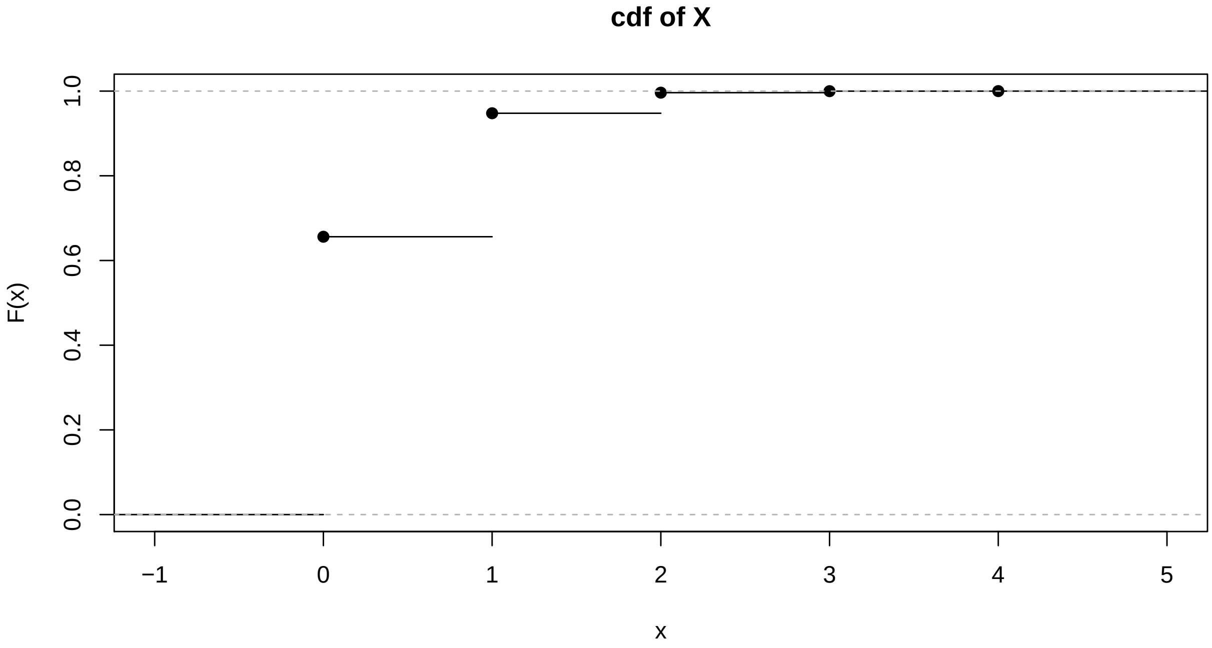
$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

$F(x)$ satisfies the following properties.

1. $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$
2. $0 \leq F(x) \leq 1$
3. if $x \leq y$, then $F(x) \leq F(y)$

For the last example, the cdf function is then

The cdf function can be plotted as



And based on cdf $F(x)$, you should be able to calculate the probabilities of the following types

$P(X < 1) =$ ______________________________

$P(X \leq 1) =$ ______________________________

$P(X \leq 3) - P(X \leq 2) =$ ______________________________

$P(1 \leq X < 3) =$ ______________________________

$P(1 < X < 3) =$ ______________________________

$P(1 < X \leq 3) =$ ______________________________

$P(1 \leq X \leq 3) =$ ______________________________

3.1.3 Mean and Variance

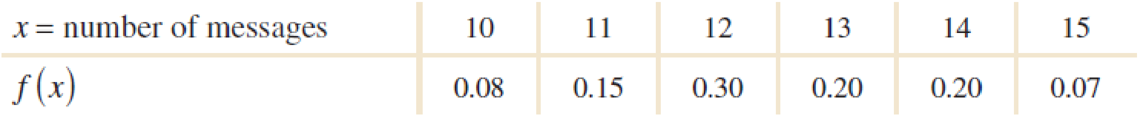
Two numbers are often used to summarize a probability distribution for a random variable X. The **mean** is a measure of the center or middle of the probability distribution, and the **variance** is a measure of the dispersion, or variability in the distribution.

The **mean** or **expected value** of the discrete random variable X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_{x} x f(x).$$

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

Example 3.1.2. The number of email messages received per hour has the following distribution:



x = number of messages	10	11	12	13	14	15
$f(x)$	0.08	0.15	0.30	0.20	0.20	0.07

Determine the mean and standard deviation of the number of messages received per hour.

$\mu =$ ______________________________

Interpretation: On average, we would expect _______ email messages per hour.

Interpretation: Over the long run, if we observed many values of Y with this pmf, then the average of these X observations would be close to _______

Let X be a discrete random variable with pmf $f(x)$. Suppose that g, g_1, g_2, ..., g_k are real-valued functions, and let c be any real constant.

$$E[g(X)] = \sum_{\text{all } x} g(x)f(x).$$

Further expectations satisfy the following (linearity) properties:

1. $E(c) = c$
2. $E[cg(X)] = cE[g(X)]$
3. $E[\sum_{j=1}^{k} g_j(X)] = \sum_{j=1}^{k} E[g_j(X)]$

For linear function $g(x) = ax + b$ where a, b are constants, we have

$$E[g(X)] = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Note: These rules are also applicable if X is continuous (coming up).

Example 3.1.3. In Example 3.1.2, suppose that each email message header reserves 15 kilobytes of memory space for storage. Let the random variable Y denote the memory space reserved for all message headers per hour (in kilobytes). Then

$$Y = \_\_\_\_\_\_\_\_\_\_$$

Thus

$$E(Y) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

The expected reserved memory space for all message headers per hour is __________

The population **variance** of X, denoted as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2 = E(X^2) - [E(X)]^2.$$

The population **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

Facts: The population variance σ^2 satisfies the following:

1. $\sigma^2 \geq 0$. $\sigma^2 = 0$ if and only if the random variable Y has a **degenerate distribution**; i.e., all the probability mass is located at one support point.
2. The larger (smaller) σ^2 is, the more (less) spread in the possible values of X about the population mean $\mu = E(X)$.
3. σ^2 is measured in $(\text{units})^2$ and σ is measured in the original units.

Let X be a discrete random variable with pmf $f(x)$. Suppose $g(x) = ax + b$ where a, b are constants, we have

$$V[aX + b] = \underline{\hspace{8cm}}$$

Note: These rules are also applicable if X is continuous (coming up).

In Example 3.1.3, we have

$$V(Y) = \underline{\hspace{8cm}}$$

The variance of reserved memory space for all message headers per hour is ______________

The measures of mean and variance do not uniquely identify a probability distribution. That is, two different distributions can have the same mean and variance. Still, these measures are simple, useful summaries of the probability distribution of X.

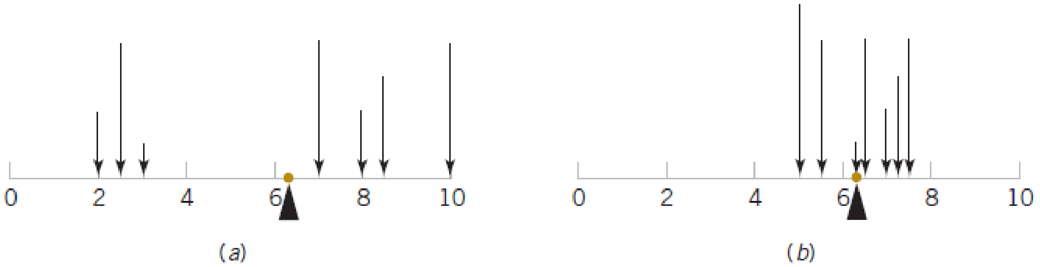


FIGURE 3-5 A probability distribution can be viewed as a loading with the mean equal to the balance point. Parts (a) and (b) illustrate equal means, but part (a) illustrates a larger variance.

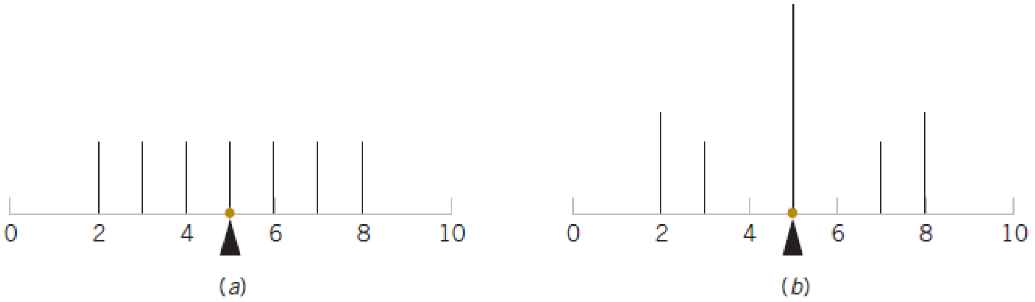


FIGURE 3-6 The probability distributions illustrated in parts (a) and (b) differ even though they have equal means and equal variances.

3.2 Bernoulli Distribution and Binomial Distribution

Let us consider the following random experiments and random variables:

1. A worn machine tool produces 1% defective parts. Let X = number of defective parts in the next 25 parts produced.

2. Each sample of air has a 10% chance of containing a particular rare molecule. Let X = the number of air samples that contain the rare molecule in the next 18 samples analyzed.

3. Of all bits transmitted through a digital transmission channel, 40% are received in error. Let X = the number of bits in error in the next five bits transmitted.

4. A multiple-choice test contains 10 questions, each with four choices, and for each question, the chance of you gets right is 90%. Let X = the number of questions answered correctly.

5. In the next 20 births at a hospital, let X = the number of female births.

Each of these random experiments can be thought of as consisting of a series of repeated, random trials:

1. The production of 25 parts in the 1st example

2. Detecting rare molecule in 18 samples of air

3. Counting errors in 5 transmitted bits

4. Answering 10 multiple-choice questions

5. Gender of the next 20 babies

Each of the repeated trials consists of two possible outcomes: (generally speaking) **success** and **failure**, and we want to know how many (generally speaking) **successes** occur in the a certain number of trials. The terms **success** and **failure** are just labels. Sometime it can mislead you. For example, in the 1st example, we are interested in the number of *defective* parts (**herein, "success" means defective**).

To model a trial with two outcomes, we typically use **Bernoulli Distribution**. We say random variable X follows a Bernoulli distribution, if it has the following probability mass function:

$$f(x) = \begin{cases} p & \text{if } x = 1, \text{ represents success} \\ 1 - p & \text{if } x = 0, \text{ represents failure} \end{cases}$$

The mean and variance of X are

$$\mu = E[X] = \underline{\hspace{8cm}}$$
$$\sigma^2 = V[X] = \underline{\hspace{8cm}}$$

Now, let us get back to our original examples. What we are interested in is the number of successes occurs in a certain number of identical trails, each trail has two possible outcomes (success and failure) with certain probability of success. Thus, we are investigating the summation of a given number of identical Bernoulli random variables.

1. In the first example: we investigate the random variable X is the summation of $n = \underline{25}$ identical Bernoulli random variables, each of which has two possible outcomes (defective = "success," indefective="failure"), with probability of success being $p = \underline{0.01}$
2. In the second example: we investigate the random variable X is the summation of $n =$ ______ identical Bernoulli random variables, each of which has two possible outcomes (____________ = "success," ____________="failure"), with probability of success being $p =$ ______
3. In the third example: we investigate the random variable X is the summation of $n =$ ______ identical Bernoulli random variables, each of which has two possible outcomes (____________ = "success," ____________="failure"), with probability of success being $p =$ ______
4. In the forth example: we investigate the random variable X is the summation of $n =$ ______ identical Bernoulli random variables, each of which has two possible outcomes (____________ = "success," ____________="failure"), with probability of success being $p =$ ______
5. In the fifth example: we investigate the random variable X is the summation of $n =$ ______ identical Bernoulli random variables, each of which has two possible outcomes (____________ = "success," ____________="failure"), with probability of success being $p =$ ______

To model these quantities, one commonly used distribution is **Binomial Distribution**: Suppose that n **independent** and **identical** Bernoulli trials are performed. Define

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

We say the X has a **Binomial Distribution** with number of trials n and success probability p. Shorthand notation is $X \sim B(n, p)$. The probability mass function of X is given by

$$f(x) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}, & x = 0, 1, 2, 3, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

where

$$\begin{pmatrix} n \\ x \end{pmatrix} = \frac{n!}{x!(n-x)!}, \text{ and } r! = r \times (r-1) \times \cdots \times 2 \times 1 \text{ (note } 0! = 1).$$

The mean and variance are

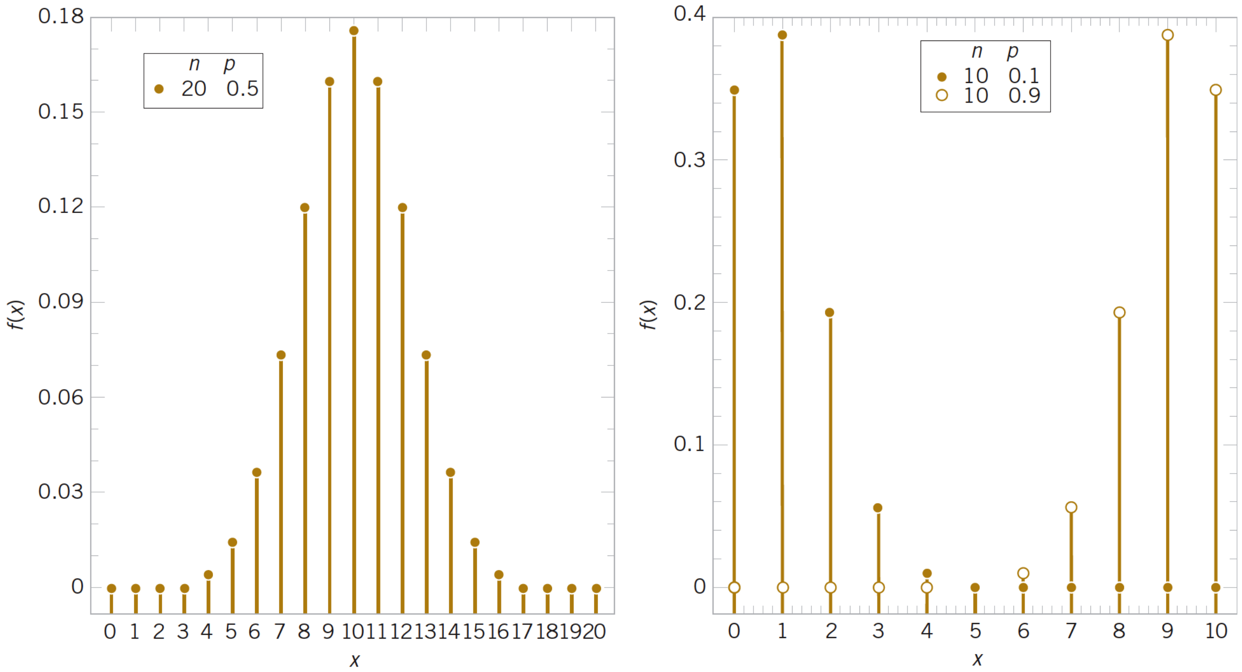
$$\mu = E[X] = \underline{\hspace{8cm}}$$
$$\sigma^2 = V[X] = \underline{\hspace{8cm}}$$

There are three key elements for correctly identifying a Bernoulli distribution:

(1) The trials are independent.

(2) Each trial results in only two possible outcomes, labeled as “success” and “failure.”

(3) The probability of a success in each trial, denoted as p, remains constant.



(a) (b)

Let us see the 3rd example: Of all bits transmitted through a digital transmission channel, 40% are received in error. Let X = the number of bits in error in the next five bits transmitted. Now we calculate $P(X = 2)$ by assuming all the transmitted bits are independent.

Thus, from above we can see that X is actually a Binomial random variable; i.e., $X \sim B(5, 0.4)$. Now let us answer the following questions:

(a) What is the probability that at least one bits are received in error?

(b) What are $E(X)$ and $V(X)$?

Now considering the first example, we have $X \sim$ ______________, what is the probability when $X \leq 10$? Computing this probability "by hand" could be very time-consuming. We will use TI-84. The codes are (in "DISTR"):

$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
binompdf$(n,\ p,\ x)$	**binomcdf**$(n,\ p,\ x)$

(a) What is the probability that there are exactly five defective parts?

(b) What is the probability that there are at least five defective parts?

(c) What is the probability that there are at most ten defective parts?

(d) What is $P(2 \leq X \leq 8)$? {Hint: a general formula $P(a < X \leq b) = F(b) - F(a)$}

3.3 Geometric Distributions

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 a “success,” denoted by p, remains constant on every trial.

However, instead of fixing a certain number of trials and then finding out how many of them are successes, trials are conducted until a success is obtained.

Suppose that Bernoulli trials are continually observed. Define

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

We say that X has a geometric distribution with success probability p. Shorthand notation is $X \sim \text{Geom}(p)$.

Example 3.3.1. The probability that a bit transmitted through a digital transmission channel is received in error is 0.1. Assume that the transmissions are independent events, and let the random variable X denote the number of bits transmitted until the first error. Calculate $P(X = 5)$.

If $X \sim \text{Geom}(p)$, then the probability mass function of X is given by

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

And the mean and variance of X are

$$\mu = E(X) = \frac{1}{p}$$
$$\sigma^2 = V(X) = \frac{1-p}{p^2}.$$

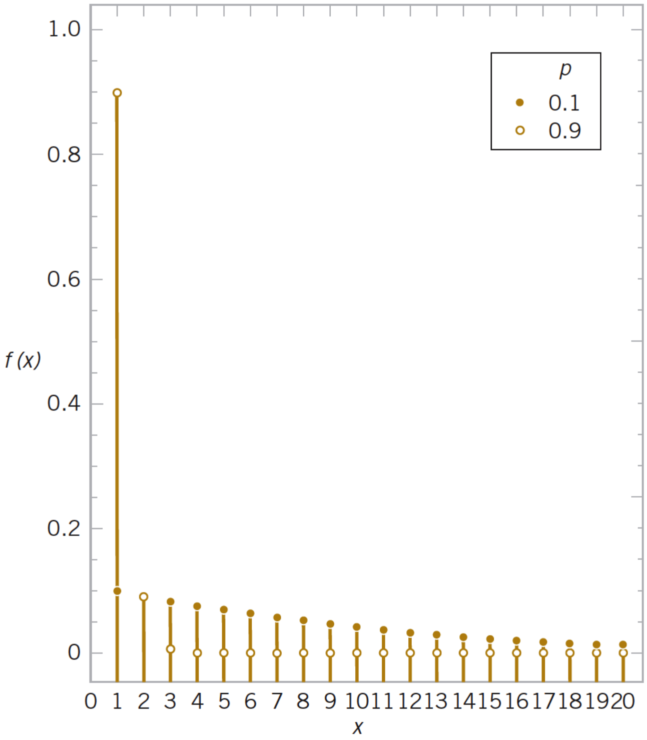


Figure 3.3.1: Geometric distribution of selected values of the parameter p.

Back to Example 3.3.1, the probability mass function is plotted in the above figure with solid dots. And

$$\mu = E(X) = \underline{\hspace{8cm}}$$
$$\sigma^2 = V(X) = \underline{\hspace{8cm}}$$

if $X \sim \text{Geom}(p)$, its cumulative distribution function is

$$F(x) = P(X \leq x) = \sum_{k=1}^{x} f(k) = \sum_{k=1}^{x} (1-p)^{k-1} p = 1 - (1-p)^x.$$

Example 3.3.2. Biology students are checking the eye color of fruit flies. For each fly, the probability of observing white eyes is $p = 0.25$. In this situation, we interpret the Bernoulli trials as

- 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

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

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

(b) What is the probability the first white-eyed fly is observed before the fourth fly is examined?

3.4 Negative Binomial Distributions

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 a "success," denoted by p, remains constant on every trial.

However, instead of fixing a certain number of trials and then finding out how many of them are successes, trials are conducted until the $\boldsymbol{r}$**th** success is obtained.

Suppose that Bernoulli trials are continually observed. Define

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

We say that X has a negative binomial distribution with waiting parameter r and success probability p. Shorthand notation is $X \sim \text{NB}(r, p)$.

Example 3.4.1. The probability that a bit transmitted through a digital transmission channel is received in error is 0.1. Assume that the transmissions are independent events, and let the random variable X denote the number of bits transmitted until the 3rd error. Calculate $P(X = 5)$.

If $X \sim \text{NB}(r,p)$, then the probability mass function of X is given by

$$f(x) = \begin{cases} \left(\begin{array}{c} x-1 \\ r-1 \end{array} \right) p^r (1-p)^{(x-r)}, & x = r, r+1, r+2, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

And the mean and variance of X are

$$\begin{aligned} \mu = E(X) &= \frac{r}{p} \\ \sigma^2 = V(X) &= \frac{r(1-p)}{p^2}. \end{aligned}$$

Note that the negative binomial distribution is a mere generalization of the geometric. If $r = 1$, then the $\text{NB}(r,p)$ distribution reduces to the $\text{Geom}(p)$.

Back to Example 3.4.1, the probability mass function is plotted in the above figure with solid dots. And

$$\begin{aligned} \mu = E(X) &= \underline{\hspace{6cm}} \\ \sigma^2 = V(X) &= \underline{\hspace{6cm}} \end{aligned}$$

To calculate $f(x) = P(X = x)$ when $X \sim \text{NB}(r,p)$, one could use the TI-84:

$$f(x) = P(X = x) = p \times \textbf{binompdf}(x-1, p, r-1).$$

Unfortunately, there is no simple TI-84 codes to calculate the cumulative distribution function of a negative binomial distribution; i.e., $F(x) = P(X \leq x)$. The only way is

$$F(x) = \sum_{k=r}^{x} p \times \textbf{binompdf}(k-1, p, r-1) = p \times \left\{ \sum_{k=r}^{x} \textbf{binompdf}(k-1, p, r-1) \right\}.$$

However, when you can use computer (like when doing homework but not in exams), you can always use R to compute this type of probability.

Type	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
$X \sim B(n,p)$	**dbinom**(x, n, p)	**pbinom**(x, n, p)
$X \sim \text{Geom}(p)$	**dgeom**($x-1$, p)	**pgeom**($x-1$, p)
$X \sim \text{NB}(r,p)$	**dnbinom**($x-r$, r, p)	**pnbinom**($x-r$, r, p)

Example 3.4.2. At an automotive paint plant, 25 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."
- $p = P(\text{"success"}) = P(\text{not conforming}) = 0.25$.

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

$$\begin{aligned} X = &\ \text{the number of batches needed to find the } \mathbf{r}\text{th nonconforming} \\ \sim &\ \text{NB}(r, p = 0.25) \end{aligned}$$

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

(b) What is the probability that **no more than two** nonconforming batches will be observed among the first 4 batches sent to the lab?

(c) What is the probability that **no more than three** nonconforming batches will be observed among the first 30 batches sent to the lab?

3.5 Hypergeometric Distribution

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 (defective/not), new born babies (boy/girl), etc.

In the population of interest, we have

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

Randomly select n objects from the population (objects are selected **at random**, random means each remain object has the same chance of getting selected, and **without replacement**). Define

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

We say that X has a hypergeometric distribution and write $X \sim \text{hyper}(N, n, K)$. The probability function of $X \sim \text{hyper}(N, n, K)$ is

$$f(x) = \begin{cases} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, & x \le K \text{ and } n - x \le N - K \\ \\ 0, & \text{otherwise.} \end{cases}$$

Further, its mean and variance are

$$\mu = E(X) = n\left(\frac{K}{N}\right) \quad \text{and} \quad \sigma^2 = V(X) = n\left(\frac{K}{N}\right)\left(\frac{N-K}{N}\right)\left(\frac{N-n}{N-1}\right).$$

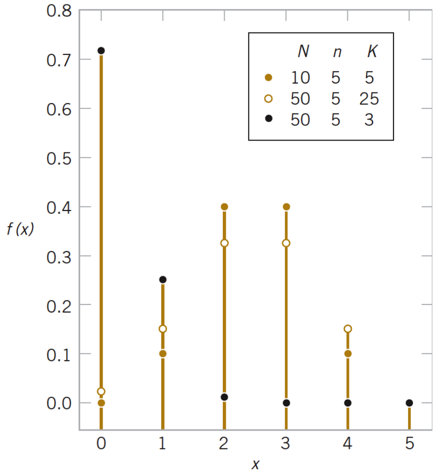


Figure 3.5.2: Hypergeometric distributions for selected values of N, K, and n.

Example 3.5.1. 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."

Suppose among the 100 parts there are 10 parts which are defective.

(a) What is the probability that the lot will be accepted?

(b) What is the probability that at least 3 of the 5 parts sampled are defective?

R codes for hypergeometric distribution:

Type	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
$X \sim \text{hyper}(N, n, K)$	**dhyper**$(x,\ K,\ N-K,\ n)$	**phyper**$(x,\ K,\ N-K,\ n)$

In the previous example, we could compute the probabilities of interest, using R, as follows:

```
> 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

The Poisson distribution is commonly used to model counts in an interval of time, an area, a volume or other unit, such as

1. the number of customers entering a post office in a given hour
2. the number of α-particles discharged from a radioactive substance in one second
3. the number of machine breakdowns per month
4. the number of insurance claims received per day
5. the number of defects on a piece of raw material.

In general, we define

$$X = \text{ the number of ``occurrences'' over a unit interval of time (or space).}$$

A Poisson distribution for X emerges if these "occurrences" obey the following rules:

(I) **the number of occurrences in non-overlapping intervals (of time or space) are independent random variables.**

(II) the probability of an occurrence in a sufficiently short interval is proportional to the length of the interval.

(III) The probability of 2 or more occurrences in a sufficiently short interval is zero.

We say that X has a Poisson distribution and write $X \sim$ Poisson(λ). A process that produces occurrences according to these rules is called a Poisson process.

If $X \sim$ Poisson(λ), then the probability mass function of X is given by

$$f(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \ldots \\ \\ 0, & \text{otherwise.} \end{cases}$$

and

$$E(X) = \lambda$$
$$V(X) = \lambda.$$

Remark: **In a Poisson process, suppose the mean of counts in one unit is λ, then the mean of counts in 2 units is 2λ, in 3 unites is 3λ, ...**

Example 3.6.1. Let X denote the number of times per month that a detectable amount of radioactive gas is recorded at a nuclear power plant. Suppose that X 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?

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

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

(d) Given the event that there are four times a detectable amount of gas is recorded in September, what is the probability that there are exactly three times a detectable amount of gas is recorded in October?

TI-84 codes for Poisson distribution:

Type	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
$X \sim$ Poisson(λ)	**poissonpdf**(λ, x)	**poissoncdf**(λ, x)

R codes for Poisson distribution:

Type	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
$X \sim$ Poisson(λ)	**dpois**(x, λ)	**ppois**(x, λ)

Example 3.6.2. Orders arrive at a Web site according to a Poisson process with a mean of 12 per hour. Determine the following:

(a) Probability of no orders in five minutes.

(b) Probability of 3 or more orders in five minutes.

(c) Length of a time interval such that the probability of no orders in an interval of this length is 0.001.

3.7 General Continuous Distribution

Recall: A **continuous random variable** is a random variable with an interval (either finite or infinite) of real numbers for its range.

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

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

If X = temperature (measured in degree ^oC), the set of all possible values of X (ignoring absolute zero and physical upper bounds) might be described as

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

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

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., $X = 3$, etc.) like we did with discrete random variables. Instead, we assign positive probability to events which are intervals (e.g., $2 < X < 4$, etc.).

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

1. $f(x) \geq 0$, that is, $f(x)$ is nonnegative.

2. The area under any pdf is equal to 1, that is,

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

3. If x_0 is a specific value of interest, then the **cumulative distribution function (cdf)** of X is given by

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx.$$

In another way, $f(x)$ can be view as the first derivative of $F(x)$; i.e.,

$$f(x) = F'(x).$$

4. If x_1 and x_2 are specific values of interest ($x_1 < x_2$), then

$$\begin{aligned} P(x_1 \leq X \leq x_2) &= \int_{x_1}^{x_2} f(x)dx \\ &= F(x_2) - F(x_1). \end{aligned}$$

5. For any specific value x_0, $P(X = x_0) = 0$. In other words, in continuous probability models, specific points are assigned zero probability (see #4 above and this will make perfect mathematical sense). An immediate consequence of this is that if X is continuous,

$$P(x_1 \leq X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2) = P(x_1 < X < x_2)$$

   and each is equal to

$$\int_{x_1}^{x_2} f(x)dx.$$

   This is not true if X has a discrete distribution because positive probability is assigned to specific values of X. Evaluating a pdf at a specific value x_0, that is, computing $f(x_0)$, does not give you a probability! This simply gives you the height of the pdf $f(x)$ at $x = x_0$.

6. The **expected value (or population mean)** of X is given by

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

   And the **population variance** of X is given by

$$\sigma^2 = V(X) = E(X^2) - \{E(X)\}^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

   The **population standard deviation** of X is given by the positive square root of the variance:

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

Let X be a continuous random variable with pdf $f(x)$. Suppose that g, g_1, g_2, ..., g_k are real-valued functions, and let c be any real constant.

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Further expectations satisfy the following (linearity) properties:

1. $E(c) = c$
2. $E[cg(X)] = cE[g(X)]$
3. $E[\sum_{j=1}^{k} g_j(X)] = \sum_{j=1}^{k} E[g_j(X)]$

For linear function $g(x) = ax + b$ where a, b are constants, we have

$$E[g(X)] = aE[X] + b \text{ and } V[aX + b] = a^2 V[X].$$

Example 3.7.1. Suppose that X has the pdf

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

(a) Find the cdf of X.

(b) Calculate $P(X < 0.3)$

(c) Calculate $P(X > 0.8)$

(d) Calculate $P(0.3 < X < 0.8)$

(e) Find the mean of X

(f) Find the standard deviation of X.

(g) If we define $Y = 3X$, find the cdf and pdf of Y. Further calculate the mean and variance of Y.

3.8 Exponential Distribution

The Exponential Distribution is commonly used to answer the following questions:

- How long do we need to wait before a customer enters a shop?
- How long will it take before a call center receives the next phone call?
- How long will a piece of machinery work without breaking down?

All these questions concern the time we need to wait before a given event occurs. We often model this waiting time by assuming it follows an exponential distribution.

> A random variable X is said to have an **exponential distribution** with parameter $\lambda > 0$ if its pdf is given by
>
> $$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$
>
> Shorthand notation is $X \sim \text{Exp}(\lambda)$. The parameter λ is called **rate** parameter.

Now, let us calculate the cumulative distribution function of $X \sim \text{Exp}(\lambda)$:

$$F(x_0) = P(X \le x_0) = \underline{\hspace{6cm}}$$

$$= \begin{cases} \underline{\hspace{13cm}} \\ \underline{\hspace{6cm}} \end{cases}$$

Thus, for any specified time x_0 (of course it is positive), the probability of the event happens no later than x_0 is

$$P(X \le x_0) = F(x_0) = \underline{\hspace{6cm}}$$

The probability of the event happens later than x_0 is

$$P(X > x_0) = 1 - F(x_0) = \underline{\hspace{6cm}}$$

Now, let us define $Y = \lambda X$. What is the cdf and pdf of Y? What are the mean and variance of Y?

Using Y, we are able to calculate the mean and variable of $X \sim \text{Exp}(\lambda)$ as

$$\mu = E(X) = \underline{\hspace{8cm}}$$

$$\sigma^2 = V(X) = \underline{\hspace{8cm}}$$

Consequently, the standard deviation of X is then $\sigma = \sqrt{\sigma^2} = \underline{\hspace{3cm}}$.

Example 3.8.1. Assume that the length of a phone call in minutes is an exponential random variable X with parameter $\lambda = 1/10$, (or the question may tell you the value of λ through the expectation; i.e., this based the expected waiting time for a phone call is 10 minuets). If someone arrives at a phone booth just before you arrive, find the probability that you will have to wait

(a) less than 5 minuets

(b) greater than 10 minuets

(c) between 5 and 10 minuets

Also compute the expected value and variance.

Memoryless property: Suppose Z is a continuous random variable whose values are all non-negative. We say Z is memoryless if for any $r \geq 0$, $s \geq 0$, we have

$$P(Z > t + s \mid Z > t) = P(Z > s).$$

INTERPRETATION: suppose Z represents the waiting time until something happens. This property says that, conditioning on that you have waited at least t time, the probability of waiting for additionally at leat s time is the same with the probability of waiting for at least s time starting from the beginning. In other words, In other words, the fact that Z has made it to time t has been "forgotten."

In the following, we show that any exponential random variable, $X \sim \text{Exp}(\lambda)$ has the memoryless property.

Example 3.8.2. In previous example, what is probability that you need wait for more than 10 minuets given the fact you have waited for more than 3 minutes?

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

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

Then,

$$W \sim \text{Exp}(\lambda).$$

NOTE THAT it is also true that the time between any two consecutive occurrences in a Poisson process follows this same exponential distribution (these are called "interarrival times").

Example 3.8.3. Suppose that customers arrive at a check-out according to a Poisson process with mean $\lambda = 12$ per hour. 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.)

3.9 Gamma Distribution

We start this subsection with a very interesting function: the Gamma Function, defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt, \text{ where } \alpha > 0.$$

When $\alpha > 1$,the gamma function satisfies the recursive relationship,

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

Therefore, if n is an integer, then

$$\Gamma(n) = (n-1)!$$

Notice that

$$\begin{aligned}
1 &= \int_0^\infty \frac{1}{\Gamma(\alpha)}t^{\alpha-1}e^{-t}dt \\
&\quad \text{(change variable } x = t/\lambda, \text{ for } \lambda > 0) \\
&= \int_0^\infty \frac{1}{\Gamma(\alpha)}(\lambda x)^{\alpha-1}e^{-\lambda x}d(\lambda x) \\
&= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}dx \\
&= \int_{-\infty}^\infty f(x)dx \quad \text{(Thus } f(x) \text{ is a valid pdf.)}
\end{aligned}$$

where

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

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

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

Shorthand notation is $X \sim \text{Gamma}(\alpha, \lambda)$. Its mean and variance are

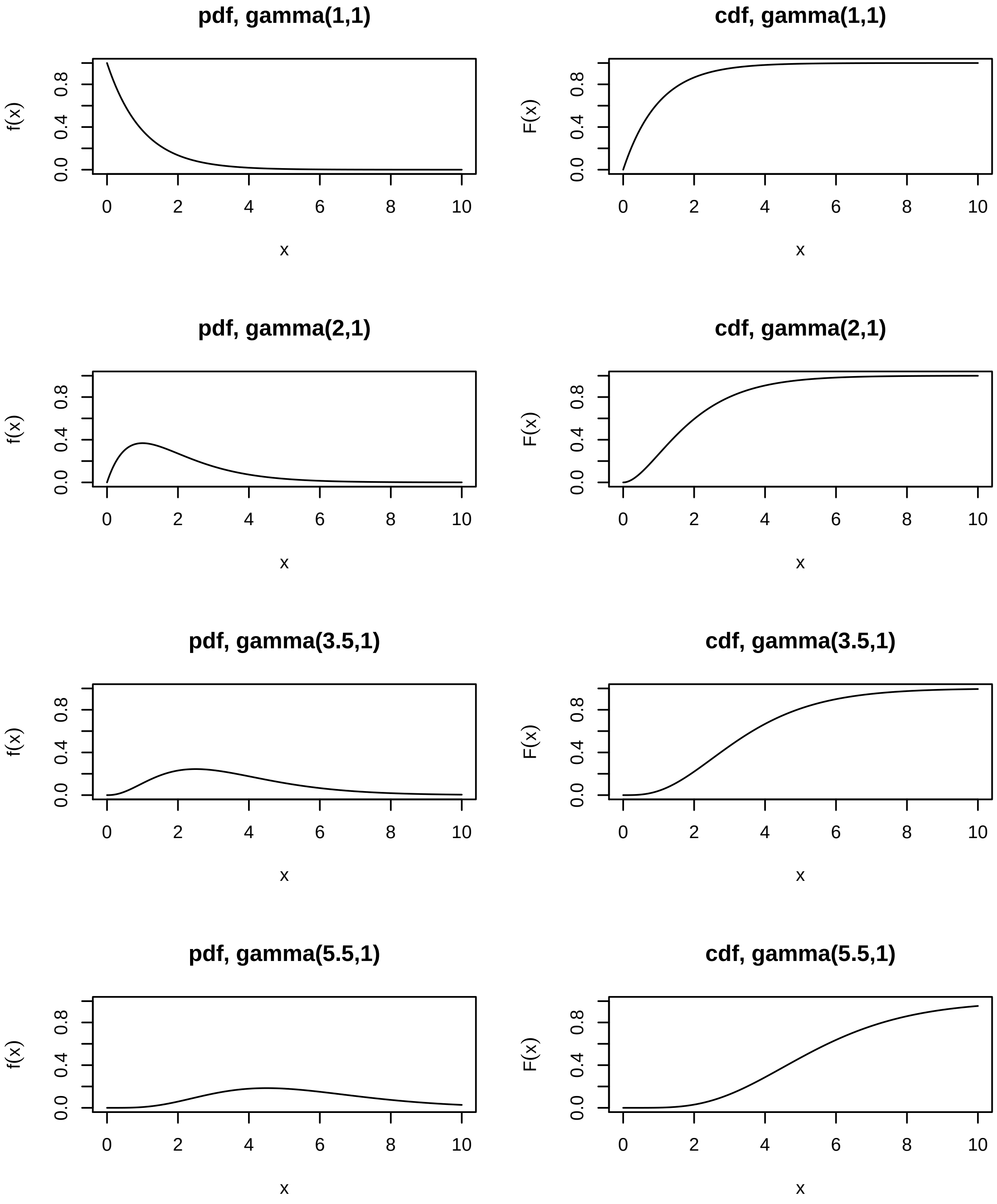
$$E(X) = \frac{\alpha}{\lambda}, V(X) = \frac{\alpha}{\lambda^2}.$$

- When $\alpha = 1$, we have

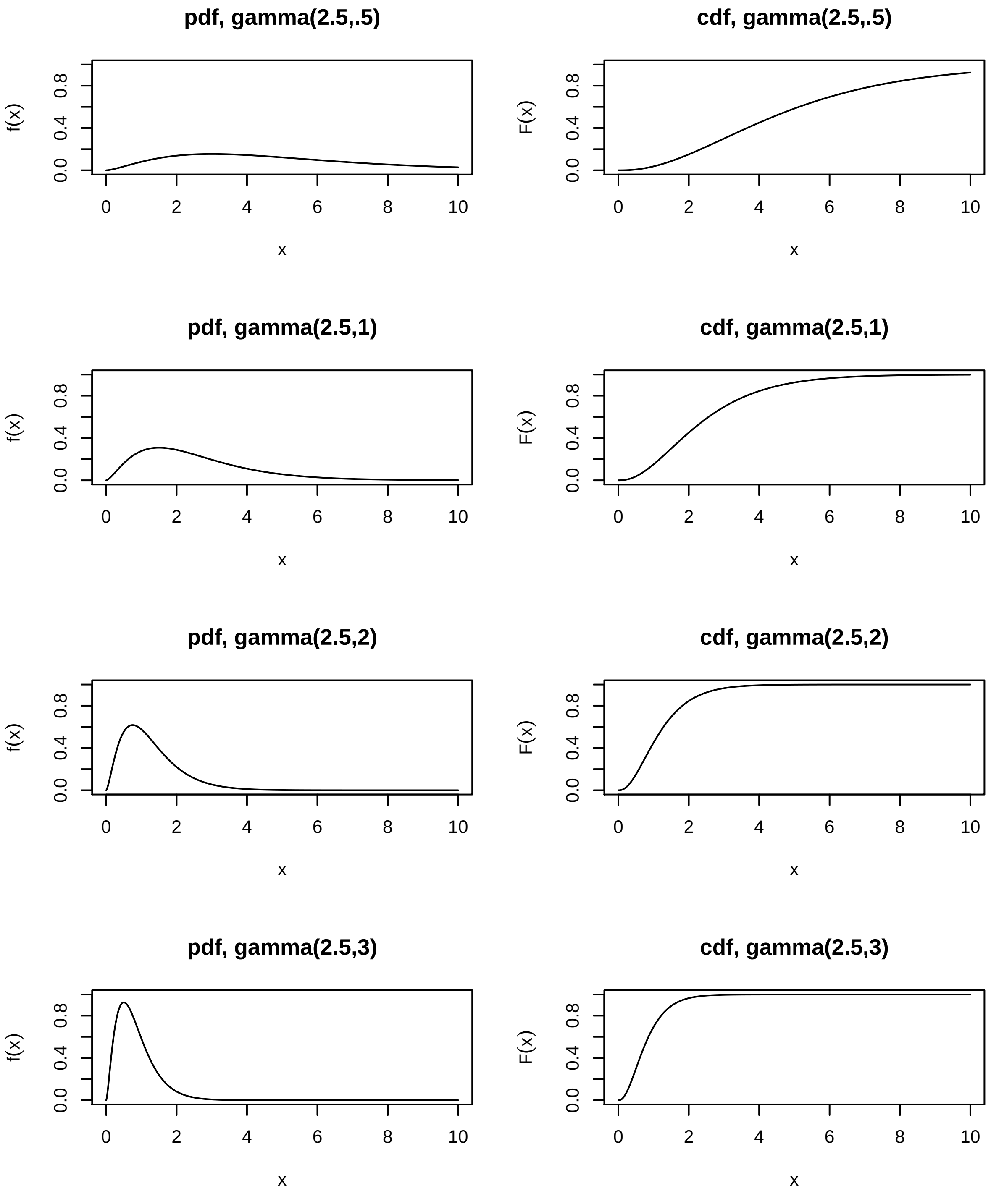
$$f(x) = \begin{cases} \frac{\lambda^1}{\Gamma(1)}x^{1-1}e^{-\lambda x} = \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Hence, exponential distribution $\text{Exp}(\lambda)$ is a special case of Gamma distribution; i.e., $\text{Gamma}(1, \lambda)$. In other words, the gamma distribution is more **flexible** than the exponential distribution.

- Plot of pdf and cdf:



- Plot of pdf and cdf:



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

  $$W = \text{ the time until the } \alpha\text{th occurrence (herein } \alpha \text{ is an integer).}$$

  Then,

  $$W \sim \text{Gamma}(\alpha, \lambda).$$

  NOTE THAT it is also true that the time between any two occurrences (unlike last subsection, these two occurrences does not need to be consecutive) in a Poisson process follows a gamma distribution.

- The **cdf** of a gamma random variable does not exist in closed form. Therefore, probabilities involving gamma random variables (when $\alpha \neq 1$) must be computed numerically (e.g., using R).

R codes for Exponential and Gamma distributions:

Type	$F(x) = P(X \leq x)$
$X \sim \text{Exp}(\lambda)$	**pexp**(x, λ) or **pgamma**(x, 1, λ)
$X \sim \text{Gamma}(\alpha, \lambda)$	**pgamma**(x, α, λ)

Example 3.9.1. Calls to the help line of a large computer distributor follow a Poisson distribution with a mean of 20 calls per minute. Determine the following:

(a) Mean time until the one-hundredth call

(b) Mean time between call numbers 50 and 80

(c) Probability that the time till the third call occur within 15 seconds

3.10 Normal Distribution

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

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

Shorthand notation is $X \sim N(\mu, \sigma^2)$. Another name for the normal distribution is the **Gaussian distribution**.

$$E(X) = \mu, V(X) = \sigma^2.$$

Example 3.10.1. If $X \sim N(1, 4)$, find the mean, variance, and standard deviation of X.

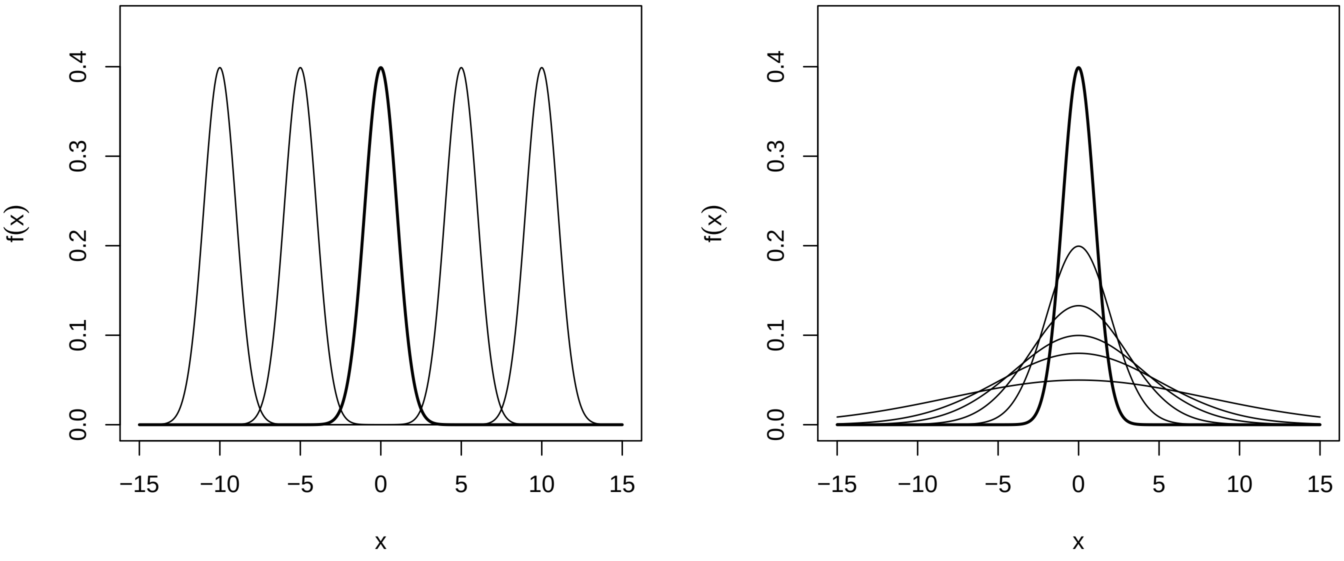


Figure 3.10.3: The left one presents the plot of pdf of $N(-10, 1)$, $N(-5, 1)$, $N(0, 1)$, $N(5, 1)$, $N(10, 1)$ (from left to right).
The right one presents the plot of pdf of $N(0, 1)$, $N(0, 2^2)$, $N(0, 3^2)$, $N(0, 4^2)$, $N(0, 5^5)$, $N(0, 8^8)$ (from top to down)

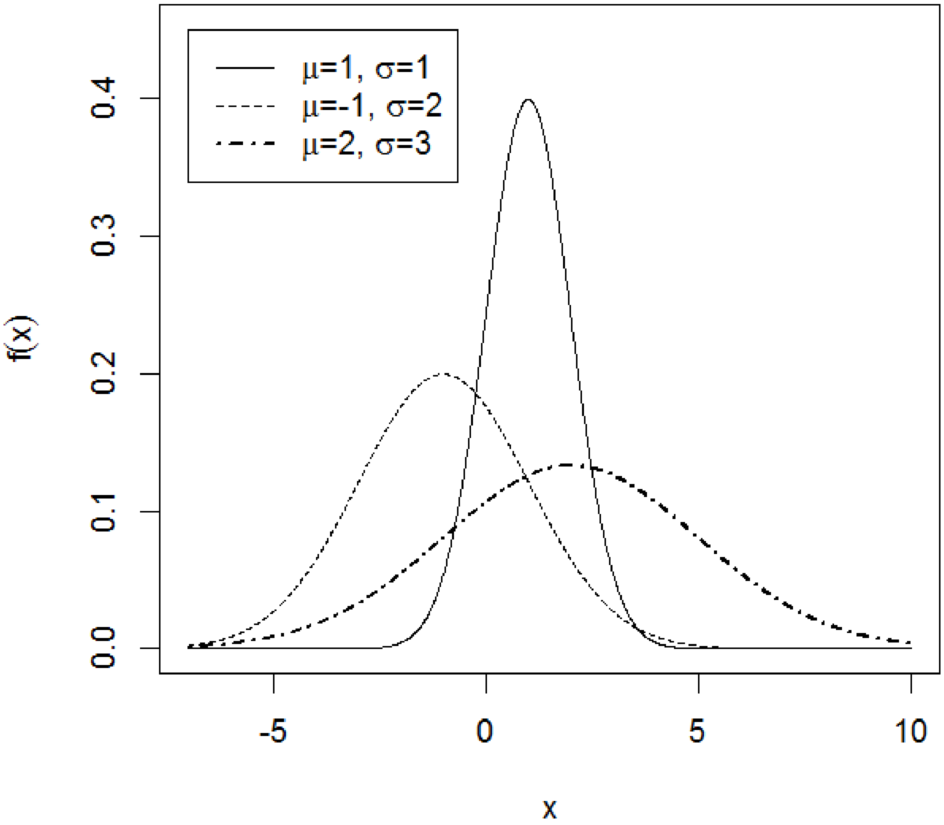
Example 3.10.2. Denote $Z = \frac{X-\mu}{\sigma}$, identify the distribution of Z.

Standard normal distribution: when $\mu = 0$ and $\sigma^2 = 1$, we say the normal distribution $N(0, 1)$ is the standard normal distribution. We denote a standard normal random variable by Z; i.e.,

$$Z \sim N(0, 1).$$

If random variable $X \sim N(\mu, \sigma^2)$, we can standardize X to get a standard normal random variable:

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



The cumulative function of a standard normal random variable is denoted as

$$\Phi(z) = P(Z \leq z).$$

However, the function $\Phi(z)$ does not exist in closed form. Actually, for any normal random variable, its cdf does not exists in closed form.

TI-84 codes for Normal distributions $N(\mu, \sigma^2)$:

Type	Commands (input σ not σ^2)
$P(a \leq X \leq b)$	**normalcdf** (a, b,μ,σ)
$P(a \leq X)$	**normalcdf** (a, 10^{99},μ,σ)
$P(X \leq b)$	**normalcdf** (-10^{99}, b,μ,σ)

For $X \sim N(\mu, \sigma^2)$,

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

In the other way, we can express

$$X = \mu + \sigma Z.$$

Thus, all the normal distribution shares a common thing which is the standard normal distribution. If we know standard normal, we know every normal distribution. For example, in terms of calculating probabilities, if $X \sim N(\mu, \sigma^2)$, we can always **standardize** it to get the standard normal Z and calculate the probabilities based on standard normal.

$P(x_1 < X < x_2) =$ ________________________

$=$ ________________________

$=$ ________________________

$=$ ________________________

Similarly, we have

$P(X > x_1) =$ ________________________, $P(X < x_2) =$ ________________________

Example 3.10.3. Find the following properties:

$Z \sim N(0,1)$	$X \sim N(1,4)$	$X \sim N(-1,9)$	
$P(-1 < Z < 1)$	$P(-1 < X < 3)$	$P(-4 < X < 2)$	
$P(-2 < Z < 2)$	$P(-3 < X < 5)$	$P(-7 < X < 5)$	
$P(-3 < Z < 3)$	$P(-5 < X < 7)$	$P(-10 < X < 8)$	

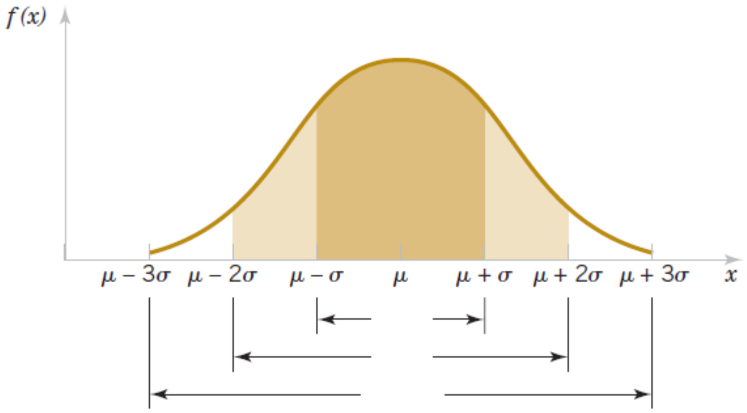
Three important things about normal distributions:

- **Empirical rule, or the 68-95-99.7% rule:** For $X \sim N(\mu, \sigma^2)$, calculate

  (a) $P(\mu - \sigma \le X \le \mu + \sigma) =$ ________________________________

  (b) $P(\mu - 2\sigma \le X \le \mu + 2\sigma) =$ ________________________________

  (c) $P(\mu - 3\sigma \le X \le \mu + 3\sigma) =$ ________________________________



Interpretation:

- about ______ of the distribution is between $\mu - \sigma$ and $\mu + \sigma$.
- about ______ of the distribution is between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- about ______ of the distribution is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

- **Symmetric:** The pdf of a normal distribution is always symmetric respect to its mean. Thus

$$P(Z > z) = P(Z < -z)$$
$$P(-z < Z < z) = 1 - 2P(Z > z) \text{ if } z > 0$$

For $X \sim N(\mu, \sigma^2)$,

$$P(X - \mu > x) = P(X - \mu < x)$$
$$P(-x < X - \mu < x) = 1 - 2P(X - \mu > x) \text{ if } x > 0.$$

- **Find the inverse of the cdf of a normal distribution:** We have already known how to compute $F(x) = P(X \leq x)$ when $X \sim N(\mu, \sigma^2)$. In the opposite way, suppose the question tells you $P(X \leq x) = \alpha$, how to find x based on the value of α?

TI-84 codes for the inverse of the cdf of $N(\mu, \sigma^2)$:

For any given $0 < \alpha < 1$,
the value of x, such that $\boldsymbol{P(X \leq x)} = \alpha$
can be found using the TI-84 code:
invNorm(α, μ, σ).

In the other way, if you need find the value of x such that $\boldsymbol{P(X > x)} = \alpha$, use

invNorm$(1 - \alpha, \mu, \sigma)$.

Example 3.10.4. If X is normally distributed with a mean of 10 and a standard deviation of 2.

(a) Find $P(2 < X < 8)$, $P(X > 10)$, $P(X < 9)$.

(b) Determine the value for x that solves each of the following:

(1) $P(X > x) = 0.5$

(2) $P(X > x) = 0.95$

(3) $P(x < X < 11) = 0.3$

(4) $P(-x < X - 10 < x) = 0.95$

(5) $P(-x < X - 10 < x) = 0.99$

Example 3.10.5. Suppose that the current measurements in a strip of wire are assumed to follow a normal distribution with a mean of 10 milliamperes and a variance of σ^2 (milliamperes)^2, where σ^2 is unknown.

(a) Suppose we know the probability that a measurement exceeds 12 milliamperes is 0.16, approximate σ^2 via the Empirical rule.

(b) Based on part (a), find the value x satisfies that the probability that a measurement exceeds x milliamperes is 0.05.

(c) Ignoring the findings in (a-b), suppose we know the probability that a measurement exceeds 13.29 milliamperes is 0.05, find σ.

3.11 Weibull Distribution

Reliability analysis is important in engineering. It deals with failure time (i.e., lifetime, time-to-event) data. For example,

- T = time from start of product service until failure
- T = time of sale of a product until a warranty claim
- T = number of hours in use/cycles until failure:

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 in practice, because reliability is related to product quality. There are many well known lifetime distributions, including

- exponential
- Weibull
- lognormal
- others: gamma, inverse Gaussian, Gompertz-Makeham, Birnbaum-Sanders, extreme value, log-logistic, etc.
- The normal (Gaussian) distribution is rarely used to model lifetime variables.

In this section, we will learn Weibull distribution.

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

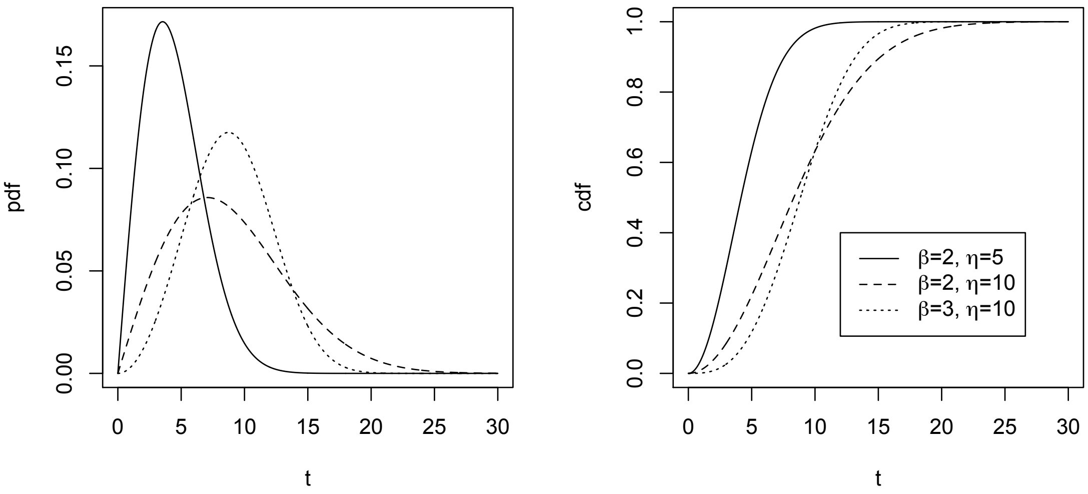
$$f_T(t) = \begin{cases} \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^{\beta}}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $T \sim \text{Weibull}(\beta, \eta)$.

- We call

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

- By changing the values of β and η, the Weibull pdf can assume many shapes. The Weibull distribution is very popular among engineers in reliability applications; e.g., here are the plots of probability density function and cumulative distribution function of several Weibull distributions.



- The cdf of T exists in closed form and is given by

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

- If $T \sim$ Weibull(β, η), then its mean and variance are

$$E(T) = \eta\Gamma\left(1 + \frac{1}{\beta}\right), \quad V(T) = \eta^2\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right\}.$$

- Note that when $\beta = 1$, the Weibull pdf reduces to the exponential$(\lambda = 1/\eta)$ pdf.

Example 3.11.1. Suppose that the lifetime of a rechargeable battery, 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 $t = 20$?

(c) What is the probability that a battery is still functional at time $t = 20$ given that the battery is functional at time $t = 10$?

(d) What is the value of t such that $P(T \leq t) = .99$?

3.12 Reliability functions

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

$$F_T(t) = \int_0^t f_T(u)du \quad \text{and} \quad S_T(t) = \int_t^\infty f_T(u)du.$$

The **hazard function** is defined as

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

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

The hazard function (or hazard rate) is a very important characteristic of a lifetime distribution. **It indicates the way the risk of failure varies with time**. Distributions with increasing hazard functions are seen in units for whom some kind of aging or "wear out" takes place. Certain types of units (e.g., electronic devices, etc.) may display a decreasing hazard function, at least in the early stages of their lifetimes. It is insightful to note that

$$
\begin{aligned}
h_T(t) &= \lim_{\epsilon \to 0} \frac{P(t \le T < t+\epsilon | T \ge t)}{\epsilon} = \lim_{\epsilon \to 0} \frac{P(t \le T < t+\epsilon)}{\epsilon P(T \ge t)} \\
&= \frac{1}{P(T \ge t)} \lim_{\epsilon \to 0} \frac{F_T(t+\epsilon) - F_T(t)}{\epsilon} = \frac{f_T(t)}{S_T(t)}.
\end{aligned}
$$

We can therefore describe the distribution of the continuous lifetime random variable T by using either $f_T(t)$, $F_T(t)$, $S_T(t)$, or $h_T(t)$.

Example 3.12.1. In this example, we find the hazard function for $T \sim \text{Weibull}(\beta, \eta)$. Recall that when $t > 0$, the pdf of T is

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

The cdf and survivor function of T are, respectively,

$$
F_T(t) = 1 - e^{-(t/\eta)^\beta} \quad \text{and} \quad S_T(t) = 1 - F_T(t) = e^{-(t/\eta)^\beta}.
$$

Therefore, the hazard function, for $t > 0$, is

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

Plots of Weibull hazard functions are given in Figure 3.12.4. It is easy to show

- $h_T(t)$ is increasing if $\beta > 1$ (wear out; population of units get weaker with aging)
- $h_T(t)$ is constant if $\beta = 1$ (constant hazard; exponential distribution)
- $h_T(t)$ is decreasing if $\beta < 1$ (infant mortality; population of units gets stronger with aging).

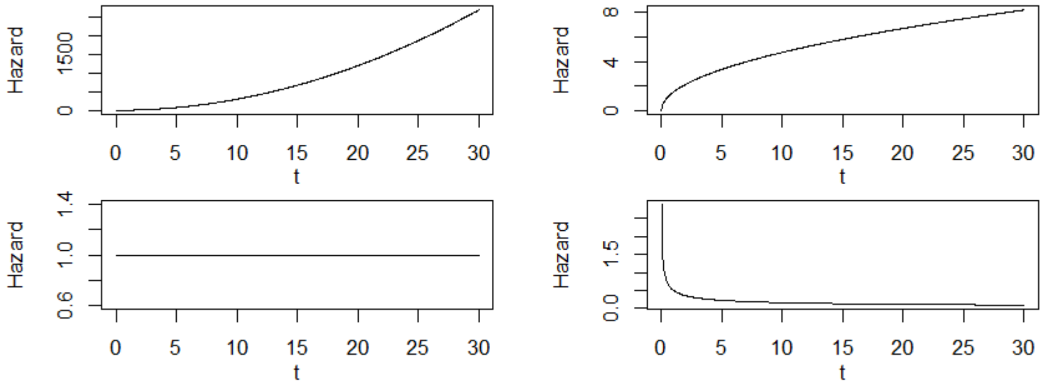


Figure 3.12.4: Weibull hazard functions with $\eta = 1$. Upper left: $\beta = 3$. Upper right: $\beta = 1.5$. Lower left: $\beta = 1$. Lower right: $\beta = 0.5$.

4 One-Sample Statistical Inference

4.1 Populations and samples

Overview: This chapter is about **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 4.1.1. Suppose that we wish to study the performance of lithium batteries used in a certain calculator. The purpose of our study is to determine the mean lifetime of these batteries so that we can place a limited warranty on them in the future. Since this type of battery has not been used in this calculator before, no one (except the Oracle) can tell us the distribution of X, the battery's lifetime. In fact, not only is the distribution not known, but all parameters which index this distribution aren't known either.

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 weight, 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 probabilistic 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. Our hope is that a random sample of individuals is "representative" of the entire population of individuals.

Notation: We will denote a random sample of observations by

$$X_1, X_2, ..., X_n.$$

That is, X_1 is the value of X for the first individual in the sample, X_2 is the value of X 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. Statisticians refer to the set of observations $X_1, X_2, ..., X_n$ generically as **data**. Lower case notation $x_1, x_2, ..., x_n$ is used when citing numerical values (or when referring to realizations of the upper case versions).

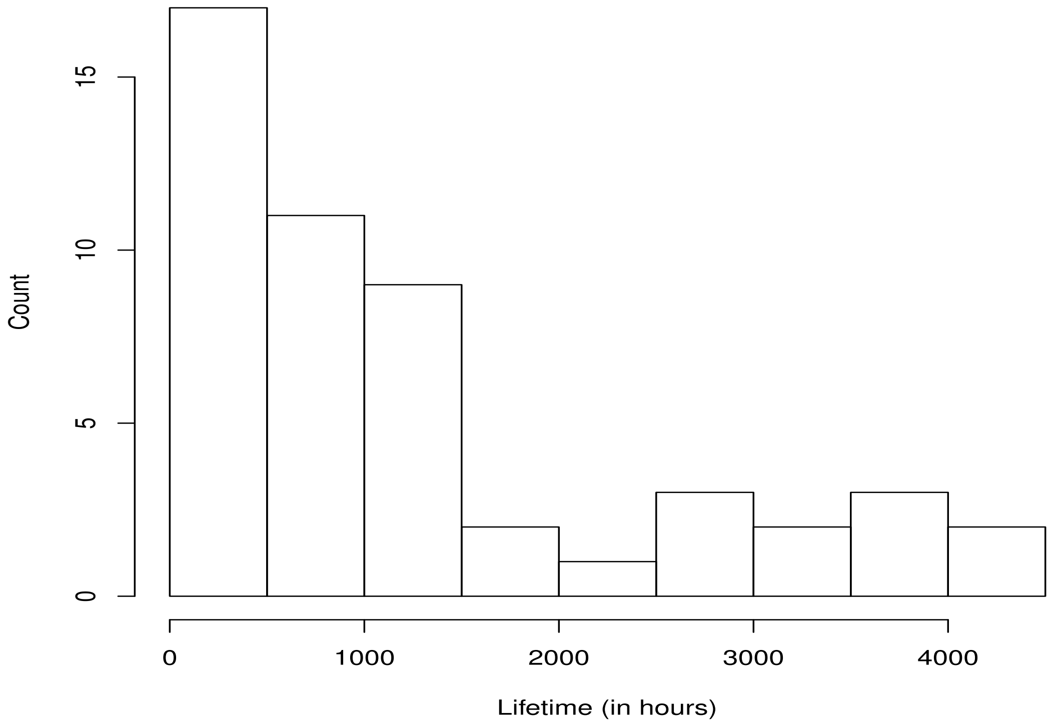


Figure 4.1.1: Histogram of battery lifetime data (measured in hours).

Example 4.1.2. *BATTERY DATA*: Consider the following random sample of $n = 50$ battery lifetimes $x_1, x_2, ..., x_{50}$ (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 4.1.1, we display a **histogram** 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 histogram?

- An exponential Exp(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 theoretical distribution which is used to model the entire population of battery lifetimes.

- In general, (population) parameters which index probability distributions (like the exponential) are unknown.

- All of the probability distributions that we discussed in Chapter 3 are meant to describe (model) population behavior.

4.2 Parameters and statistics

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

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

All of the probability distributions that we talked about in Chapter 3 were indexed by population (model) parameters. For example,

- the $N(\mu, \sigma^2)$ 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 two parameters, the size n and the population proportion of successes p.

Suppose that $X_1, X_2, ..., X_n$ is a random sample from a population. The **sample mean** is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The **sample variance** is

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

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

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

Important: These quantities can be computed from a sample of data $X_1, X_2, ..., X_n$.

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

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

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

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

In R, the following codes can help you calculate the sample mean, sample variance, and sample standard deviation. In the following codes, the battery data is saved in the variable with the name "battery."

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

Summary: The table below succinctly summarizes the salient 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.

These are both necessary to construct **confidence intervals** and to perform **hypothesis tests**, two important exercises discussed in this chapter.

4.3 Point estimators and sampling distributions

NOTATION: 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 (Chapter 6-7).

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

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

Important: It is important to note that, in general, an estimator $\widehat{\theta}$ is a statistic, so it depends on the sample of data $X_1, X_2, ..., X_n$.

- The data $X_1, X_2, ..., X_n$ come from the sampling process; e.g., different random samples will yield different data sets $X_1, X_2, ..., X_n$.
- In this light, because the sample values $X_1, X_2, ..., X_n$ will vary from sample to sample, the value of $\widehat{\theta}$ will too! It therefore makes sense to think about all possible values of $\widehat{\theta}$; that is, the **distribution** of $\widehat{\theta}$.

The distribution of an estimator $\widehat{\theta}$ (a statistic) is called its **sampling distribution**. A sampling distribution describes mathematically how $\widehat{\theta}$ would vary in repeated sampling. We will study many sampling distributions in this chapter.

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. Note that unbiasedness is a characteristic describing the center of a sampling distribution. This deals with **accuracy**.

RESULT: Mathematics shows that when $X_1, X_2, ..., X_n$ is a random sample,

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

That is, $\overline{X}$ 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).

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,

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

4.4 Sampling distributions involving $\overline{X}$

Sampling distribution of $\overline{X}$ from normal distribution: Suppose that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. The sample mean $\overline{X}$ has the following **sampling distribution**:

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

- This result reminds us that
$$E(\overline{X}) = \mu.$$
That is, the sample mean $\overline{X}$ is an **unbiased estimator** of the population mean μ.

- This result also shows that the **standard error** of $\overline{X}$ (as a point estimator) is
$$\text{se}(\overline{X}) = \sqrt{\text{var}(\overline{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Example 4.4.1. Suppose

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

We assume

$$X \sim N(\mu = 1.5, \sigma^2 = 0.16).$$

We call this the **population distribution**, because it describes the distribution of values of X for all individuals in the population (here, in-traffic drivers).

Question: suppose that we take a random sample of $n = 5$ drivers with times $X_1, X_2, ..., X_5$. What is the distribution of the sample mean $\overline{X}$?
Solution: If the sample size is $n = 5$, then with $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Longrightarrow \quad \overline{X} \sim N(1.5, 0.032).$$

This distribution describes the values of $\overline{X}$ 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 $\overline{X}$ each time.
Question: Suppose that we take a random sample of $n = 25$ drivers with times $X_1, X_2, ..., X_{25}$. What is the distribution of the sample mean $\overline{X}$?
Solution: If the sample size is $n = 25$, then with $\mu = 1.5$ and $\sigma^2 = 0.16$, we have

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Longrightarrow \quad \overline{X} \sim N(1.5, 0.0064).$$

Central Limit Theorem: Suppose that $X_1, X_2, ..., X_n$ is a random sample from a population distribution (does not have to be normal distribution) with mean μ and variance σ^2 (not necessarily a normal distribution). When the sample size n is large, we have

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

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

- **Sampling distribution of $\overline{X}$ from normal distribution** guarantees that when the underlying population distribution is $N(\mu, \sigma^2)$, the sample mean

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

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

- The central limit theorem demonstrates that, for the population mean μ, the point estimator $\overline{X}$ works consistently well in terms of that, $\text{se}(\overline{X}) = \sigma/\sqrt{n}$ converges to zero as sample size n increases (which is reasonable, since a larger sample size means that it contain more information about the population, thus the resulting estimator should be more accurate).

- **Waring:** The central limit theorem says no matter which distribution the samples are collected from, the sample mean follows approximately normal **when sample size is large**. However, this **does not** mean that when sample size is large, any distribution becomes normal.

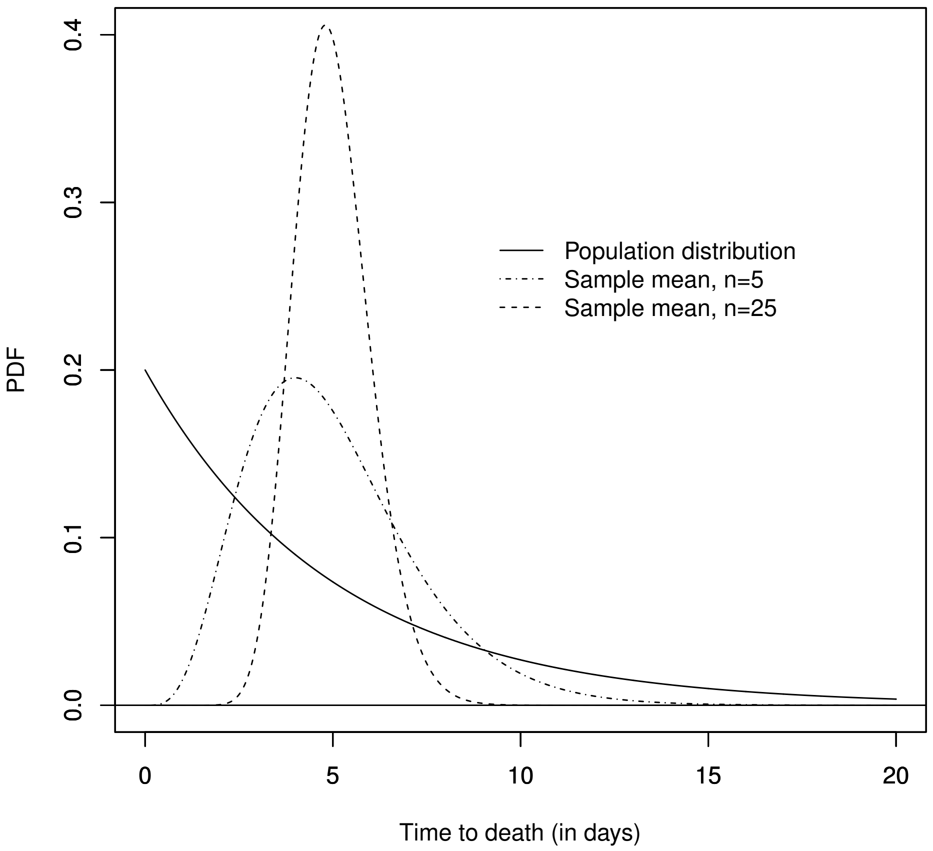


Figure 4.4.2: Rat death times. Population distribution: $X \sim \text{Exp}(\lambda = 1/5)$. Also depicted are the sampling distributions of $\overline{X}$ when $n = 5$ and $n = 25$.

Example 4.4.2. The time to death for rats injected with a toxic substance, denoted by X (measured in days), follows an exponential distribution with $\lambda = 1/5$. That is,

$$X \sim \text{Exp}(\lambda = 1/5).$$

This is the **population distribution**, that is, this distribution describes the time to death for all rats in the population.

- In Figure 4.4.2, I have shown the Exp(1/5) population distribution (solid curve). I have also depicted the theoretical sampling distributions of $\overline{X}$ when $n = 5$ and when $n = 25$.
- **Main point:** Notice how the sampling distribution of $\overline{X}$ begins to (albeit distantly) resemble a normal distribution when $n = 5$. When $n = 25$, the sampling distribution of $\overline{X}$ looks very much to be normal. This is precisely what is conferred by the CLT. The larger the sample size n, the better a normal distribution approximates the true sampling distribution of $\overline{X}$.

Example 4.4.3. When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch (measured in grams) is a random variable X with the following population parameters:

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

Suppose that $n = 50$ batches are prepared (independently). What is the probability that the sample mean impurity amount $\overline{X}$ is greater than 4.2 grams?
Solution: With $n = 50$, $\mu = 4$, and $\sigma^2 = (1.5)^2$, the CLT says that

$$\overline{X} \sim \mathcal{A}N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Longrightarrow \quad \overline{X} \sim \mathcal{A}N(4, 0.045).$$

Therefore,

$$P(\overline{X} > 4.2) \quad \approx \quad \texttt{normalcdf(4.2,}10^{99}\texttt{,4,}\sqrt{0.045}\texttt{)} = \texttt{0.1728893}.$$

Important: Note that in making this (approximate) probability calculation, we never made an assumption about the underlying population distribution shape.

4.5 Confidence intervals for a population mean μ

Before getting to confidence intervals, I need introduce a new definition:

Upper quantiles of a distribution: we say x is the **upper α-th quantile** of a distribution of random variable X, if

$$P(X > x) = \alpha.$$

(Lower) quantiles of a distribution: we say x is the **(lower) α-th quantile** of a distribution of random variable X, if

$$P(X \leq x) = \alpha.$$

Quantiles of the standard normal distribution. Recall in Section 3.10, I have introduced that, when $X \sim N(\mu, \sigma^2)$, Find x such that $P(X > x) = \alpha$ can use the commend "invNorm$(1 - \alpha, \mu, \sigma)$." Thus, for standard normal distribution; i.e., $Z \sim N(0, 1)$, we denote its **upper α-th quantile** as z_α which can be calculated as

$$z_\alpha = \text{invNorm}(1 - \alpha, 0, 1).$$

Based on the symmetric of standard normal distribution (with respect to 0), we have

$$\text{the lower } \alpha\text{-th quantile of the standard normal distribution} = -z_\alpha.$$

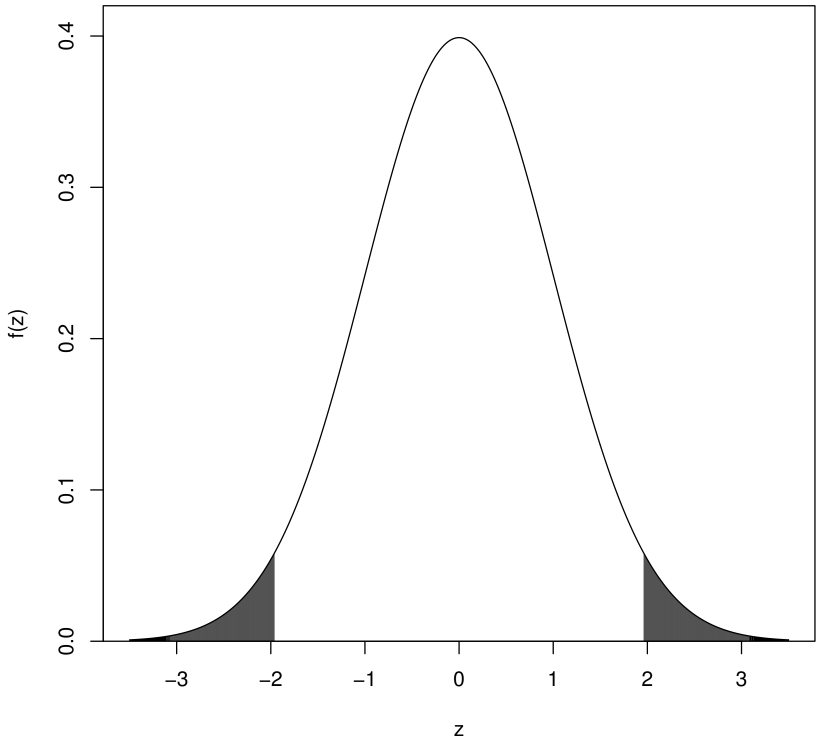


Figure 4.5.3: $N(0,1)$ pdf. The upper 0.025 and lower 0.025 areas have been shaded. The associated quantiles are the upper 0.025-th quantile as $z_{0.025} \approx 1.96$ and the lower 0.025-th quantile as $-z_{0.025} \approx -1.96$, respectively.

For example, if $\alpha = 0.05$ (see Figure 4.5.3), we know that the upper $\alpha/2$-th quantile and the lower $\alpha/2$-th quantile are

$$\begin{aligned} z_{0.05/2} = z_{0.025} = \text{invNorm}(1-0.025, 0, 1) &\approx 1.96 \\ -z_{0.05/2} = -z_{0.025} &\approx -1.96. \end{aligned}$$

Then, it is easy to see that,

$$1-\alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}).$$

It means that the probability of a standard random variable Z follows in the interval $(-z_{\alpha/2}, z_{\alpha/2})$ is $1-\alpha$. When $\alpha = 0.05$, we have the probability of a standard random variable Z follows in the interval $(-1.96, 1.96)$ is 95%.

Example 4.5.1. When $\alpha = 0.01$, find the upper and lower $\alpha/2$-th quantiles.

$$\begin{aligned} \text{upper: } z_{0.01/2} = z_{0.005} = \text{invNorm}(1-0.005, 0, 1) &\approx 2.576 \\ \text{lower: } -z_{0.01/2} = -z_{0.005} &\approx -2.576. \end{aligned}$$

4.5.1 Known population variance σ^2

To get things started, we will assume that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ population distribution. We will assume that

- the population variance σ^2 is **known** (largely unrealistic).
- the goal is to estimate the population mean μ.

We already know that $\overline{X}$ is an **unbiased** (point) **estimator** for μ, that is,

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

However, reporting $\overline{X}$ alone does not acknowledge that there is variability attached to this estimator. For example, in Example 4.4.3, for with the $n = 50$ measured amount of impurity, reporting

$$\overline{x} \approx 4.099 \text{ g}$$

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

- the 50 batches measured were drawn randomly from a population of all pipes, and
- different samples would give different sets of pipes (and different values of $\overline{x}$).

In other words, using a point estimator only **ignores important information**; namely, how variable the population of the amount of impurity in a batch is.

To avoid this problem (i.e., to account for the uncertainty in the sampling procedure), we therefore pursue the topic of **interval estimation** (also known as **confidence intervals**). The main difference between a point estimate 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."

We start our discussion by revisiting the sampling distribution of $\overline{X}$ from normal distribution in the last subsection. Recall that if $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, then the sampling distribution of $\overline{X}$ is

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

and therefore

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

Now we start building a confidence interval estimator of μ. In general, for any value of α, $0 < \alpha < 1$,

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu-\overline{X} > -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu > \overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

We call

$$\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

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

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- Note the form of the interval:

$$\underbrace{\text{point estimate}}_{\overline{X}} \pm \underbrace{\text{quantile}}_{z_{\alpha/2}} \times \underbrace{\text{standard error}}_{\sigma/\sqrt{n}}.$$

  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" in confidence interval 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.

- That is, "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 ($\alpha = 0.10$) $\Longrightarrow$ $z_{0.05} \approx 1.645$
  - 95 percent ($\alpha = 0.05$) $\Longrightarrow$ $z_{0.025} \approx 1.96$
  - 99 percent ($\alpha = 0.01$) $\Longrightarrow$ $z_{0.005} \approx 2.58$.

- The **length** of the $100(1-\alpha)$ percent confidence interval

  $$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  is equal to

  $$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

  Therefore,

  - the larger the sample size n, the smaller the interval length.
  - the larger the population variance σ^2, the larger the interval length.
  - the larger the confidence level $100(1-\alpha)$, the larger the interval length.

  Clearly, shorter confidence intervals are preferred. They are more informative!

Example 4.5.2. Civil engineers have found that the ability to see and read a sign at night depends in part on its "surround luminance;" i.e., the light intensity near the sign. The data below are $n = 30$ measurements of the random variable Y, the surround luminance (in candela per m^2). The 30 measurements constitute a random sample from all signs in a large metropolitan area.

10.9	1.7	9.5	2.9	9.1	3.2	9.1	7.4	13.3	13.1
6.6	13.7	1.5	6.3	7.4	9.9	13.6	17.3	3.6	4.9
13.1	7.8	10.3	10.3	9.6	5.7	2.6	15.1	2.9	16.2

Based on past experience, the engineers assume a normal population distribution (for the population of all signs) with known population variance $\sigma^2 = 20$.
QUESTION. Find a 90 percent confidence interval for μ, the mean surround luminance.
SOLUTION. We first use TI-84 to calculate the sample mean $\overline{x}$. Put numbers in a list L_1, then call "1-Var Stats." It gives us $\overline{x} = 8.62$. For a 90 percent confidence level; i.e., with $\alpha = 0.10$, we use

$$z_{0.10/2} = z_{0.05} \approx 1.645.$$

This can be determined from TI-84: invNorm(1-.05,0,1)$\approx$ 1.645. With $n = 30$ and $\sigma^2 = 20$, a 90 percent confidence interval for the mean surround luminance μ is

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \Longrightarrow \quad 8.62 \pm 1.645\left(\frac{\sqrt{20}}{\sqrt{30}}\right) \quad \Longrightarrow \quad (7.28, 9.96) \text{ candela/m}^2.$$

Interpretation (no credits if no interpretation): We are 90 percent confident that the mean surround luminance μ for all signs in the population is between 7.28 and 9.96 candela/m^2.

One-sided confidence bounds: A $100(1-\alpha)\%$ **upper-confidence bound** for μ is

$$\mu \leq \overline{x} + z_\alpha \sigma/\sqrt{n}$$

and a $100(1-\alpha)\%$ **lower-confidence bound** for μ is

$$\overline{x} - z_\alpha \sigma/\sqrt{n} \leq \mu$$

Example 4.5.3. ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at 60^oC are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3. Assume that impact energy is normally distributed with $\sigma = 1$J.

QUESTION. Construct a lower, one-sided 95% confidence interval for the mean impact energy.

SOLUTION. The interval is

$$\overline{x} - z_\alpha \sigma/\sqrt{n} \leq \mu \Rightarrow 64.46 - 1.64\frac{1}{\sqrt{10}} \leq \mu \Rightarrow 63.94 \leq \mu.$$

Practical Interpretation: The lower limit for the two-sided 95% confidence interval in last example was 63.84. Because $z_\alpha < z_{\alpha/2}$, the lower limit of a one-sided interval is always greater than the lower limit of a two-sided interval of equal confidence. The one-sided interval does not bound μ from above so that it still achieves 95% confidence with a slightly larger lower limit. If our interest is only in the lower limit for μ, then the one-sided interval is preferred because it provides equal confidence with a greater limit. Similarly, a one-sided upper limit is always less than a two-sided upper limit of equal confidence.

TI-84 can help you calculate confident intervals.
Make sure you know it.

Note that, we assume that we know the population distribution is normal. And based on normality, we have all the confidence intervals derived. However, what if we do not know that? Is this a serious cause for concern? Probably not. Recall that even if the population distribution (here, the distribution of all light intensity measurements in the city) is not perfectly normal, we still have

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

for n large, by the Central Limit Theorem. Therefore, our confidence interval is still approximately valid. A sample of size $n = 30$ is "pretty large." In other words, at $n = 30$, the CLT approximation above is usually "kicking in" rather well unless the underlying population distribution is grossly skewed (and I mean very grossly).

4.5.2 Sample size determination

MOTIVATION: In the planning stages of an experiment or investigation, it is often of interest to 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 for a population mean μ so that the interval length is no more than 5 units (e.g., days, inches, dollars, etc.). 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.

Suppose that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ population, where σ^2 is known. In this (known σ^2) situation, recall that a $100(1-\alpha)$ percent confidence interval for μ is given by

$$\overline{X} \pm \underbrace{z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)}_{=E,\text{ say}}.$$

The quantity E is called the **margin of error**.

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 (or an educated guess at its value; e.g., from past information, etc.)
- the confidence level, $100(1-\alpha)$
- the margin of error, E.

This is true because

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \iff n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

Choice of Sample Size: If $\overline{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\overline{x} - \mu|$ will not exceed a specified amount E when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Note that n must be an integer. Herein, always round your calculation up.

Example 4.5.4. 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 $E = 2$ 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 $z_{0.05/2} = z_{0.025} \approx 1.96$, $E = 2$, and $\sigma = 8$, the desired sample size to estimate μ is

$$n = \left(\frac{z_{\alpha/2}\sigma}{B}\right)^2 = \left(\frac{1.96 \times 8}{2}\right)^2 \approx 61.46 \approx 62.$$

4.5.3 Unknown population variance σ^2

In last section, we assume that we know the value of the population variance, but what if we do not know (which is a more common)?

What if σ^2 is unknown: Suppose that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. Result 1 says the sample mean $\overline{X}$ has the following sampling distribution:

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

If we standardize $\overline{X}$, we obtain

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

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

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

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

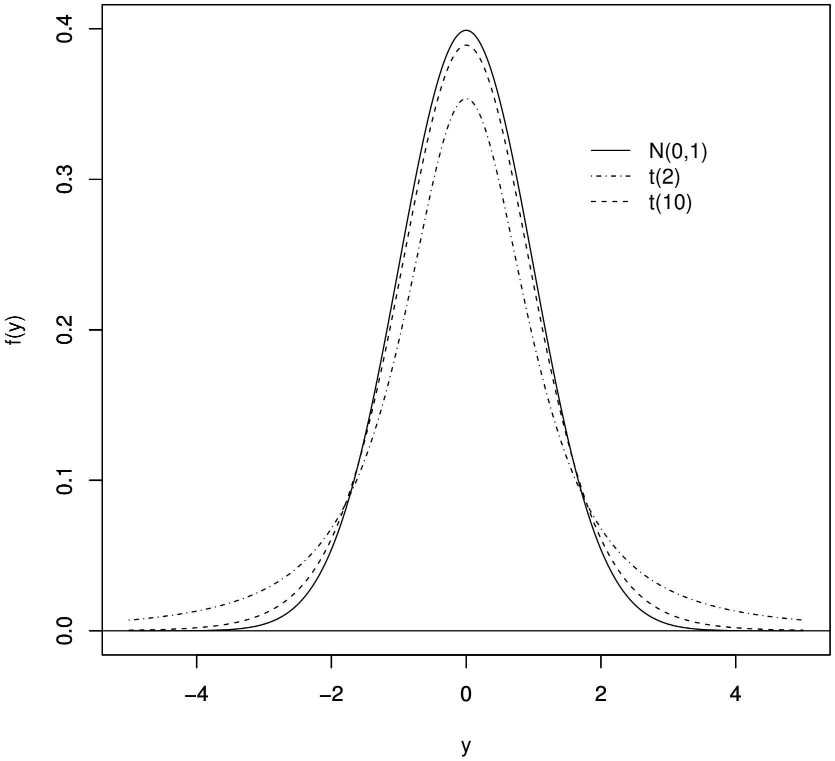


Figure 4.5.4: Probability density functions of $N(0, 1)$, $t(2)$, and $t(10)$.

The t distribution has the following characteristics:

- It is continuous and symmetric about 0 (just like the standard normal distribution).
- 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 $\nu \to \infty$, $t(\nu) \to N(0,1)$; thus, when ν becomes larger, the $t(\nu)$ and the $N(0,1)$ distributions look more alike.
- When compared to the standard normal distribution, the t distribution, in general, is less peaked and has more probability (area) in the tails.
- The t pdf formula is complicated and is unnecessary for our purposes. R will compute t probabilities and quantiles from the t distribution.

We continue to assume that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ population distribution.

- Our goal is the same; namely, to write a $100(1-\alpha)$ percent confidence interval for the population mean μ.
- However, we will no longer make the (rather unrealistic) assumption that population variance σ^2 is known.

Recall: If you look back in the notes at the "known σ^2 case," you will see that to derive a $100(1-\alpha)$ percent confidence interval for μ, we started with the following distributional result:

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

This led us to the following confidence interval formula:

$$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

The obvious problem is that, because σ^2 is now unknown, we can not calculate the interval. Not to worry; we just need a different starting point. Recall that

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

where S is the sample standard deviation (a point estimator for the population standard deviation). This result is all we need; in fact, it is straightforward to reproduce the "known σ^2" derivation and tailor it to this (now more realistic) case. A $100(1-\alpha)$ **percent confidence interval** for μ is given by

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

The symbol $t_{n-1,\alpha/2}$ denotes the **upper** $\alpha/2$ **quantile** from a t distribution with $\nu = n-1$ degrees of freedom.

- We see that the interval again has the same form:

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

  We interpret the interval in the same way.

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

TI-84 can help you calculate confident intervals.
Make sure you know it.

Example 4.5.5. 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 X (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 below:

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).

Question. Find a 99 percent confidence interval for μ, the mean level of airborne cadmium.

Solution. Input this dataset into Calculator List L_1. Then press STAT, go to TESTS, and choose TINTERVAL. Input, choose DATA, List is L_1, Freq: 1, C-level: 0.99 (since the question is asking for a 99% confidence interval). Press CALCULATE. The results are

$$\begin{aligned} (.04417, .0544) &\text{ This is your 99\% confidence interval} \\ \overline{x} = .492857143 &\text{ (sample mean } \overline{x}) \\ Sx = .0110893999 &\text{ (sample standard deviation } s) \\ n = 35 &\text{ (sample size } n) \end{aligned}$$

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

4.6 Confidence interval for a population proportion p

We now switch gears and focus on a new parameter: the **population proportion** p. This parameter emerges 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," denoted by p, $0 < p < 1$, 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.

Point estimator: A natural point estimator for p, the **population proportion**, is

$$\widehat{p} = \underline{\hspace{6cm}}$$

the **sample proportion**. This statistic is simply the proportion of "successes" in the sample (out of n individuals).

Fairly simple arguments can be used to show the following results:

$$
\begin{aligned}
E(\widehat{p}) &= \underline{\hspace{6cm}} \\
\text{se}(\widehat{p}) &= \underline{\hspace{6cm}}
\end{aligned}
$$

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.

Sample Distribution: 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 CLT) which says that

$$\widehat{p} \sim \mathcal{AN}\left[p,\ \frac{p(1-p)}{n}\right],$$

when the sample size n is large.

Result: An approximate $100(1-\alpha)$ **percent confidence interval** for p is given by

$$\underline{\hspace{9cm}}$$

TI-84 codes: "1-PropZInt." Three inputs:
x: ______________________________;
n: ______________________________;
C-level: _________________________.

- This interval should be used only when the sample size n is "large." A common **rule of thumb** (to use this interval formula) 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 approximate the true sampling distribution of $\widehat{p}$, thereby making the confidence interval formula above approximately valid.

- Note again 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."

- The value $z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the $N(0,1)$ distribution.

Example 4.6.1. One source of water pollution is gasoline leakage from underground storage tanks. In Pennsylvania, a random sample of $n = 74$ gasoline stations is selected 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.

Question: Suppose that 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?**

Sample size determination: To determine the necessary sample size, we first need to specify two pieces of information:

- the confidence level $100(1-\alpha)$
- the margin of error:

$$E = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

A small problem arises. Note that 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

$$E = z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}.$$

Solving this equation for n, we get

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 p_0(1-p_0).$$

This is the desired sample size n to find a $100(1-\alpha)$ percent confidence interval for p with a prescribed margin of error (roughly) equal to E. I say "roughly," because there may be additional uncertainty arising from our use of p_0 (our best guess).

Conservative approach: 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

$$n = n(p_0) = \left(\frac{z_{\alpha/2}}{E}\right)^2 p_0(1 - p_0),$$

when viewed as a function of p_0, is maximized when $p_0 = 0.5$; i.e.,

$n =$ ______________________________

Example 4.6.2. 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.

Question. 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 $E = 0.02$. 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.

4.7 Confidence interval for a population variance σ^2

MOTIVATION: In many situations, one is concerned not with the mean of an underlying (continuous) population distribution, but with the variance σ^2 instead. If 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. In a laboratory setting, chemical 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.

Suppose that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. The quantity

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

a χ^2 **distribution** with $\nu = n - 1$ degrees of freedom.

The χ^2 distribution has the following characteristics:

- It is continuous, skewed to the right, and always positive.
- 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.

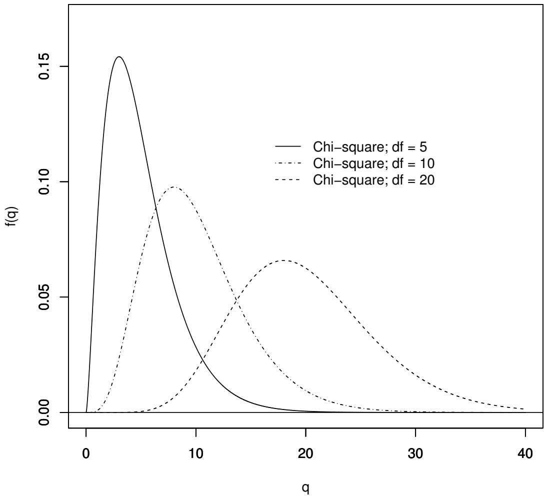


Figure 4.7.5: χ^2 probability density functions with different degrees of freedom.

Goal: Suppose that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution. We would like to write a $100(1-\alpha)$ percent confidence interval for σ^2.

Notation: Let $\chi^2_{n-1,\alpha/2}$ denote the **upper** $\alpha/2$ **quantile** and let $\chi^2_{n-1,1-\alpha/2}$ denote the **lower** $\alpha/2$ **quantile** of the $\chi^2(n-1)$ distribution; i.e., $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ satisfy

$$\begin{aligned} P(Q > \chi^2_{n-1,\alpha/2}) &= \alpha/2 \\ P(Q < \chi^2_{n-1,1-\alpha/2}) &= \alpha/2, \end{aligned}$$

respectively. Note that, unlike the $N(0,1)$ and t distributions, the χ^2 distribution is **not symmetric**. Therefore, different notation is needed to identify the quantiles of χ^2 distributions (this is nothing to get worried about). **Use the chi-square table to find these values. (Unfortunately, TI-84 cannot help you in this case).**

Example 4.7.1. Using the chi-square table, find the following values: $\chi^2_{n-1,\alpha/2}$, and $\chi^2_{n-1,1-\alpha/2}$, for all combination of $n = 10, 20, 30$ and $\alpha = 0.1, 0.05, 0.01$.

DERIVATION: Because $Q \sim \chi^2(n-1)$, we write

$$\begin{aligned} 1-\alpha &= P(\chi^2_{n-1,1-\alpha/2} < Q < \chi^2_{n-1,\alpha/2}) \\ &= P\left[\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right] \\ &= P\left[\frac{1}{\chi^2_{n-1,1-\alpha/2}} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{n-1,\alpha/2}}\right] \\ &= P\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right] \\ &= P\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]. \end{aligned}$$

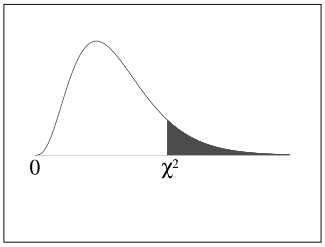
This argument shows that

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

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."

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Important: 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,

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}\right)$$

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.).

Major warning: Unlike the z and t confidence intervals for a population mean μ, the χ^2 interval for σ^2 (and for σ) **is not robust** to departures from normality. 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. Therefore, it is very important to "check" the normality assumption with these interval procedures (e.g., use a qq-plot which will be introduced later).

Example 4.7.2. Indoor swimming pools are noted for their poor acoustical properties. Suppose your goal is to design a pool in such a way that

- the population mean time that it takes for a low-frequency sound to die out is $\mu = 1.3$ seconds
- the population standard deviation for the distribution of die-out times is $\sigma = 0.6$ seconds.

Computer simulations of a preliminary design are conducted to see whether these standards are being met; here are data from $n = 20$ independently-run simulations. The data are obtained on the time (in seconds) it takes for the low-frequency sound to die out.

1.34	2.56	1.28	2.25	1.84	2.35	0.77	1.84	1.80	2.44
0.86	1.29	0.12	1.87	0.71	2.08	0.71	0.30	0.54	1.48

Question. Find a 95 percent confidence interval for the population standard deviation of times σ. What does this interval suggest about whether the preliminary design conforms to specifications (with respect to variability)?

4.8 Statistical Hypotheses

In the previous sections we have discussed how one could use data, observed from an experiment, to estimate unknown parameters. Often, we will also wish to use this data to make a decision about a statement about the parameters; e.g.,

- The mean number of car accidents that occur in Atlanta during rush hour on a given day is greater than 3.
- The mean number of car accidents that occur in Atlanta during rush hour on a given day is less than 5.
- The proportion of people who voted Democrat in the last election is greater than the proportion of people who voted Republican.

In all of these situations, we are forming a hypothesis about the structure of the underlying population(s) and we will develop data driven techniques that will allow us to decide whether or not our hypothesis is reasonable.

A **statistical hypothesis** is a statement about the parameters of one or more populations.

Example 4.8.1. Suppose that we are interested in the burning rate of solid propellant used to power aircrew escape systems; burning rate is a random variable that can be described by a probability distribution. Suppose that we are interested in the mean burning rate (a parameter of this distribution). Specifically, we are interested in whether or not the mean burning rate is $50cm/s$. This can be formally expressed as

$$\begin{aligned} H_0 &: \mu = 50cm/s \\ H_a &: \mu \neq 50cm/s. \end{aligned}$$

The statement $H_0 : \mu = 50cm/s$ is referred to as the **null hypothesis**, and the statement $H_a : \mu = 50cm/s$ is called the **alternative hypothesis**. The alternative hypothesis, as expressed above, is referred to as a two-sided alternative hypothesis since it specifies values of μ both greater and less than $50cm/s$. In some situations we may wish to formulate a one-sided alternative hypothesis; e.g.,

$$\begin{aligned} H_0 &: \mu = 50cm/s \\ H_a &: \mu < 50cm/s. \end{aligned}$$

or

$$\begin{aligned} H_0 &: \mu = 50cm/s \\ H_a &: \mu > 50cm/s. \end{aligned}$$

In general, we will collect data and summarize it in an effort to assess the validity of either H_0 or H_a. For example, suppose that a sample of n specimens are tested an that the mean burning rate $\overline{x}$ is observed. The sample mean is an estimate of the population mean μ. A value of $\overline{x}$ that is "reasonable" with respect to $H_0(H_a)$ tends to suggest that the null(alternative) hypothesis might be favorable. So what is reasonable? Consider defining a region such that if $\overline{x}$ falls in this region we will decide to reject H_0 in favor of H_a, we refer to this region as the critical region, and subsequently refer to the boundaries of this region as the critical values. For example, we might decide to reject $H_0 : \mu = 50cm/s$ if $\overline{x} < 48$ or $\overline{x} > 52$, consequently the values of $\overline{x}$ that are less than 48 and greater than 52 are the critical region and the critical values are 48 and 52.

This decision process can naturally lead one to make either one of the two following types of mistakes:

- Rejecting the null hypothesis, H_0, when it is in fact true (**Type I error**).
- Failing to reject the null hypothesis, H_0, when the alternative hypothesis, H1, is in fact true (**Type II error**).

Since the decision making process is based on random variables, probabilities can be associated with the Type I and II errors. Typically the probability of making a type I error is denoted by α, i.e,

$$\alpha = P(\text{type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true}) :$$

The type I error is often referred to as the **significance level** or size of the test. Likewise, we can quantify the probability of making a type II error as

$$1 - \beta = P(\text{type II error}) = P(\text{Fail to reject } H_0 \text{ when } H_0 \text{ is false}).$$

where $\beta = P(\text{reject } H_0 \text{ when } H_0 \text{ is false})$ is often called as **power** of the test.

Example 4.8.2. The FDA's policy is that the pharmaceutical company must provide substantial evidence that a new drug is safe prior to receiving FDA approval, so that the FDA can confidently certify the safety of the drug to potential customers. The null hypothesis is typically make by the status quo. For any new drug, FDA always assume it is unsafe. For testing the new drug, the null hypothesis is: "the new drug is unsafe"; and the alternative hypothesis is "the new drug is safe."

(a) Given the choice of null and alternative hypothesis, describe type I and II errors in terms of this application.

(b) If the FDA wants to very confident that the drug is safe before premitting to be marketed, is it more important that α or $1 - \beta$ be small? Explain.

A few important points with regard to type I and II errors:

- Type I and II errors are related. A decrease in the probability of a type I error always results in an increase in the probability of a type II error, and vice versa, provided the sample size n does not change.
- An increase in the sample size will generally reduce both α and $1-\beta$.
- When the null hypothesis is false, β decreases as the true value of the parameter approaches the value hypothesized in the null hypothesis. The value of β increases as the difference between the true parameter and the hypothesized value increases.

Typically, when designing a hypothesis test one controls (or specifies) the type I error probability. It is common to specify values of 0.10, 0.05, and 0.01 for α, but these values may not be appropriate in all situations.

There are typically three ways to test hypotheses: the confidence interval approach, the critical value approach, and the P-value approach.

4.9 Testing hypotheses on the mean μ with known variance σ^2

In this case, the population variance σ^2 is given, like in Section 4.5.1. Suppose $X_1, \ldots, X_n$ is a random sample of size n from a population. The population is normally distributed, (if it is not, the conditions of the central limit theorem apply, and n is sufficiently large.). Then, we have the sample mean $\overline{X}$ as a point estimator for μ, and its sampling distribution as

$$\overline{X} \sim N(\mu, \sigma^2/n).$$

To test hypotheses on the mean μ, we use this point estimator and its sample distribution.

In this course, the null hypothesis is always

$$H_0 : \mu = \mu_0$$

But, we have three different types of alternative

$$H_a : \mu \neq \mu_0$$
$$H_a : \mu > \mu_0$$
$$H_a : \mu < \mu_0$$

Firstly, we introduce the **confidence interval approach**:

1. Find the right type of confidence interval.

2. Find the corresponding rejection region:

Alternative hypothesis	Reject Criterion
$H_a : \mu \neq \mu_0$	$\mu_0 \notin \left[\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$: μ_0 is not in two-sided confidence interval
$H_a : \mu > \mu_0$	$\overline{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} > \mu_0$: μ_0 is less than the lower confidence bound
$H_a : \mu < \mu_0$	$\overline{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}} < \mu_0$: μ_0 is larger than the upper confidence bound

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population mean is _______ (less than, greater than, or differs from) _______.

Secondly, we introduce **the critical value approach**:

1. Comput the test statistics:
$$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

2. Find the corresponding rejection region:

Alternative hypothesis	Rejection Criterion
$H_a : \mu \neq \mu_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_a : \mu > \mu_0$	$z_0 > z_{\alpha}$
$H_a : \mu < \mu_0$	$z_0 < -z_{\alpha}$

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population mean is _______ (less than, greater than, or differs from) _______.

Why the first two approaches are equivalent? We only show it for the two-sided alternative, i.e., $H_a : \mu \neq \mu_0$.

$$\begin{aligned}
& z_0 > z_{\alpha/2} \text{ or } z_0 < -z_{\alpha/2} \\
\Leftrightarrow\ & \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \text{ or } \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \\
\Leftrightarrow\ & \overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu_0 \text{ or } \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0 \\
\Leftrightarrow\ & \mu_0 \notin \left[\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].
\end{aligned}$$

Finally, we have **the P-value approach**:

1. Comput the test statistics:
$$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

2. Find the corresponding rejection region:

Alternative hypothesis	P-value	Rejection Criterion
$H_a : \mu \neq \mu_0$	$2\{1 - \Phi(\lvert z_0\rvert)\}$	if P-value $< \alpha$
$H_a : \mu > \mu_0$	$1 - \Phi(z_0)$	if P-value $< \alpha$
$H_a : \mu < \mu_0$	$\Phi(z_0)$	if P-value $< \alpha$

   Recall that $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution; i.e., $\Phi(z) = P(Z \leq z)$.

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At $\alpha\%$ significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population mean is _______ (less than, greater than, or differs from) _______.

Why the three approaches are equivalent? We only show the P-value approach and the critical value approach are the same for the two-sided alternative, i.e., $H_a : \mu \neq \mu_0$:

$$\begin{aligned}
& z_0 > z_{\alpha/2} \text{ or } z_0 < -z_{\alpha/2} \\
\Leftrightarrow\ & P(Z \geq |z_0|) < P(Z \geq z_{\alpha/2}) \\
\Leftrightarrow\ & 1 - P(Z \leq |z_0|) < \frac{\alpha}{2} \\
\Leftrightarrow\ & 2\{1 - P(Z \leq |z_0|)\} < \alpha \\
\Leftrightarrow\ & 2\{1 - \Phi(|z_0|)\} < \alpha \\
\Leftrightarrow\ & \text{P-value } < \alpha
\end{aligned}$$

With the same spirit, we can see that these three approaches are essentially the same.

TI- 84 for P-value: STAT→TESTS→ "1: Z-test". INPUT has two options: DATA or STATS. Like in confidence intervals, correctly input the rest information, and correctly choose the type of your alternative hypothesis, press CALCULATE. The output P is the P-value. Then compare it with α.

Example 4.9.1. Civil engineers have found that the ability to see and read a sign at night depends in part on its "surround luminance;" i.e., the light intensity near the sign. It is believed that the mean surround luminance is 10 candela per m^2 in a large metropolitan area. The data below are $n = 30$ measurements of the random variable X, the surround luminance (in candela per m^2). The 30 measurements constitute a random sample from all the signs in the large metropolitan area in question:

10.9	1.7	9.5	2.9	9.1	3.2	9.1	7.4	13.3	13.1
6.6	13.7	1.5	6.3	7.4	9.9	13.6	17.3	3.6	4.9
13.1	7.8	10.3	10.3	9.6	5.7	2.6	15.1	2.9	16.2

Based on past experience, the engineers assume a normal population distribution (for the population of all signs) with known population variance $\sigma^2 = 20$. From this data what conclusions should we draw about the hypothesized mean surround luminance, at the $\alpha = 0.05$ significance level.

Solution. In order to solve this problem one can complete the following 6-step procedure:

1. Identify the parameter of interest: In this case it is the population mean μ.

2. Identify the null hypothesis: In this case $H_0 : \mu = \mu_0$, where $\mu_0 = 10$.

3. Identify the alternative hypothesis: In this case $H_a : \mu \neq \mu_0$, where $\mu_0 = 10$.

4. Follow one of the three approach to perform test.

   - If choosing the confidence interval approach, for the question, we should use the (two-sided) $100 \times (1 - \alpha)\% = 95\%$ confidence interval, which is

     $$[6.9344, 10.19].$$

     We can see it covers μ_0; i.e., $\mu_0 = 10 \in [6.9344, 10.19]$.

   - If choosing the critical value approach, first compute the test statistics:

     $$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{8.62 - 10}{4.47/\sqrt{20}} \approx -1.69.$$

     Since $\alpha = 0.05$, the $H_a : \mu \neq \mu_0$, the rejection region is $z_0 < -z_{\alpha_2} = -1.96$ or $z_0 > z_{\alpha/2} = 1.96$. We can see that our computed z_0 is not in the rejection region.

   - If choosing P-value approach: TI-84 shows that P-value is P ≈ 0.091 which is not smaller than $\alpha = 0.05$.

5. Compare: since

   - μ_0 is covered by the $100 \times (1 - \alpha)\%$ two-sided confidence interval;
   - $z_0 = -1.69$ is not in the rejection region;
   - the P-value is not less than $\alpha = 0.05$;

   we would fail to reject the null hypothesis.

6. Conclusion: At the 0.05 significance level, the data do not provide sufficient evidence to conclude that the mean surround luminance for the specified large metropolitan area differs from 10 candela per m^2.

4.10 Testing hypotheses on the mean μ with unknown variance σ^2

In this case, the population variance σ^2 is not given (which is more common in real applications), like in Section 4.5.3. Suppose $X_1, \ldots, X_n$ is a random sample of size n from a population. The population is normally distributed, (if it is not, the conditions of the central limit theorem apply, and n is sufficiently large.). Then, we have

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

Using this fact, we are now able to test the null hypothesis

$$H_0 : \mu = \mu_0$$

versus one of three different types of alternative

$$H_a : \mu \neq \mu_0$$
$$H_a : \mu > \mu_0$$
$$H_a : \mu < \mu_0$$

Firstly, we introduce the **confidence interval approach**:

1. Find the right type of confidence interval.

2. Find the corresponding rejection region:

Alternative hypothesis	Reject Criterion
$H_a : \mu \neq \mu_0$	$\mu_0 \notin \left[\overline{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$: μ_0 is not in two-sided confidence interval
$H_a : \mu > \mu_0$	$\overline{x} - t_{n-1,\alpha}\frac{s}{\sqrt{n}} > \mu_0$: μ_0 is less than the lower confidence bound
$H_a : \mu < \mu_0$	$\overline{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}} < \mu_0$: μ_0 is larger than the upper confidence bound

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population mean is _______ (less than, greater than, or differs from) _______.

Secondly, we introduce **the critical value approach**:

1. Comput the test statistics:

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

2. Find the corresponding rejection region:

Alternative hypothesis	Rejection Criterion
$H_a : \mu \neq \mu_0$	$t_0 > t_{n-1,\alpha/2}$ or $t_0 < -t_{n-1,\alpha/2}$
$H_a : \mu > \mu_0$	$t_0 > t_{n-1,\alpha}$
$H_a : \mu < \mu_0$	$t_0 < -t_{n-1,\alpha}$

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population mean is _______ (less than, greater than, or differs from) _______.

The value of $t_{n-1,a}$ for $a = \alpha$ or $\alpha/2$ can be found in the following T-table in the same spirit of founding $\chi^2_{n-1,a}$ via the "Chi-squre Distribution Table"

Finally, we have **the P-value approach**:

1. Comput the test statistics:

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

2. Find the corresponding rejection region:

Alternative hypothesis	P-value	Rejection Criterion
$H_a : \mu \neq \mu_0$	$2P(t_{n-1} > \|t_0\|)$	if P-value $< \alpha$
$H_a : \mu > \mu_0$	$P(t_{n-1} > t_0)$	if P-value $< \alpha$
$H_a : \mu < \mu_0$	$P(t_{n-1} < t_0)$	if P-value $< \alpha$

Recall that $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution; i.e., $\Phi(z) = P(Z \leq z)$.

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population mean is _______ (less than, greater than, or differs from) _______.

Using similar argument in previous section, we can see that these three approaches are essentially the same.

TI- 84 for P-value: STAT→TESTS→ "2: T-test". INPUT has two options: DATA or STATS. Like in confidence intervals, correctly input the rest information, and correctly choose the type of your alternative hypothesis, press CALCULATE. The output P is the P-value. Then compare it with α.

t Table upper-tail probability:

df	.25	.10	.05	.025	.01	.005
1	1.0000	3.0777	6.314	12.706	31.821	63.657
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.925
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778

t Table upper-tail probability:

df	.25	.10	.05	.025	.01	.005
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.6772	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.6772	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.6771	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.6771	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.6771	1.2905	1.6611	1.9853	2.3662	2.6286
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259
110	0.6767	1.2893	1.6588	1.9818	2.3607	2.6213
120	0.6765	1.2886	1.6577	1.9799	2.3578	2.6174
∞	0.6745	1.2816	1.6449	1.9600	2.3264	2.5758

Example 4.10.1. Civil engineers have found that the ability to see and read a sign at night depends in part on its "surround luminance;" i.e., the light intensity near the sign. It is believed that the mean surround luminance is 10 candela per m^2 in a large metropolitan area. The data below are $n = 30$ measurements of the random variable X, the surround luminance (in candela per m^2). The 30 measurements constitute a random sample from all the signs in the large metropolitan area in question:

10.9	1.7	9.5	2.9	9.1	3.2	9.1	7.4	13.3	13.1
6.6	13.7	1.5	6.3	7.4	9.9	13.6	17.3	3.6	4.9
13.1	7.8	10.3	10.3	9.6	5.7	2.6	15.1	2.9	16.2

Based on past experience, the engineers assume a normal population distribution (for the population of all signs). From this data what conclusions should we draw about the hypothesized mean surround luminance, at the $\alpha = 0.05$ significance level.

4.11 Testing hypotheses on the population proportion p

To test the null hypothesis

$$H_0 : p = p_0$$

versus one of three different types of alternative

$$H_a : p \neq p_0$$
$$H_a : p > p_0$$
$$H_a : p < p_0$$

We only introduce the critical approach and the P-value approach:

To estimate the population proportion p, we have our point estimator; i.e., sample proportion $\hat{p}$. When $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, the CLT says we have

$$\hat{p} \sim \mathcal{A}N\left(p, \frac{p(1-p)}{n}\right)$$

To control the type I error, recalling that the type I error is under the assumption H_0 is in fact true. So when H_0 is true, we have $p = p_0$. Thus,

$$\hat{p} \sim \mathcal{A}N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

Then

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathcal{A}N(0,1).$$

Based on this, we can design the testing procedures as the following.

Firstly, we introduce **the critical value approach**:

1. Comput the test statistics:
$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

2. Find the corresponding rejection region:

Alternative hypothesis	Rejection Criterion
$H_a : p \neq p_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_a : p > p_0$	$z_0 > z_\alpha$
$H_a : p < p_0$	$z_0 < -z_\alpha$

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population proportion is _______

(less than, greater than, or differs from) ______.

Then, we have **the P-value approach**:

1. Comput the test statistics:
$$z_0 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$
2. Find the corresponding rejection region:

Alternative hypothesis	P-value	Rejection Criterion
$H_a : p \neq p_0$	$2\{1 - \Phi(\lvert z_0 \rvert)\}$	if P-value $< \alpha$
$H_a : p > p_0$	$1 - \Phi(z_0)$	if P-value $< \alpha$
$H_a : p < p_0$	$\Phi(z_0)$	if P-value $< \alpha$

   Recall that $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution; i.e., $\Phi(z) = P(Z \leq z)$.

3. Compare: find whether or not the rejection criterion is satisfied
4. **Conclusions (no credits if no conclusion**: At α% significance level, the data ______ (do or do not) provide sufficient evidence to conclude that the real population proportion is ______ (less than, greater than, or differs from) ______.

TI- 84 for P-value: STAT→TESTS→ "5: 1-PropZTest". Like in confidence intervals, correctly input the rest information, and correctly choose the type of your alternative hypothesis, press CALCULATE. The output P is the P-value. Then compare it with α.

Example 4.11.1. Suppose that 500 parts are tested in manufacturing and 10 are rejected. Using both the approaches (critical value approach and P-value approach) to test the hypothesis $H_0 : p = 0.03$ against $H_a : p < 0.03$ at significance level $\alpha = 0.05$.

4.12 Testing hypotheses on the population variance σ^2

In Section 4.7, we have introduced that

Suppose that $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution (normality is important in this case). The quantity

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

a χ^2 **distribution** with $\nu = n-1$ degrees of freedom.

Using this fact, we are now able to test the null hypothesis

$$H_0 : \sigma = \sigma_0 \text{ (or } H_0 : \sigma^2 = \sigma_0^2)$$

against one of three different types of alternative

$$H_a : \sigma \neq \sigma_0 \text{ (or } H_a : \sigma^2 \neq \sigma_0^2)$$
$$H_a : \sigma > \sigma_0 \text{ (or } H_a : \sigma^2 > \sigma_0^2)$$
$$H_a : \sigma < \sigma_0 \text{ (or } H_a : \sigma^2 < \sigma_0^2)$$

Since one can easily transform testing hypotheses on the population standard deviation σ to testing on the population variance σ^2. Thus, we only need know how to test on σ^2. We have the confidence interval approach and the critical value approach. Unfortunately, TI-84 cannot help us on this case. Thus the P-value approach is not required.

Firstly, we introduce the **confidence interval approach**:

1. Find the right type of confidence interval.

2. Find the corresponding rejection region:

Alternative hypothesis	Reject Criterion
$H_a : \sigma^2 \neq \sigma_0^2$	$\sigma_0^2 \notin \left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right]$: σ_0^2 is not in two-sided confidence interval
$H_a : \sigma^2 > \sigma_0^2$	$\frac{(n-1)s^2}{\chi^2_{n-1,\alpha}} > \sigma_0^2$: σ_0^2 is less than the lower confidence bound
$H_a : \sigma^2 < \sigma_0^2$	$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}} < \sigma_0^2$: σ_0^2 is larger than the upper confidence bound

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population variance is _______ (less than, greater than, or differs from) _______.

Secondly, we introduce **the critical value approach**:

1. Comput the test statistics:

$$Q_0 = \frac{(n-1)s^2}{\sigma_0^2}$$

2. Find the corresponding rejection region:

Alternative hypothesis	Rejection Criterion
$H_a : \sigma^2 \neq \sigma_0^2$	$Q_0 > \chi^2_{n-1,\alpha/2}$ or $Q_0 < \chi^2_{n-1,1-\alpha/2}$
$H_a : \sigma^2 > \sigma_0^2$	$Q_0 > \chi^2_{n-1,\alpha}$
$H_a : \sigma^2 < \sigma_0^2$	$Q_0 < \chi^2_{n-1,1-\alpha}$

3. Compare: find whether or not the rejection criterion is satisfied

4. **Conclusions (no credits if no conclusion**: At α% significance level, the data _______ (do or do not) provide sufficient evidence to conclude that the real population variance is _______ (less than, greater than, or differs from) _______.

The value of $\chi^2_{n-1,a}$ can be found via the "Chi-square Distribution Table"

Example 4.12.1. An automated filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ (fluid ounces)^2. If the variance of fill volume exceeds 0.01 (fluid ounces)^2, an unacceptable proportion of bottles will be underfilled or overfilled. Is there evidence in the sample data to suggest that the manufacturer has a problem with underfilled or overfilled bottles? Use $\alpha = 0.05$, and assume that fill volume has a normal distribution (use both the confidence interval and critical value approaches).

5 Two-sample Statistical Inference

5.1 For the difference of two population means $\mu_1 - \mu_2$: Independent samples

REMARK: In practice, it is very common to compare the same characteristic (mean, proportion, variance) from two different distributions. For example, we may wish to compare

- the **mean** starting salaries of male and female engineers (compare μ_1 and μ_2)
- the **proportion** of scrap produced from two manufacturing processes (compare p_1 and p_2)
- the **variance** of sound levels from two indoor swimming pool designs (compare σ_1^2 and σ_2^2).

Our previous work is applicable only for a single distribution (i.e., a single mean μ, a single proportion p, and a single variance σ^2). We therefore need to extend these procedures to handle two distributions. We start with comparing two means.

Two-sample problem: Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad X_{11}, X_{12}, ..., X_{1n_1} \sim N(\mu_1, \sigma_1^2) \text{ random sample} \\ \text{Sample 2}: &\quad X_{21}, X_{22}, ..., X_{2n_2} \sim N(\mu_2, \sigma_2^2) \text{ random sample.} \end{aligned}$$

GOAL: Construct statistical inference, including a $100(1-\alpha)$ percent confidence interval and hypothesis test at significance level α for the difference of population means $\mu_1 - \mu_2$.

Point Estimators: We define the statistics

$$\begin{aligned} \overline{X}_{1\cdot} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j} &= \text{sample mean for sample 1} \\ \overline{X}_{2\cdot} = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j} &= \text{sample mean for sample 2} \\ S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(X_{1j} - \overline{X}_{1\cdot})^2 &= \text{sample variance for sample 1} \\ S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(X_{2j} - \overline{X}_{2\cdot})^2 &= \text{sample variance for sample 2.} \end{aligned}$$

5.1.1 Known variance case: both σ_1^2 and σ_2^2 are known

Goal: We want to write a confidence interval for $\mu_1 - \mu_2$, but how this interval is constructed depends on the values of σ_1^2 and σ_2^2. In particular, we consider three cases:

- We know the values of σ_1^2 and σ_2^2.
- We do not know the values of σ_1^2 or σ_2^2. However, we know $\sigma_1^2 = \sigma_2^2$; that is, the two population variances are **equal**.
- We do not know the values of σ_1^2 or σ_2^2, also $\sigma_1^2 \neq \sigma_2^2$; that is, the two population variances are **not equal**.

We first consider the equal variance case. Addressing this case requires us to start with the following (sampling) distribution result:

$$Z = \frac{(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \underline{\hspace{3cm}}$$

Some comments are in order:

- For this sampling distribution to hold (exactly), we need
  - the two samples to be independent
  - the two population distributions to be normal (Gaussian)
- The sampling distribution $Z \sim N(0,1)$ should suggest to you that confidence interval quantiles will come from the standard normal distribution; note that this distribution depends on the **sample sizes** from both samples.
- In particular, because $Z \sim N(0,1)$, we can find the value $z_{\alpha/2}$ that satisfies

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

- Substituting Z into the last expression and performing algebraic manipulations, we obtain

$$\underline{\hspace{6cm}}$$

  This is a $100(1-\alpha)$ **percent confidence interval** for the mean difference $\mu_1 - \mu_2$.
  TI-84: STAT: TESTS: 2-SampZInt

- We see that the interval again has the same form:

$$\underbrace{\text{point estimate}}_{\overline{X}_{1\cdot} - \overline{X}_{2\cdot}} \pm \underbrace{\text{quantile}}_{z_{\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

  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."

- The $100(1-\alpha)$ percent confident upper bound of $\mu_1 - \mu_2$ is

  ____

  and the $100(1-\alpha)$ percent confident lower bound of $\mu_1 - \mu_2$ is

  ____

- **Testing hypothesis**: Following the six steps:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

versus each one of the following:

$$H_0 : \mu_1 - \mu_2 \neq \Delta_0$$
$$H_0 : \mu_1 - \mu_2 > \Delta_0$$
$$H_0 : \mu_1 - \mu_2 < \Delta_0$$

Confidence interval approach

Alternative hypothesis	Reject Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$\Delta_0 \notin \left[(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$
$H_a : \mu_1 - \mu_2 > \Delta_0$	$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} > \Delta_0$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \Delta_0$

Critical value approach

$z_0 = \left[(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - \Delta_0\right] / \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Alternative hypothesis	Rejection Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_a : \mu_1 - \mu_2 > \Delta_0$	$z_0 > z_{\alpha}$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$z_0 < -z_{\alpha}$

P-value approach **TI-84**: STAT: TESTS: 2-SampZTest

Alternative hypothesis	Rejection Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	P-value $= 2\{1 - \Phi(\lvert z_0 \rvert)\} < \alpha$
$H_a : \mu_1 - \mu_2 > \Delta_0$	P-value $= 1 - \Phi(z_0) < \alpha$
$H_a : \mu_1 - \mu_2 < \Delta_0$	P-value $= \Phi(z_0) < \alpha$

Example 5.1.1. In the vicinity of a nuclear power plant, environmental engineers from the EPA would like to determine if there is a difference between the 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:	22.0	20.6	15.4	17.9	24.4	15.6	11.4	17.5		

Suppose we know both distributions are normal and $\sigma_1 = 7$, $\sigma_2 = 4$.

(a) Construct a 90 percent confidence interval for the mean difference $\mu_1 - \mu_2$. Here, μ_1 (μ_2) denotes the population mean weight of all fish at location 1 (2).

(b) Test whether the two distribution have the same mean at $\alpha = 0.05$.

5.1.2 Unknown variance case, but we know they are equal; i.e., $\sigma_1^2 = \sigma_2^2$

We now consider the equal variance (**but unknown**) case. Addressing this case requires us to start with the following (sampling) distribution result:

$$T = \frac{(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2),$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Some comments are in order:

- For this sampling distribution to hold (exactly), we need
  - the two 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. Algebraically, it is simply 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).
- The sampling distribution $T \sim t(n_1 + n_2 - 2)$ should suggest to you 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 $T \sim t(n_1 + n_2 - 2)$, we can find the value $t_{n_1+n_2-2,\alpha/2}$ that satisfies

$$P(-t_{n_1+n_2-2,\alpha/2} < T < t_{n_1+n_2-2,\alpha/2}) = 1 - \alpha.$$

- Substituting T into the last expression and performing algebraic manipulations, we obtain

$$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

  This is a $100(1-\alpha)$ **percent confidence interval** for the mean difference $\mu_1 - \mu_2$.
  TI-84: Stat: Tests: 2-SampTInt (Pooled: Yes)

- We see that the interval again has the same form:

$$\underbrace{\text{point estimate}}_{\overline{X}_{1\cdot}-\overline{X}_{2\cdot}} \pm \underbrace{\text{quantile}}_{t_{n_1+n_2-2,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

  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."

- The $100(1-\alpha)$ percent confident upper bound of $\mu_1 - \mu_2$ is

  ____

  and the $100(1-\alpha)$ percent confident lower bound of $\mu_1 - \mu_2$ is

  ____

- **Testing hypothesis**: Following the six steps:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

versus each one of the following:

$$H_0 : \mu_1 - \mu_2 \neq \Delta_0$$
$$H_0 : \mu_1 - \mu_2 > \Delta_0$$
$$H_0 : \mu_1 - \mu_2 < \Delta_0$$

Confidence interval approach

Alternative hypothesis	Reject Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$\Delta_0 \notin \left[(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]$
$H_a : \mu_1 - \mu_2 > \Delta_0$	$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - t_{n_1+n_2-2,\alpha}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} > \Delta_0$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) + t_{n_1+n_2-2,\alpha}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} < \Delta_0$

Critical value approach

$t_0 = \left[(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - \Delta_0\right] / \sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Alternative hypothesis	Rejection Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$t_0 > t_{n_1+n_2-2,\alpha/2}$ or $t_0 < -t_{n_1+n_2-2,\alpha/2}$
$H_a : \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{n_1+n_2-2,\alpha}$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{n_1+n_2-2,\alpha}$

P-value approach

TI-84: STAT: TESTS: 2-SampTTest (Pooled: Yes)

Alternative hypothesis	Rejection Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	P-value $< \alpha$
$H_a : \mu_1 - \mu_2 > \Delta_0$	P-value $< \alpha$
$H_a : \mu_1 - \mu_2 < \Delta_0$	P-value $< \alpha$

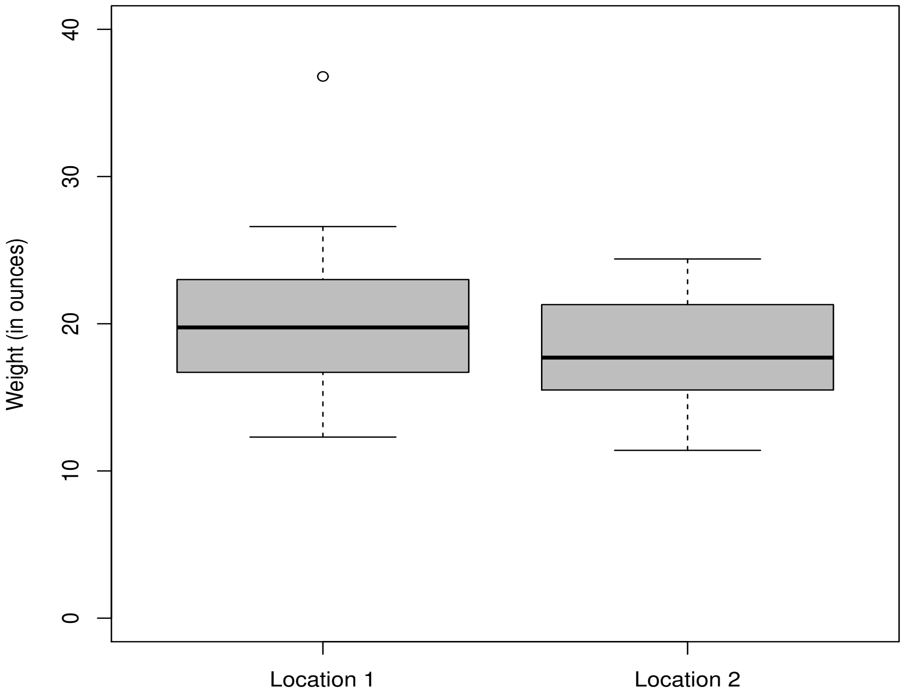


Figure 5.1.1: Boxplots of fish weights by location in Example 5.1.2.

Example 5.1.2. In the vicinity of a nuclear power plant, environmental engineers from the EPA would like to determine if there is a difference between the 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:	22.0	20.6	15.4	17.9	24.4	15.6	11.4	17.5		

Suppose we know both distributions are normal and $\sigma_1^2 = \sigma_2^2$.

(a) Construct a 90 percent confidence interval for the mean difference $\mu_1 - \mu_2$. Here, μ_1 (μ_2) denotes the population mean weight of all fish at location 1 (2).

(b) Test whether the two distribution have the same mean at $\alpha = 0.05$.

5.1.3 Unknown and unequal variance case: $\sigma_1^2 \neq \sigma_2^2$

REMARK: When $\sigma_1^2 \neq \sigma_2^2$, the problem of constructing a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ becomes more difficult theoretically. However, we can still write an **approximate** confidence interval.

An approximate $100(1-\alpha)$ percent confidence interval $\mu_1 - \mu_2$ is given by

$$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

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).
- No one in their right mind would calculate this interval "by hand" (particularly nasty is the formula for ν). **TI-84**: Stat: Tests: 2-SampTInt (Pooled: No) will produce the interval on request.
- We see that the interval again has the same form:

$$\underbrace{\text{point estimate}}_{\overline{X}_{1\cdot}-\overline{X}_{2\cdot}} \pm \underbrace{\text{quantile}}_{t_{\nu,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}.$$

  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."

- The $100(1-\alpha)$ percent confident upper bound of $\mu_1 - \mu_2$ is

  ____

  and the $100(1-\alpha)$ percent confident lower bound of $\mu_1 - \mu_2$ is

  ____

- **Testing hypothesis**: Following the six steps:

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

versus each one of the following:

$$H_0 : \mu_1 - \mu_2 \neq \Delta_0$$
$$H_0 : \mu_1 - \mu_2 > \Delta_0$$
$$H_0 : \mu_1 - \mu_2 < \Delta_0$$

Confidence interval approach

Alternative hypothesis	Reject Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$\Delta_0 \notin \left[(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right]$
$H_a : \mu_1 - \mu_2 > \Delta_0$	$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) - t_{\nu,\alpha}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} > \Delta_0$
$H_a : \mu_1 - \mu_2 < \Delta_0$	$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) + t_{\nu,\alpha}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \Delta_0$

P-value approach **TI-84**: Stat: Tests: 2-SampTTest (Pooled: No)

Alternative hypothesis	Rejection Criterion
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	P-value $< \alpha$
$H_a : \mu_1 - \mu_2 > \Delta_0$	P-value $< \alpha$
$H_a : \mu_1 - \mu_2 < \Delta_0$	P-value $< \alpha$

Remark: In the last two subsections, we have presented two confidence intervals for $\mu_1 - \mu_2$. One assumes $\sigma_1^2 = \sigma_2^2$ (equal variance assumption) and one that assumes $\sigma_1^2 \neq \sigma_2^2$ (unequal variance assumption). **If you are unsure about which interval to use, go 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 $\sigma_1^2 \neq \sigma_2^2$.

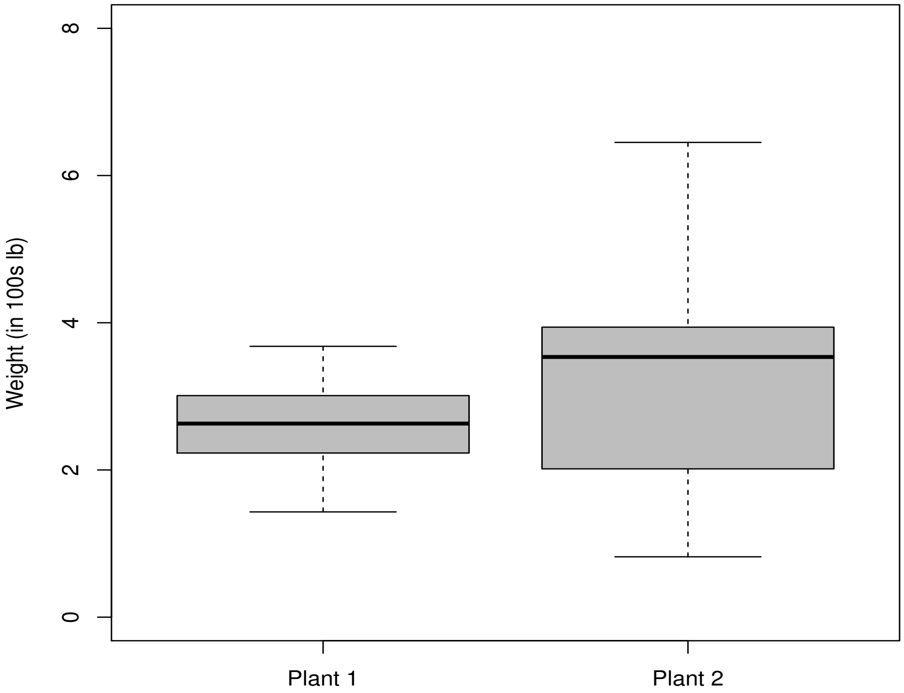


Figure 5.1.2: Boxplots of discarded white paper amounts (in 100s lb) in Example 5.1.3.

Example 5.1.3. 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:	3.79	2.08	3.66	1.53	4.07	4.31	2.62	4.52	3.80	5.30
	3.41	0.82	3.03	1.95	6.45	1.86	1.87	3.78	2.74	3.81

QUESTION. Are there differences in the mean amounts of white paper discarded by employees at the two plants? (use $\alpha = 0.05$).

5.2 For the difference of two population proportions $p_1 - p_2$: Independent samples

We also can extend our confidence interval procedure for a single population proportion p to **two populations**. Define

$$\begin{aligned} p_1 &= \text{population proportion of "successes" in Population 1} \\ p_2 &= \text{population proportion of "successes" in Population 2.} \end{aligned}$$

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

$$\begin{aligned} Y_1 &= \text{number of "successes" in Sample 1 (out of } n_1 \text{ individuals)} \sim b(n_1, p_1) \\ Y_2 &= \text{number of "successes" in Sample 2 (out of } n_2 \text{ individuals)} \sim b(n_2, p_2). \end{aligned}$$

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 write a $100(1-\alpha)$ percent confidence interval and conduct hypothesis test for $p_1 - p_2$, the difference of two population proportions.

To accomplish this goal, we need the following distributional 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 $p_1 - p_2$.

- For the Z sampling distribution to hold approximately (and therefore for the interval above to be useful), we need

- the two 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 CLT should adequately approximate the true sampling distribution of Z, thereby making the confidence interval formula above approximately valid.
- 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 $N(0,1)$ distribution.

- This confidence interval can be calculated through 2-PropZInt in TI-84.

Hypothesis testing: In two-sample situations, it is often of interest to see if the proportions p_1 and p_2 are different; i.e.,

$$H_0 : p_1 = p_2$$

versus one of the following alternatives; i.e.,

$$\begin{aligned} H_a &: p_1 \neq p_2 \\ H_a &: p_1 > p_2 \\ H_a &: p_1 < p_2 \end{aligned}$$

To do this, we need find a testing statistics. Remember that, while we are designing a test, we always want to bound the Type I Error, which is defined as "reject H_0 when H_0 in fact is true." So, when H_0 is in fact true, we can view the two independent samples are from the same population; i.e., $p_1 = p_2 = p$. Then we have Z (define above) can be written as

$$Z_0 = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(\widehat{p}_1 - \widehat{p}_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{AN}(0,1).$$

Further, a natural estimator of p is then

$$\widehat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Thus, we have

$$Z_0 = \frac{(\widehat{p}_1 - \widehat{p}_2)}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim \mathcal{AN}(0,1).$$

Thus, the testing procedure can be summarized as below:

Critical value approach

$z_0 = (\widehat{p}_1 - \widehat{p}_2)/\sqrt{\widehat{p}(1-\widehat{p})(n_1^{-1}+n_2^{-1})}$

Alternative hypothesis	Rejection Criterion
$H_a : p_1 \neq p_2$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_a : p_1 > p_2$	$z_0 > z_\alpha$
$H_a : p_1 < p_2$	$z_0 < -z_\alpha$

P-value approach	**TI-84**: Stat: Tests: 2-PropZTest
Alternative hypothesis	Rejection Criterion
$H_a : p_1 \neq p_2$	P-value $= 2\{1 - \Phi(\lvert z_0\rvert)\} < \alpha$
$H_a : p_1 > p_2$	P-value $= 1 - \Phi(z_0) < \alpha$
$H_a : p_1 < p_2$	P-value $= \Phi(z_0) < \alpha$

Example 5.2.1. A programmable lighting control system is being designed. The purpose of the system is to reduce electricity consumption costs in buildings. The system eventually will entail the use of a large number of transceivers (a device comprised of both a transmitter and a receiver). Two types of transceivers are being considered. In life testing, 200 transceivers (randomly selected) were tested for each type.

Transceiver 1:	20 failures were observed (out of 200)
Transceiver 2:	14 failures were observed (out of 200).

Question. Define p_1 (p_2) to be the population proportion of Transceiver 1 (Transceiver 2) failures. Write a 95 percent confidence interval for $p_1 - p_2$. Is there a significant difference between the failure rates p_1 and p_2?

5.3 For the ratio of two population variances σ_2^2/σ_1^2: Independent samples

You will 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 different 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. This confidence interval is used to compare the **population variances** in two independent samples.

Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim N(\mu_1, \sigma_1^2) \text{ random sample} \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim N(\mu_2, \sigma_2^2) \text{ random sample.} \end{aligned}$$

Goal: Construct a $100(1-\alpha)$ percent confidence interval and conduct hypothesis test for the **ratio of population variances** σ_2^2/σ_1^2.

To accomplish this, we need the following sampling distribution result:

$$R = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1),$$

an F **distribution** with (numerator) $\nu_1 = n_1 - 1$ and (denominator) $\nu_2 = n_2 - 1$ degrees of freedom.

Facts: The F distribution has the following characteristics:

- continuous, skewed right, and always positive
- indexed by two **degree of freedom** parameters ν_1 and ν_2; these are usually integers and are often related to sample sizes
- The F pdf formula is complicated and is unnecessary for our purposes.

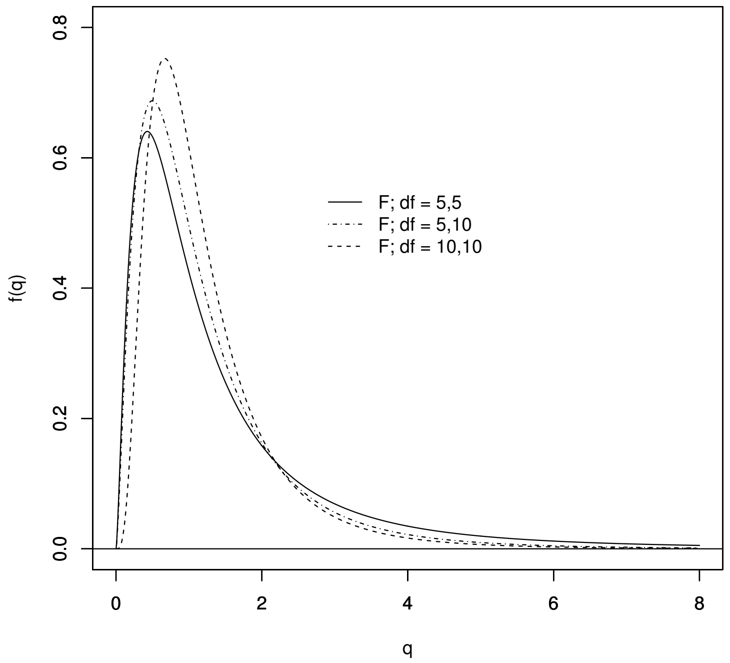


Figure 5.3.3: F probability density functions with different degrees of freedom.

Notation: Let $F_{n_1-1,n_2-1,\alpha/2}$ and $F_{n_1-1,n_2-1,1-\alpha/2}$ denote the upper and lower quantiles, respectively, of the $F(n_1-1,n_2-1)$ distribution; i.e., these values satisfy

$$\begin{aligned} P(R > F_{n_1-1,n_2-1,\alpha/2}) &= \alpha/2 \\ P(R < F_{n_1-1,n_2-1,1-\alpha/2}) &= \alpha/2, \end{aligned}$$

respectively. Similar to the χ^2 distribution, the F distribution is **not symmetric**. Therefore, different notation is needed to identify the quantiles of F distributions.

Because

$$R = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1,n_2-1),$$

we can write

$$\begin{aligned} 1-\alpha &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq R \leq F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \leq F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(\frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,1-\alpha/2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,\alpha/2}\right). \end{aligned}$$

This shows that

$$\left[\frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,1-\alpha/2},\ \frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,\alpha/2}\right]$$

is a $100(1-\alpha)$ **percent confidence interval** for the ratio of the population variances σ_2^2/σ_1^2. Also, we can obtained $100(1-\alpha)$ percent confidence interval for σ_1^2/σ_2^2 as

$$\left[\frac{S_1^2}{S_2^2} \times \frac{1}{F_{n_1-1,n_2-1,\alpha/2}},\ \frac{S_1^2}{S_2^2} \times \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right]$$

Based on the following relationship,

$$F_{\nu_2,\nu_1,a} = \frac{1}{F_{\nu_1,\nu_2,1-a}},$$

we can simplify the above results as

- $100(1-\alpha)$ percent confidence interval for σ_1^2/σ_2^2 is

$$\left[\frac{S_1^2}{S_2^2} \times \frac{1}{F_{n_1-1,n_2-1,\alpha/2}},\ \frac{S_1^2}{S_2^2} \times F_{n_2-1,n_1-1,\alpha/2}\right]$$

  We interpret the interval in the same way.

  > "We are $100(1-\alpha)$ percent confident that the ratio σ_1^2/σ_2^2 is in this interval."

- Taking square root of the interval, we have

$$\left[\frac{S_1}{S_2} \times \sqrt{\frac{1}{F_{n_1-1,n_2-1,\alpha/2}}},\ \frac{S_1}{S_2} \times \sqrt{F_{n_2-1,n_1-1,\alpha/2}}\right]$$

  as a $100(1-\alpha)$ percent confidence interval for σ_1/σ_2

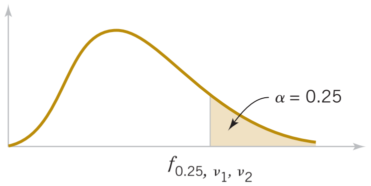


TABLE • VI Percentage Points f_{α,ν_1,ν_2} of the F Distribution

$f_{0.25,\nu_1,\nu_2}$

	ν_1	Degrees of freedom for the numerator (ν_1)																		
	ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Degrees of freedom for the denominator (ν_2)	1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
	2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
	3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
	4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
	6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
	7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.70	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
	8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
	9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
	10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
	11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
	12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
	13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
	14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
	15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
	16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
	17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
	18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
	19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
	20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
	21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
	22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
	23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
	24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
	25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
	26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.25
	27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
	28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
	29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
	30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
	40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19
	60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
	120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
	∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

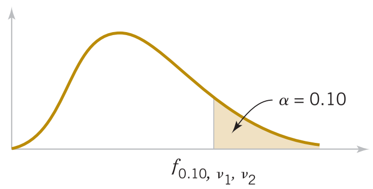


TABLE • VI Percentage Points f_{α,ν_1,ν_2} of the F Distribution (*Continued*)

$f_{0.10,\nu_1,\nu_2}$

ν_1 / ν_2	Degrees of freedom for the numerator (ν_1)																		
Degrees of freedom for the denominator (ν_2)	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

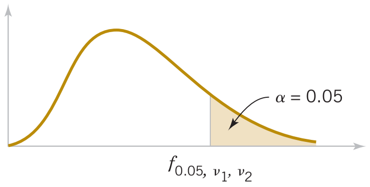


TABLE • VI Percentage Points f_{α,ν_1,ν_2} of the F Distribution (*Continued*)

f_{α,ν_1,ν_2}

ν_1 \ ν_2 Degrees of freedom for the denominator (ν_2)	Degrees of freedom for the numerator (ν_1)																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

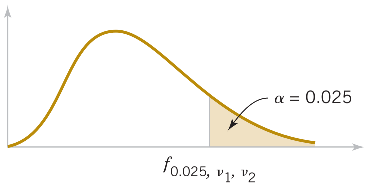


TABLE • VI Percentage Points f_{α,ν_1,ν_2} of the *F* Distribution (*Continued*)

f_{α,ν_1,ν_2}

ν_2 \ ν_1	Degrees of freedom for the numerator (ν_1)																		
Degrees of freedom for the denominator (ν_2)	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

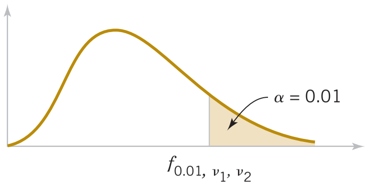


TABLE • VI Percentage Points f_{α,ν_1,ν_2} of the F Distribution (*Continued*)

f_{α,ν_1,ν_2}

	ν_1	Degrees of freedom for the numerator (ν_1)																		
	ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
Degrees of freedom for the denominator (ν_2)	1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.46
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
	15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.59
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Hypothesis testing: Now we are interested in testing

$$H_0 : \sigma_1^2 = \sigma_2^2$$

versus one of the following:

$$\begin{aligned} H_a :&\sigma_1^2 \neq \sigma_2^2 \\ H_a :&\sigma_1^2 > \sigma_2^2 \\ H_a :&\sigma_1^2 < \sigma_2^2 \end{aligned}$$

Noting that, when H_0 is true, we have

$$r_0 = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Thus

Critical value approach
$r_0 = s_1^2/s_2^2$

Alternative hypothesis	Rejection Criterion
$H_a : \sigma_1^2 \neq \sigma_2^2$	$r_0 > F_{n_1-1,n_2-1,\alpha/2}$ or $r_0 < F_{n_1-1,n_2-1,1-\alpha/2} = 1/F_{n_2-1,n_1-1,\alpha/2}$
$H_a : \sigma_1^2 > \sigma_2^2$	$r_0 > F_{n_1-1,n_2-1,\alpha}$
$H_a : \sigma_1^2 < \sigma_2^2$	$r_0 < F_{n_1-1,n_2-1,1-\alpha} = 1/F_{n_2-1,n_1-1,\alpha}$

Confidence interval approach

Alternative hypothesis	Rejection Criterion
$H_a : \sigma_1^2 \neq \sigma_2^2$	$1 \notin \left[\frac{S_1^2}{S_2^2} \times \frac{1}{F_{n_1-1,n_2-1,\alpha/2}},\ \frac{S_1^2}{S_2^2} \times F_{n_2-1,n_1-1,\alpha/2}\right]$
$H_a : \sigma_1^2 > \sigma_2^2$	$1 < \frac{S_1^2}{S_2^2} \times \frac{1}{F_{n_1-1,n_2-1,\alpha}}$
$H_a : \sigma_1^2 < \sigma_2^2$	$1 > \frac{S_1^2}{S_2^2} \times F_{n_2-1,n_1-1,\alpha}$

P-value approach **TI-84**: Stat: Tests: 2-SampFTest

Alternative hypothesis	Rejection Criterion
$H_a : \sigma_1^2 \neq \sigma_2^2$	P-value $< \alpha$
$H_a : \sigma_1^2 > \sigma_2^2$	P-value $< \alpha$
$H_a : \sigma_1^2 < \sigma_2^2$	P-value $< \alpha$

Some statisticians recommend to use this "equal/unequal variance test" before deciding which confidence interval or testing procedure to use for $\mu_1 - \mu_2$. Some statisticians do not.

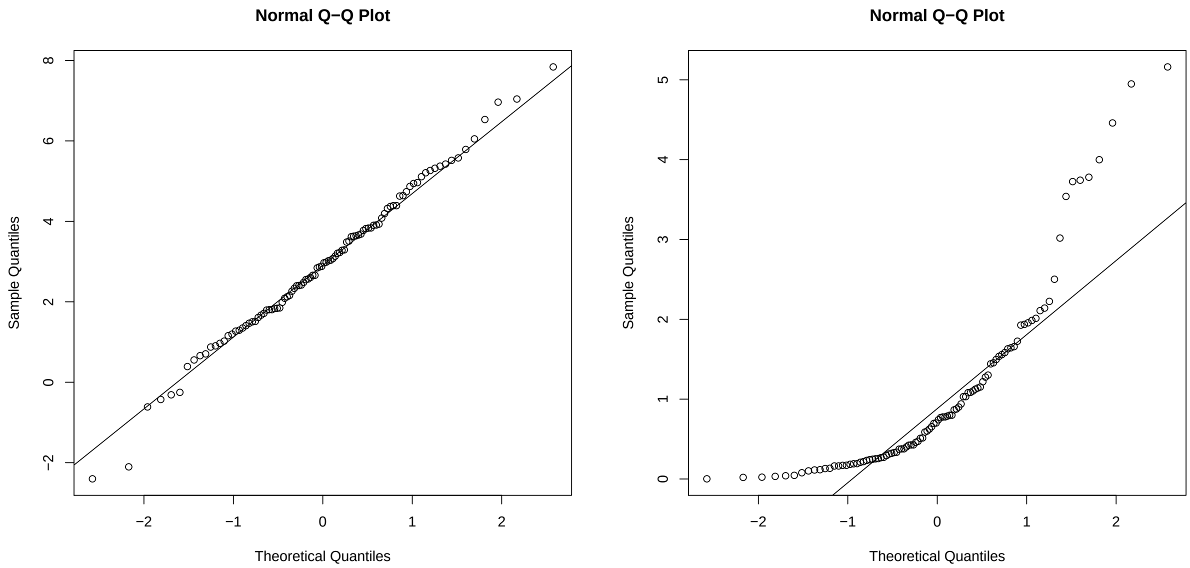
Major warning: Like the χ^2 interval for single population variance σ^2, the two-sample F interval for the ratio of two variances **is not robust** to departures from normality. If the underlying population distributions are non-normal (non-Guassian), then this interval should not be used. One easy solution, to check whether the sample is from a normal distribution or not, is the so-called "**normal qq plot**." It plots sample quantiles against theoretical normal quantiles.

- If it fits well by a line, it is normal.
- Otherwise, non-normal.

You do not need know details about how to construct it. In R program (a famous statistical software), use commend *qqnorm* to draw the plot. such as

```
data=rnorm(100,3,2)  #generate 100 samples from N(3,4)
qqnorm(data)      #plot the dots
qqline(data)      #fit the dots by a line
```

The output is in Figure 5.3.4 (a).



(a) Data from N(3,4)

(b) Data from Exp(1)

Figure 5.3.4: Normal qq plots for normal and exponentional

If it is not normal, what happens?

```
data=rexp(100,1)  #generate 100 samples from Exp(1)
qqnorm(data)      #plot the dots
qqline(data)      #fit the dots by a line
```

The output is in Figure 5.3.4 (b).

Example 5.3.1. We consider again the recycling project in Example 5.1.3 that examined the amount of white paper discarded per employee at two large plants. The data (presented in Example 4.13) were 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 ($n_1 = 25$ and $n_2 = 20$) 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:	3.79	2.08	3.66	1.53	4.07	4.31	2.62	4.52	3.80	5.30
	3.41	0.82	3.03	1.95	6.45	1.86	1.87	3.78	2.74	3.81

The boxplots in Figure 5.1.2 did suggest that the population variances may be different, and the following are the normal qq plots.

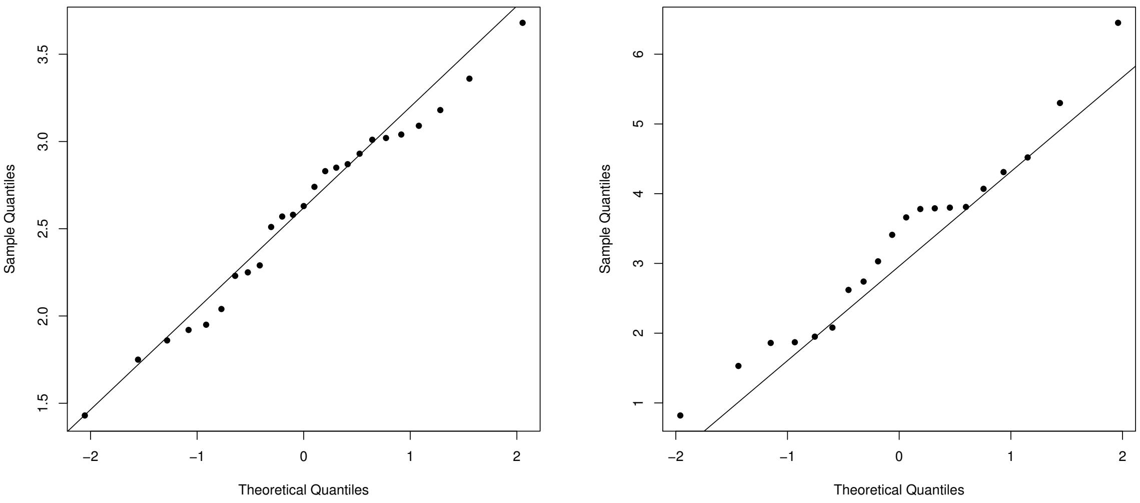


Figure 5.3.5: Normal quantile-quantile (qq) plots for employee recycle data for two plants.

Find a 95 percent confidence interval for σ_2^2/σ_1^2, the ratio of the population variances. Here, σ_1^2 (σ_2^2) denotes the population variance of the amount of white paper by employees at Plant 1 (Plant 2).

5.4 For the difference of two population means $\mu_1 - \mu_2$: Dependent samples (Matched-pairs)

Example 5.4.1. Creatine is an organic acid that helps to supply energy to cells in the body, primarily muscle. Because of this, it is commonly used by those who are weight training to gain muscle mass. *Does it really work?* Suppose that we are designing an experiment involving USC male undergraduates who exercise/lift weights regularly.

Design 1 (Independent samples): Recruit 30 students who are representative of the population of USC male undergraduates who exercise/lift weights. For a single weight training session, we will

- assign 15 students to take creatine.
- assign 15 students an innocuous substance that looks like creatine (but has no positive/negative effect on performance).

For each student, we will record

$$Y = \text{ maximum bench press weight (MBPW).}$$

We will then have two samples of data (with $n_1 = 15$ and $n_2 = 15$):

$$\begin{array}{rl} \text{Sample 1 (Creatine):} & Y_{11}, Y_{12}, ..., Y_{1n_1} \\ \text{Sample 2 (Control):} & Y_{21}, Y_{22}, ..., Y_{2n_2}. \end{array}$$

To compare the population means

$$\begin{array}{rcl} \mu_1 & = & \text{population mean MBPW for students taking creatine} \\ \mu_2 & = & \text{population mean MBPW for students not taking creatine,} \end{array}$$

we could construct a **two-sample** t confidence interval for $\mu_1 - \mu_2$ using

$$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

or

$$(\overline{X}_{1\cdot} - \overline{X}_{2\cdot}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

depending on our underlying assumptions about σ_1^2 and σ_2^2.

Design 2 (Matched Pairs): Recruit 15 students who are representative of the population of USC male undergraduates who exercise/lift weights.

- Each student will be assigned first to take either creatine or the control substance.
- For each student, we will then record his value of Y (MBPW).
- After a period of recovery (e.g., 1 week), we will then have each student take the other "treatment" (creatine/control) and record his value of Y again (but now on the other treatment).
- In other words, for each individual student, we will measure Y under both conditions.

Note: In Design 2, because MBPW measurements are taken on the same student, the difference between the measurement (creatine/control) should be **less variable** than the difference between a creatine measurement on one student and a control measurement on a different student.

- In other words, the student-to-student variation inherent in the latter difference is not present in the difference between MBPW measurements taken on the same individual student.

Table 5.1: Creatine example. Sources of variation in the two independent sample and matched pairs designs.

Design	Sources of Variation
Two Independent Samples	among students, within students
Matched Pairs	within students

Matched pairs: In general, by obtaining a pair of measurements on a single individual (e.g., student, raw material, machine, etc.), where one of measurement corresponds to "Treatment 1" and the other measurement corresponds to "Treatment 2," you eliminate variation **among** the individuals. This allows you to compare the two experimental conditions (e.g., creatine/control, biodegradability treatments, operators, etc.) under more **homogeneous** conditions where only variation within individuals is present (that is, the variation arising from the difference in the two experimental conditions).

Advantage: When you remove extra variability, this enables you to do a better job at comparing the two experimental conditions (treatments). By "better job," I mean, you can **more precisely estimate** the difference between the treatments (excess variability that naturally arises among individuals is not getting in the way). This gives you a better chance of identifying a difference between the treatments if one really exists.

Note: In matched pairs experiments, 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.

Table 5.2: Creatine data. Maximum bench press weight (in lbs) for creatine and control treatments with 15 students. NOTE: These are not real data.

Student j	Creatine MBPW	Control MBPW	Difference ($D_j = Y_{1j} - Y_{2j}$)
1	230	200	30
2	140	155	-15
3	215	205	10
4	190	190	0
5	200	170	30
6	230	225	5
7	220	200	20
8	255	260	-5
9	220	240	-20
10	200	195	5
11	90	110	-20
12	130	105	25
13	255	230	25
14	80	85	-5
15	265	255	10

Implementation: (*Basically, you take the difference, then perform a T-test on the differences*) 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:

$$D_1, D_2, ..., D_n,$$

the so-called **data differences**. The one sample $100(1-\alpha)$ percent confidence interval

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

where $\overline{D}$ and S_D are the sample mean and sample standard deviation of the differences, respectively, is an interval estimate for

$$\mu_D = \text{ mean difference between the 2 treatments.}$$

We interpret the interval in the same way.

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

Testing hypothesis: (*Basically, you need run a T-test on the differences*) Following the six steps (now we have $\mu_D = \mu_1 - \mu_2$), test

$$H_0 : \mu_D = \Delta_0$$

versus each one of the following:

$$H_0 : \mu_D \neq \Delta_0$$
$$H_0 : \mu_D > \Delta_0$$
$$H_0 : \mu_D < \Delta_0$$

Confidence interval approach

Alternative hypothesis	Reject Criterion
$H_a : \mu_D \neq \Delta_0$	$\Delta_0 \notin \left[\overline{D} \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}}\right]$
$H_a : \mu_D > \Delta_0$	$\Delta_0 < \overline{D} - t_{n-1,\alpha}\frac{S_D}{\sqrt{n}}$
$H_a : \mu_D < \Delta_0$	$\Delta_0 > \overline{D} + t_{n-1,\alpha}\frac{S_D}{\sqrt{n}}$

Critical value approach

$t_0 = \frac{\overline{D}-\Delta_0}{S_D/\sqrt{n}}$

Alternative hypothesis	Rejection Criterion
$H_a : \mu_D \neq \Delta_0$	$t_0 > t_{n-1,\alpha/2}$ or $t_0 < -t_{n-1,\alpha/2}$
$H_a : \mu_D > \Delta_0$	$t_0 > t_{n-1,\alpha}$
$H_a : \mu_D < \Delta_0$	$t_0 < -t_{n-1,\alpha}$

P-value approach **TI-84**: Use T-test on the differences

Alternative hypothesis	Rejection Criterion
$H_a : \mu_D \neq \Delta_0$	P-value $< \alpha$
$H_a : \mu_D > \Delta_0$	P-value $< \alpha$
$H_a : \mu_D < \Delta_0$	P-value $< \alpha$

Now let us finish the example of Creatine. Find a 95% confidence interval for μ_D and test whether there is a difference using $\alpha = 0.05$.

5.5 One-way analysis of variance

So far in this chapter, we have discussed confidence intervals for a single population mean μ and for the difference of two population means $\mu_1 - \mu_2$. When there are two means, we have recently seen that the design of the experiment/study completely determines how the data are to be analyzed.

- 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.

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 differences among the individuals.

Blocking: 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, individuals themselves are treated as "blocks."

The analysis of data from experiments involving blocking will not be covered in this course (see, e.g., STAT 506, STAT 525, and STAT 706). We focus herein on a simpler setting, that is, a **one-way classification model**. This is an extension of the two independent-sample design to more than two treatments.

One-way Classification: Consider an experiment to compare $t \geq 2$ treatments set up as follows:

- We obtain a random sample of individuals and randomly assign them to treatments. Samples corresponding to the treatment groups are **independent** (i.e., the individuals in each treatment sample are unrelated).
- In **observational studies** (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.
- 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 design, the only way in which individuals are "classified" is by the treatment group assignment. Hence, such an arrangement is called a **one-way classification**. When individuals are thought to be "basically alike" (other than the possible effect of treatment), experimental error consists only of the variation among the individuals themselves, that is, there are no other **systematic** sources of variation.

Example 5.5.1. Four types of mortars: (1) ordinary cement mortar (OCM), polymer impregnated mortar (PIM), resin mortar (RM), and (4) polymer cement mortar (PCM), were 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 the data are in Figure 5.5.6.

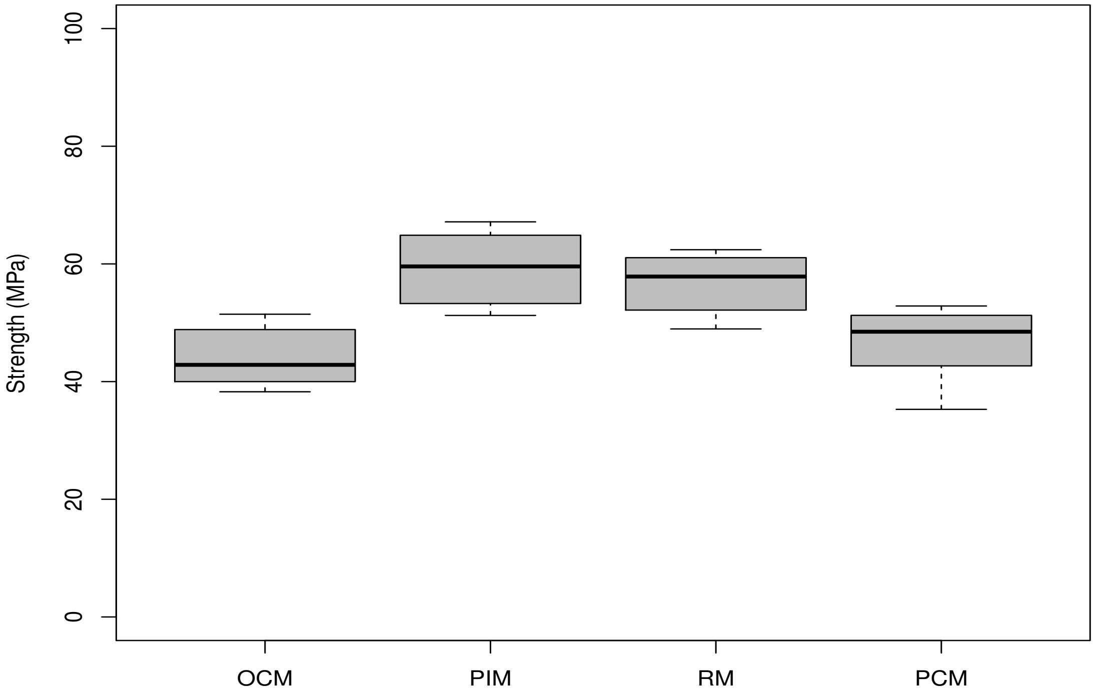


Figure 5.5.6: Boxplots of strength measurements (MPa) for four mortar types.

In this example,

- "Treatment" = mortar type (OCM, PIM, RM, and PCM). There are $t = 4$ treatment groups.
- Individuals = mortar specimens
- This is an example of an **observational study**; not an experiment. That is, we do not physically apply a treatment here; instead, the mortar specimens are inherently different to begin with. We simply take random samples of each mortar type.

Query: An initial question that we might have is the following:

> "Are the treatment (mortar type) population means equal? Or, are the treatment population means different?"

This question can be answered by performing a **hypothesis test**, that is, by testing

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$\text{versus}$$
$$H_a : \text{the population means } \mu_i \text{ are not all equal.}$$

Goal: We now develop a statistical procedure that allows us to test this type of hypothesis in a one-way classification model.

ASSUMPTIONS: We have **independent** random samples from $t \geq 2$ **normal** distributions, each of which has the **same variance** (but possibly different means):

$$\begin{array}{ll} \text{Sample 1:} & Y_{11}, Y_{12}, ..., Y_{1n_1} \sim N(\mu_1, \sigma^2) \\ \text{Sample 2:} & Y_{21}, Y_{22}, ..., Y_{2n_2} \sim N(\mu_2, \sigma^2) \\ \vdots & \vdots \\ \text{Sample } t\text{:} & Y_{t1}, Y_{t2}, ..., Y_{tn_t} \sim N(\mu_t, \sigma^2). \end{array}$$

STATISTICAL HYPOTHESIS: Our goal is to develop a procedure to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$
$$\text{versus}$$
$$H_a : \text{the population means } \mu_i \text{ are not all equal.}$$

- The **null hypothesis** H_0 says that there is "no treatment difference," that is, all treatment population means are the same.
- The **alternative hypothesis** H_a says that a difference among the t population means exists somewhere (but does not specify how the means are different).
- The goal is to decide which hypothesis is more supported by the observed data.

To do this, I need introduce some notation:

Notation: Let t denote the number of treatments 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 **replications** for treatment i
- 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

$$\begin{aligned}\overline{X}_{i\cdot} &= \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\\ S_i^2 &= \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{X}_{i\cdot})^2\\ \overline{X}_{\cdot\cdot} &= \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}.\end{aligned}$$

  The statistics $\overline{X}_{i}$. and S_i^2 are simply the sample mean and sample variance, respectively, of the ith sample. The statistic $\overline{X}$.. is the sample mean of all the data (across all t treatment groups).

The procedure we develop is formulated by deriving two estimators for σ^2. These two estimators are formed by (1) looking at the variance of the observations **within** samples, and (2) looking at the variance of the sample means **across** the t samples.

- **"WITHIN" Estimator**: To estimate σ^2 **within** samples, we take a weighted average (weighted by the sample sizes) of the t sample variances; that is, we "pool" all variance estimates together to form one estimate. Define

$$\begin{aligned}SS_{res} &= (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_t - 1)S_t^2\\ &= \sum_{i=1}^{t}\underbrace{\sum_{j=1}^{n_i}(Y_{ij} - \overline{X}_{i\cdot})^2}_{(n_i-1)S_i^2}.\end{aligned}$$

  We call SS_{res} the **residual sum of squares**. Mathematics shows that

$$E\left(\frac{SS_{res}}{\sigma^2}\right) = N - t \implies E(MS_{res}) = \sigma^2,$$

where

$$MS_{res} = \frac{SS_{res}}{N-t}.$$

Important: MS_{res} is an unbiased estimator of σ^2 regardless of whether or not H_0 is true. We call MS_{res} the **residual mean squares**.

- **"ACROSS" Estimator**: To derive the "across-sample" estimator, we assume a common sample size $n_1 = n_2 = \cdots = n_t = n$ (to simplify notation). Recall that if a sample arises from a normal population, then the sample mean is also normally distributed, i.e.,

$$\overline{X}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma^2}{n}\right).$$

NOTE: If all the treatment population means are equal, that is,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t = \mu, \quad \text{say},$$

is true, then

$$\overline{X}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

If H_0 is true, then the t sample means $\overline{X}_{1\cdot}, \overline{X}_{2\cdot}, ..., \overline{X}_{t\cdot}$ are a random sample of size t from a normal distribution with mean μ and variance σ^2/n. The sample variance of this "random sample" is given by

$$\frac{1}{t-1}\sum_{i=1}^{t}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

and has expectation

$$E\left[\frac{1}{t-1}\sum_{i=1}^{t}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2\right] = \frac{\sigma^2}{n}.$$

Therefore,

$$MS_{trt} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2}_{SS_{trt}},$$

is an unbiased estimator of σ^2; i.e., $E(MS_{trt}) = \sigma^2$, when H_0 is true. We call SS_{trt} the **treatment sums of squares** and MS_{trt} the **treatment mean squares**. MS_{trt} is our second point estimator for σ^2. Recall that MS_{trt} is an unbiased estimator of σ^2 **only when $\boldsymbol{H_0 : \mu_1 = \mu_2 = \cdots = \mu_t}$ is true (this is important!)**. If we have different sample sizes, we simply adjust MS_{trt} to

$$MS_{trt} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2}_{SS_{trt}}.$$

This is still an unbiased estimator for σ^2 when H_0 is true.

Motivation:

- **When H_0 is true** (i.e., the treatment means are the same), then

$$\begin{aligned} E(MS_{trt}) &= \sigma^2 \\ E(MS_{res}) &= \sigma^2. \end{aligned}$$

These two facts suggest that when H_0 is true,

$$F = \frac{MS_{trt}}{MS_{res}} \approx 1.$$

- **When H_0 is not true** (i.e., the treatment means are different), then

$$\begin{aligned} E(MS_{trt}) &> \sigma^2 \\ E(MS_{res}) &= \sigma^2. \end{aligned}$$

These two facts suggest that when H_0 is not true,

$$F = \frac{MS_{trt}}{MS_{res}} > 1.$$

Sampling distribution: When H_0 is true, the F statistic

$$F = \frac{MS_{trt}}{MS_{res}} \sim F(t-1, N-t).$$

DECISION: We "reject H_0" and conclude the treatment population means are different if the F statistic is far out in the right tail of the $F(t-1, N-t)$ distribution. Why? Because a large value of F is not consistent with H_0 being true! Large values of F (far out in the right tail) are more consistent with H_a. Thus, to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

versus

$$H_a : \text{the population means } \mu_i \text{ are not all equal,}$$

The rejection criterion would be

$$F = \frac{MS_{trt}}{MS_{res}} > F_{t-1,N-1,\alpha}.$$

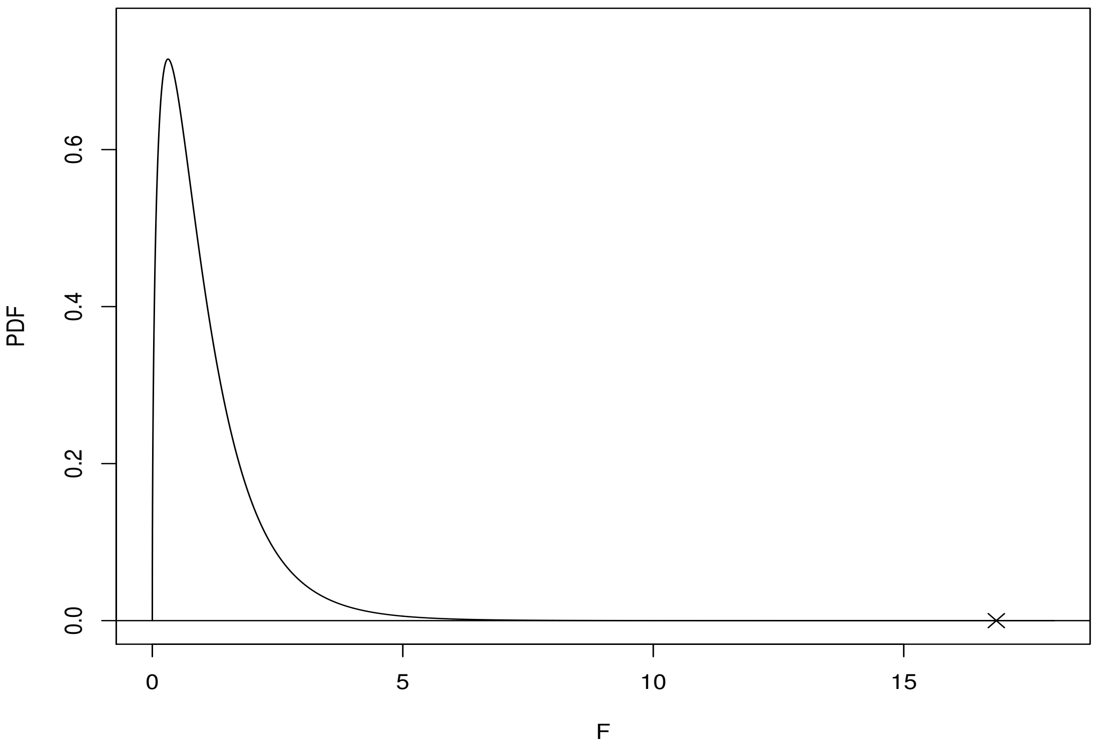


Figure 5.5.7: The $F(3, 32)$ probability density function. This is the distribution of F in Example 4.17 if H_0 is true. An "×" at $F = 16.848$ has been added.

Example 5.5.2. (continued Example 5.5.1.) **Morta Data**: We now use TI-84 to do this test for the strength/mortar type data in Example 5.5.1 at $\alpha = 0.05$.

```
Input data in to list L1, L2, L3, and L4. Then go to Stat, to Tests, select ANOVA
‘‘ANOVA(L1, L2, L3, L4)" The output are then:
One-way ANOVA
   F=16.84834325 (This the F value used for test)
   p=9.5764486E-7 (P-value for the test)
   Factor
     df=3 (t-1)
     SS=1520.87591 (SStrt)
     MS=506.958637 (MStrt)
   Error
     df=32 (N-t)
     SS=962.864785 (SSres)
     MS=30.0895245 (MSres)
   Sxp=5.48539192 (estimate of the common standard deviation sigma)
```

Conclusion: P-value is significantly smaller than $\alpha = 0.05$ (i.e.; $F = 16.848$ is not an observation we would expect from the $F(3, 32)$ distribution (the distribution of F when H_0 is true); see Figure 5.5.7). Therefore, we reject H_0 and conclude the population mean strengths for the four mortar types are different. In other words, at $\alpha = 0.05$ (note that, not only for $\alpha = 0.05$, but also for $\alpha = 0.01, 0.005, 0.0005$, the evidence from the data provides sufficient evidence to reject H_0.

As we have just seen (from the recent R analysis), 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}	$MS_{trt} = \frac{SS_{trt}}{t-1}$	$F = \frac{MS_{trt}}{MS_{res}}$
Residuals	$N-t$	SS_{res}	$MS_{res} = \frac{SS_{res}}{N-t}$	
Total	$N-1$	SS_{total}		

For example, we can re-organize the above output as,

```
Analysis of Variance Table
            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
Total       35 2483.74
```

- It is easy to show that
$$SS_{total} = SS_{trt} + SS_{res}.$$
- SS_{total} measures how observations vary about the overall mean, without regard to treatments; that is, it 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 treatments
  - 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.

6 Linear regression

6.1 Introduction

IMPORTANCE: A problem that arises 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 continuous random variable Y 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 $g : \mathbb{R}^k \to \mathbb{R}$. 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 $g(x_1, x_2, ..., x_k)$.

LINEAR MODELS: 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 β_0, β_1, ..., β_k. We call this a **linear regression model**.

- The **response variable** Y is assumed to be random (but we do get to observe its value).
- The **regression parameters** β_0, β_1, ..., β_k are assumed to be fixed and unknown.
- The **independent variables** $x_1, x_2, ..., x_k$ are assumed to be fixed (not random).
- The **error term** ϵ is assumed to be random (and not observed).

DESCRIPTION: 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:

$$\begin{aligned} Y &= \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}_{g(x_1,x_2)} + \epsilon. \end{aligned}$$

Main point: The term "linear" does not refer to the shape of the regression function g. It refers to how the regression parameters β_0, β_1, ..., β_k enter the g function.

6.2 Simple linear regression

Terminology: A **simple linear regression model** includes only one independent variable x and is of the form

$$\begin{aligned} Y &= g(x) + \epsilon \\ &= \beta_0 + \beta_1 x + \epsilon. \end{aligned}$$

The regression function

$$g(x) = \beta_0 + \beta_1 x$$

is a straight line with intercept β_0 and slope β_1. 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 these interpretations for the regression parameters β_0 and β_1:

- β_0 quantifies the mean of Y when $x = 0$.
- β_1 quantifies the change in $E(Y)$ brought about by a one-unit change in x.

Example 6.2.1. As part of a waste removal project, a new compression machine for processing sewage sludge is being studied. In particular, 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 $n = 20$ observations of (x, Y); 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 6.3: Sewage data. Moisture (Y, measured as a percentage) and machine filtration rate (x, measured in kg-DS/m/hr). There are $n = 20$ observations.

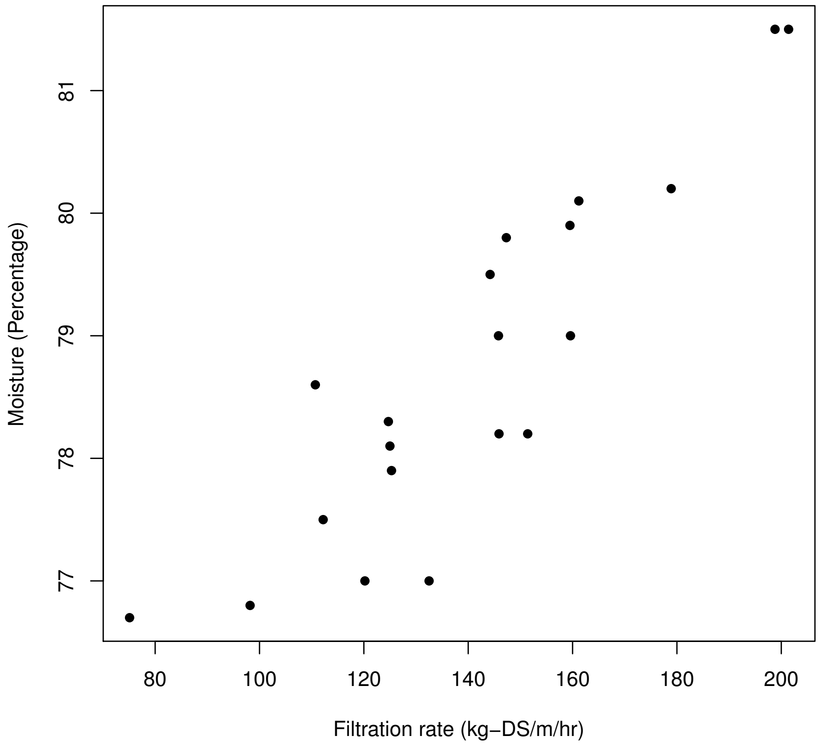


Figure 6.2.1: Scatterplot of pellet moisture Y (measured as a percentage) as a function of machine filtration rate x (measured in kg-DS/m/hr).

Figure 6.2.1 displays the data in a **scatterplot**. This is the most common graphical display for bivariate data like those seen above. From the plot, we see that

- the variables Y and x are **positively related**, that is, an increase in x tends to be associated with an increase in Y.
- the variables Y and x are **linearly related**, although there is a large amount of variation that is unexplained.
- this is an example where a simple linear regression model may be adequate.

6.2.1 Least squares estimation

Terminology: When we say, "fit a regression model," we mean that we would like to estimate the regression parameters in the model with the observed data. Suppose that we collect (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 each $i = 1, 2, ..., n$. Our first goal is to estimate β_0 and β_1. Formal assumptions for the error terms ϵ_i will be given later.

Least Squares: A widely-accepted method of estimating the model 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 $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively, that is, the values of β_0 and β_1 that minimize $Q(\beta_0, \beta_1)$. A two-variable minimization argument can be used to find closed-form expressions for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Taking partial derivatives of $Q(\beta_0, \beta_1)$, we obtain

$$\begin{aligned}
\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0 \\
\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i \overset{\text{set}}{=} 0.
\end{aligned}$$

Solving for β_0 and β_1 gives the **least squares estimators**

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x} \\
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}}.
\end{aligned}$$

In real life, it is rarely necessary to calculate $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by hand, TI-84 and R automate the entire model fitting process and subsequent analysis.

Example 6.2.2. (continued Example 6.2.1). We now use R to calculate the equation of the least squares regression line for the sewage sludge data in Example 6.1. Here is the output:

```
> fit = lm(moisture˜filtration.rate)
> fit
lm(formula = moisture ˜ filtration.rate)
Coefficients:
    (Intercept)  filtration.rate
       72.95855          0.04103
```

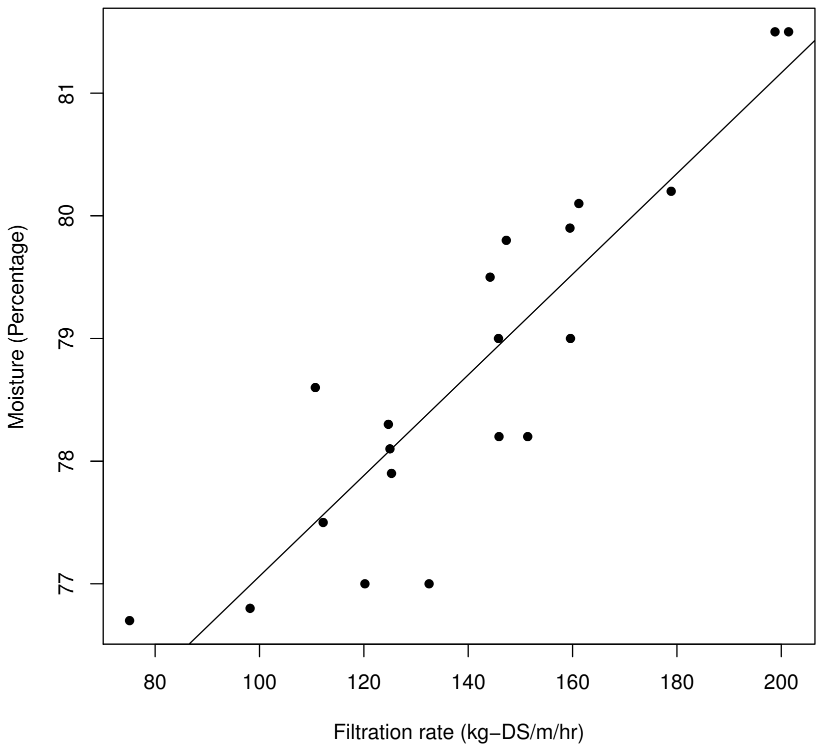


Figure 6.2.2: Scatterplot of pellet moisture Y (measured as a percentage) as a function of filtration rate x (measured in kg-DS/m/hr). The least squares line has been added.

From the output, we see the least squares estimates (to 3 dp) for the sewage data are

$$\begin{aligned}\widehat{\beta}_0 &= 72.959\\ \widehat{\beta}_1 &= 0.041.\end{aligned}$$

Therefore, the equation of the least squares line that relates moisture percentage Y to the filtration rate x is

$$\widehat{Y} = 72.959 + 0.041x,$$

or, in other words,

$$\widehat{\texttt{Moisture}} = 72.959 + 0.041 \times \ \texttt{Filtration rate}.$$

The textbook authors call the least squares line the **prediction equation**. This is because we can predict the value of Y (moisture) for any 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

$$\widehat{Y}(150) = 72.959 + 0.041(150) \approx 79.109.$$

6.2.2 Model assumptions and properties of least squares estimators

We wish to investigate the properties of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as estimators of the true 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 assumptions on the error terms ϵ_i. Specifically, 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.

Under these assumptions,

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Fact 1. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are **unbiased estimators** of β_0 and β_1, respectively, that is,

$$\begin{aligned} E(\widehat{\beta}_0) &= \beta_0 \\ E(\widehat{\beta}_1) &= \beta_1. \end{aligned}$$

Fact 2. The least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have the following sampling distributions:

$$\widehat{\beta}_0 \sim N(\beta_0, c_{00}\sigma^2) \quad \text{and} \quad \widehat{\beta}_1 \sim N(\beta_1, c_{11}\sigma^2),$$

where

$$c_{00} = \frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}} \quad \text{and} \quad c_{11} = \frac{1}{SS_{xx}}.$$

Knowing these sampling distributions is critical if we want to write confidence intervals and perform hypothesis tests for β_0 and β_1.

6.2.3 Estimating the error variance

Goal: In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$, we now turn our attention to estimating σ^2, the **error variance**.

Terminology: In the simple linear regression model, define the ith **fitted value** by

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least squares estimators. Each observation has its own fitted value. Geometrically, an observation's fitted value is the (perpendicular) projection of its Y value, upward or downward, onto the least squares line.

Terminology: We define the ith **residual** by

$$e_i = Y_i - \widehat{Y}_i.$$

Each observation has its own residual. Geometrically, an observation's residual is the vertical distance (i.e., length) between its Y value and its fitted value.

- If an observation's Y value is above the least squares regression line, its residual is positive.
- If an observation's Y value is below the least squares regression line, its residual is negative.
- In the simple linear regression model (provided that the model includes an intercept term β_0), we have the following algebraic result:

  $$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i) = 0,$$

  that is, the sum of the residuals (from a least squares fit) is equal to zero.

SEWAGE DATA: In Table 6.2, I have used R to calculate the fitted values and residuals for each of the $n = 20$ observations in the sewage sludge data set.

Obs	x	Y	$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$	$e = Y - \widehat{Y}$	Obs	x	Y	$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$	$e = Y - \widehat{Y}$
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 6.4: Sewage data. Fitted values and residuals from the least squares fit.

Terminology: We define the **residual sum of squares** by

$$SS_{res} \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

Fact 3. In the simple linear regression model,

$$MS_{res} = \frac{SS_{res}}{n-2}$$

is an unbiased estimator of σ^2, that is, $E(MS_{res}) = \sigma^2$. The quantity

$$\widehat{\sigma} = \sqrt{MS_{res}} = \sqrt{\frac{SS_{res}}{n-2}}$$

estimates σ and is called the **residual standard error**.

Example 6.2.3. SEWAGE DATA: For the sewage data in Example 6.2.1, we use R to calculate MS_{res}:

```
> fitted.values = predict(fit)
> residuals = moisture-fitted.values
> # Calculate MS_res
> sum(residuals^2)/18
[1] 0.4426659
```

For the sewage data, an (unbiased) estimate of the error variance σ^2 is

$$MS_{res} \approx 0.443.$$

The residual standard error is

$$\widehat{\sigma} = \sqrt{MS_{res}} = \sqrt{0.4426659} \approx 0.6653.$$

This estimate can also be seen in the following R output:

```
> 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 ***

Residual standard error: 0.6653 on 18 degrees of freedom
Multiple R-squared: 0.7999,     Adjusted R-squared: 0.7888
F-statistic: 71.97 on 1 and 18 DF,  p-value: 1.052e-07
```

6.2.4 Inference for β_0 and β_1

In the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

the regression parameters β_0 and β_1 are unknown. It is therefore of interest to construct confidence intervals and perform hypothesis tests for these parameters.

- 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.
- 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 for β_1: Under our model assumptions, the following sampling distribution arises:

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{MS_{res}/SS_{xx}}} \sim t(n-2).$$

This result can be used to derive a $100(1-\alpha)$ **percent confidence interval** for β_1, which is given by

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{MS_{res}/SS_{xx}}.$$

- 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:

$$\underbrace{\text{point estimate}}_{\widehat{\beta}_1} \pm \underbrace{\text{quantile}}_{t_{n-2,\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{MS_{res}/SS_{xx}}}.$$

  We interpret the interval in the same way.

  > "We are $100(1-\alpha)$ percent confident that the population regression slope β_1 is in this interval."

- When interpreting the interval, of particular interest to us is the value $\beta_1 = 0$.
  - If $\beta_1 = 0$ is in the confidence interval, this suggests that Y and x are not linearly related.
  - If $\beta_1 = 0$ is not in the confidence interval, this suggests that Y and x are linearly related.
- The $100(1-\alpha)$ percent lower and upper confidence bounds are, respectively,

$$\widehat{\beta}_1 - t_{n-2,\alpha}\sqrt{MS_{res}/SS_{xx}} \quad \text{and} \quad \widehat{\beta}_1 + t_{n-2,\alpha}\sqrt{MS_{res}/SS_{xx}}.$$

Hypothesis test for β_1: If our interest was to test

$$H_0 : \beta_1 = 0$$

versus one of the following

$$H_a : \beta_1 \neq 0$$
$$H_a : \beta_1 > 0$$
$$H_a : \beta_1 < 0$$

where 0 is a fixed value (often, $0 = 0$), we would focus our attention on

$$t_0 = \frac{\widehat{\beta}_1}{\sqrt{MS_{res}/SS_{xx}}}.$$

Critical value approach

Alternative hypothesis	Rejection Criterion
$H_a : \beta_1 \neq 0$	$t_0 > t_{n-2,\alpha/2}$ or $t_0 < -t_{n-2,\alpha/2}$
$H_a : \beta_1 > 0$	$t_0 > t_{n-2,\alpha}$
$H_a : \beta_1 < 0$	$t_0 < -t_{n-2,\alpha}$

P-value approach

Alternative hypothesis	Rejection Criterion
$H_a : \beta_1 \neq 0$	P-value $< \alpha$
$H_a : \beta_1 > 0$	P-value $< \alpha$
$H_a : \beta_1 < 0$	P-value $< \alpha$

Confidence interval approach

Alternative hypothesis	Reject Criterion
$H_a : \beta_1 \neq 0$	$0 \notin \left[\widehat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{MS_{res}/SS_{xx}}\right]$
$H_a : \beta_1 > 0$	$0 < \widehat{\beta}_1 - t_{n-2,\alpha}\sqrt{MS_{res}/SS_{xx}}$
$H_a : \beta_1 < 0$	$0 > \widehat{\beta}_1 + t_{n-2,\alpha}\sqrt{MS_{res}/SS_{xx}}$

Example 6.2.4. (continued Example 6.2.1). We now use R to test

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0,$$

for the sewage sludge data in Example 6.2.1.

```
> 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 ***
```

Analysis: We have $t = 8.484 > t_{n-2,\alpha/2} = t_{18,0.025} = 2.1009$ and P-value $= 0.000000105$. In other words, there is strong evidence to reject H_0. Also a 95 percent confidence interval for β_1 is calculated as follows:

$$\widehat{\beta}_1 \pm t_{18,0.025}\text{se}(\widehat{\beta}_1) \quad\Longrightarrow\quad 0.0410 \pm 2.1009(0.0048) \quad\Longrightarrow\quad (0.0309, 0.0511).$$

We are 95 percent confident that population regression slope β_1 is between 0.0309 and 0.0511. Note that this interval does not include "0."

6.2.5 Confidence and prediction intervals for a given $x = x_0$

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. We are often interested in using the fitted model to learn about the response variable Y at a certain setting for the independent variable $x = x_0$, say. For example, in our sewage sludge example, we might be 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 mean response** of Y when $x = x_0$. This mean response is denoted by $E(Y|x_0)$. This value is the mean of the following probability distribution:

$$Y(x_0) \sim N(\beta_0 + \beta_1 x_0, \sigma^2).$$

- We might be interested in **predicting a new response** Y when $x = x_0$. This predicted response is denoted by $Y^*(x_0)$. This value is a new outcome from

$$Y(x_0) \sim N(\beta_0 + \beta_1 x_0, \sigma^2).$$

In the first problem, we are interested in **estimating** the mean of the response variable Y at a certain value of x. In the second problem, we are interested in **predicting** the value of a new random variable Y at a certain value of x. Conceptually, the second problem is far more difficult than the first.

Goals: We would like to create $100(1-\alpha)$ percent intervals for the mean $E(Y|x_0)$ and for the new value $Y^*(x_0)$. The former is called a **confidence interval** (since it is for a mean response) and the latter is called a **prediction interval** (since it is for a new random variable).

Point estimator/Predictor: To construct either interval, we start with the same quantity:

$$\widehat{Y}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0,$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ 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 mean $E(Y|x_0)$ is given by

$$\widehat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{MS_{res}\left[\frac{1}{n}+\frac{(x_0-\overline{x})^2}{SS_{xx}}\right]}.$$

Prediction interval: A $100(1-\alpha)$ **percent prediction interval** for the new response $Y^*(x_0)$ is given by

$$\widehat{Y}(x_0) \pm t_{n-2,\alpha/2}\sqrt{MS_{res}\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{SS_{xx}}\right]}.$$

- **Comparison**: The two intervals are identical except for the extra "1" in the standard error part of the prediction interval. This extra "1" arises from the additional uncertainty associated with predicting a new response from the $N(\beta_0+\beta_1 x_0, \sigma^2)$ distribution. Therefore, a $100(1-\alpha)$ percent prediction interval for $Y^*(x_0)$ will be wider than the corresponding $100(1-\alpha)$ percent confidence interval for $E(Y|x_0)$.
- **Interval length**: The length of both intervals clearly depends on the value of x_0. In fact, the standard error of $\widehat{Y}(x_0)$ 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.

It is sometimes desired to estimate $E(Y|x_0)$ or predict $Y^*(x_0)$ based on the fit of the model for values of x_0 outside the range of x values used in the experiment/study. This is called **extrapolation** and can be very dangerous. In order for our inferences to be valid, we must believe that the straight line relationship 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 6.2.5. (Continued Example 6.2.1). In our sewage sludge 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 mean moisture percentage for compressed pellets when the machine filtration rate is $x_0 = 150$ kg-DS/m/hr. In other words, if we were to repeat the experiment over and over again, each time using a filtration rate of $x_0 = 150$ kg-DS/m/hr, then $E(Y|x_0)$ denotes the mean value of Y (moisture percentage) that would be observed.

- $Y^*(x_0)$ denotes a possible value of Y for a single run of the machine when the filtration rate is set at $x_0 = 150$ 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:

$$\widehat{Y}(x_0 = 150) = 72.959 + 0.041(150) \approx 79.11361.$$

- 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 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 run of the experiment will be between 77.68 and 80.55 percent.

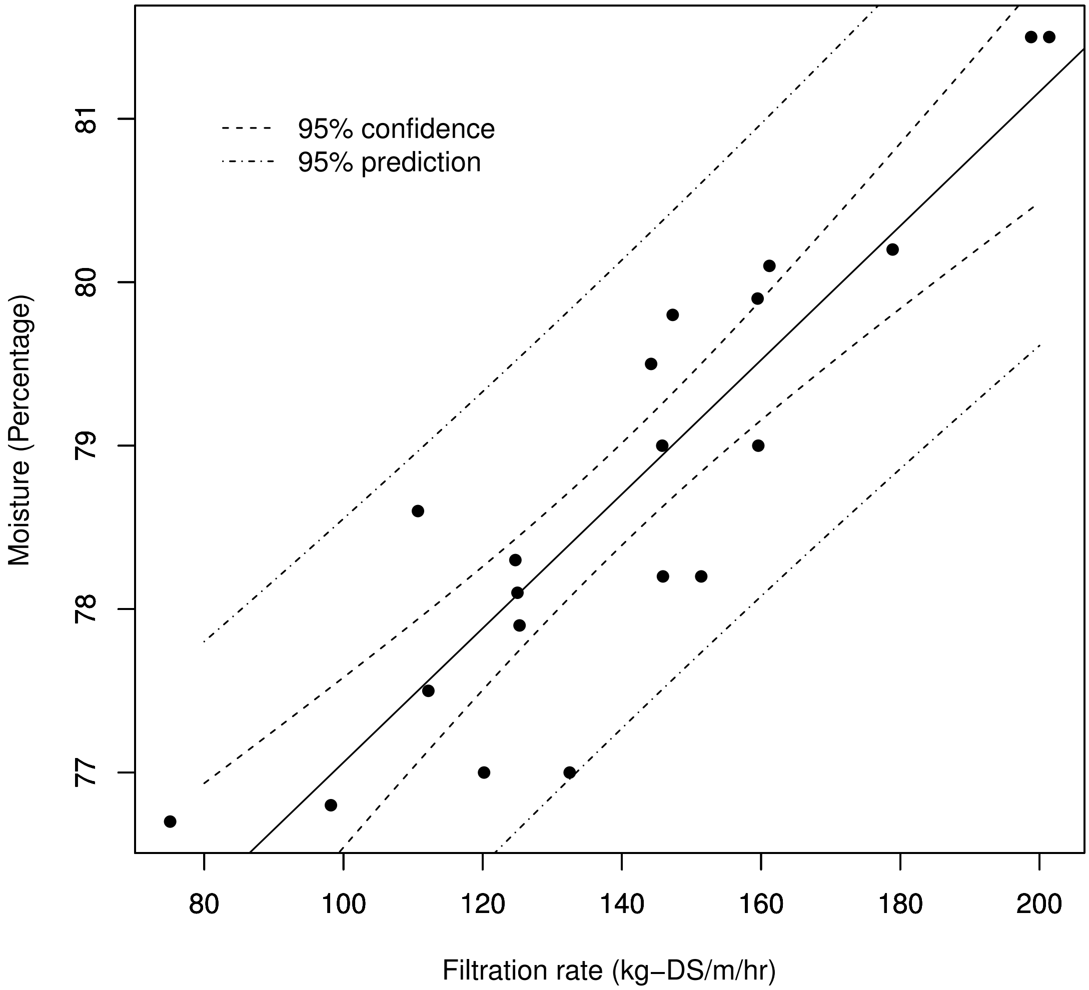


Figure 6.2.3: Scatterplot of pellet moisture Y as a function of machine filtration rate x, including the least squares regression line and ninety-five percent confidence/prediction bands.

6.3 Multiple linear regression

6.3.1 Introduction

PREVIEW: We have already considered the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim N(0, \sigma^2)$. We now extend this basic model to include multiple independent variables $x_1, x_2, ..., x_k$. This is much more realistic because, in practice, often Y depends on many different factors (i.e., not just one). Specifically, we consider models of the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. We call this a **multiple linear regression model**.

- There are now $p = k + 1$ regression parameters β_0, β_1, ..., β_k. These are unknown and are to be estimated with the observed data.
- Schematically, 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 $x_1, x_2, ..., x_k$.
- We continue to assume that $\epsilon_i \sim N(0, \sigma^2)$.
- We also assume that the independent variables $x_1, x_2, ..., x_k$ are fixed and measured without error.

PREVIEW: To fit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

we again use the method of **least squares**. Simple computing formulae for the least squares estimators are no longer available (as they were in simple linear regression). This is hardly a big deal because we will use computing to automate all analyses. For instructional purposes, it is advantageous to express multiple linear regression models in terms of matrices and vectors. This streamlines notation and makes the presentation easier.

6.3.2 Least square estimator

The notion of **least squares** is the same as it was in the simple linear regression model. To fit a multiple linear regression model, we want to find the values of $\beta_0, \beta_1, ..., \beta_k$ that minimize

$$Q(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})]^2.$$

This could be done by solving a system of linear equations,

$$\frac{\partial Q(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_j} = 0, \text{ for } j = 0, \ldots, k.$$

And the R package can help you find the solution. See the following example.

Example 6.3.1. 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, samples 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 several tasters. Data were collected on the following variables:

$Y =$ taste score (TASTE)
$x_1 =$ concentration of acetic acid (ACETIC)
$x_2 =$ concentration of hydrogen sulfide (H2S)
$x_3 =$ concentration of lactic acid (LACTIC).

Variables ACETIC and H2S were both measured on the log scale. The variable LACTIC has not been transformed. Table 6.5 contains concentrations of the various chemicals in $n = 30$ specimens of cheddar cheese and the observed taste score.

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 6.5: Cheese data. ACETIC, H2S, and LACTIC are independent variables. The response variable is TASTE.

MODEL: Researchers postulate that each of the three chemical composition variables x_1, x_2, and x_3 is important in describing the taste and consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. We now use R to fit this model using the method of least squares:

```
> fit = lm(taste˜acetic+h2s+lactic)
> summary(fit)

Call:
lm(formula = taste ˜ acetic + h2s + lactic)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
h2s           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

This output gives the values of the least squares estimates

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix} = \begin{pmatrix} -28.877 \\ 0.328 \\ 3.912 \\ 19.670 \end{pmatrix}.$$

Therefore, the fitted least squares regression model is

$$\widehat{Y} = -28.877 + 0.328x_1 + 3.912x_2 + 19.670x_3,$$

or, in other words,

$$\widehat{\texttt{TASTE}} = -28.877 + 0.328\ \texttt{ACETIC} + 3.912\ \texttt{H2S} + 19.670\ \texttt{LACTIC}.$$

6.3.3 Estimating the error variance

Terminology: Define the **residual sum of squares** by

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

In the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$,

$$MS_{res} = \frac{SS_{res}}{n-p}$$

is an **unbiased estimator** of σ^2, that is,

$$E(MS_{res}) = \sigma^2.$$

The quantity

$$\widehat{\sigma} = \sqrt{MS_{res}} = \sqrt{\frac{SS_{res}}{n-p}}$$

estimates σ and is called the **residual standard error**.

Example 6.3.2. For the cheese data in Example 6.3.1, we use R to calculate MS_{res}:

```
> fit = lm(taste˜acetic+h2s+lactic)
> summary(fit)
Call:
lm(formula = taste ˜ acetic + h2s + lactic)
Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
h2s           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Table 6.6: Analysis of variance table for linear regression.

Source	df	SS	MS	F
Regression	k	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{k}$	$F = \frac{MS_{reg}}{MS_{res}}$
Residual	$n-p$	SS_{res}	$MS_{res} = \frac{SS_{res}}{n-p}$	
Total	$n-1$	SS_{total}		

6.3.4 Analysis of variance for linear regression

IDENTITY: Algebraically, it can be shown that

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{SS_{total}} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}_{SS_{reg}} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{SS_{res}}.$$

- SS_{total} is the **total sum of squares**. SS_{total} is the numerator of the sample variance of $Y_1, Y_2, ..., Y_n$. It measures the total variation in the response data.

- SS_{reg} is the **regression sum of squares**. SS_{reg} measures the variation in the response data explained by the linear regression model.

- SS_{res} is the **residual sum of squares**. SS_{res} measures the variation in the response data not explained by the linear regression model.

ANOVA TABLE: We can combine all of this information to produce an **analysis of variance** (**ANOVA**) table. Such tables are standard in regression analysis.

- The **degrees of freedom** (df) add down.

  - SS_{total} can be viewed as a statistic that has "lost" a degree of freedom for having to estimate the overall mean of Y with the sample mean $\overline{Y}$. Recall that $n-1$ is our divisor in the sample variance S^2.

  - There are k degrees of freedom associated with SS_{reg} because there are k independent variables.

  - The degrees of freedom for SS_{res} can be thought of as the divisor needed to create an unbiased estimator of σ^2. Recall that

    $$MS_{res} = \frac{SS_{res}}{n-p} = \frac{SS_{res}}{n-k-1}$$

    is an unbiased estimator of σ^2

- The **sum of squares** (SS) also add down. This follows from the algebraic identity noted earlier.

- **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.

Coefficient of determination: Since

$$SS_{total} = SS_{reg} + SS_{res},$$

the proportion of the total variation in the data explained by the linear regression model is

$$R^2 = \frac{SS_{reg}}{SS_{total}} = 1 - \frac{SS_{res}}{SS_{total}}..$$

This statistic is called the **coefficient of determination**. Clearly,

$$0 \leq R^2 \leq 1.$$

The larger the R^2, the better the regression model explains the variability in the data.

IMPORTANT: It is critical to understand what R^2 does and does not measure. Its value is computed under the assumption that the multiple linear regression model **is correct** and assesses how much of the variation in the data may be attributed to that relationship rather than to inherent variation.

- If R^2 is small, it may be that there is a lot of random inherent variation in the data, so that, although the multiple linear regression model is reasonable, it can explain only so much of the observed overall variation.

- Alternatively, R^2 may be close to 1; e.g., in a simple linear regression model fit, but this may not be the best model. In fact, R^2 could be very "high," but ultimately not relevant because it assumes the simple linear regression model is correct. In reality, a better model may exist (e.g., a quadratic model, etc.).

- One draw back of R^2 is that, it can never decrease when a new x is added. Thus, it can be difficult to judge whether the increase is telling us anything useful about the new x. Instead, many regression users prefer to use an **adjusted** R^2 statistic:

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-p)}{SS_{total}/(n-1)}.$$

F statistics: The F statistic in the ANOVA table is used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus

$$H_a : \text{at least one of the } \beta_j \text{ is nonzero.}$$

In other words, F tests whether or not at least one of the independent variables $x_1, x_2, ..., x_k$ is important in describing the response Y. 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 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true,

$$F = \frac{MS_{reg}}{MS_{res}} \sim F(k, n - p).$$

Therefore, we can gauge the evidence against H_0 by comparing F to this distribution. Values of F far out in the (right) upper tail are evidence against H_0. R automatically produces the value of F and produces the corresponding p-value. Recall that small p-values are evidence against H_0 (the smaller the p-value, the more evidence).

Example 6.3.3. Example 6.3.1 (continued). For the cheese data, we fit the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. The ANOVA table, obtained using **R**, is shown below.

```
> fit = lm(taste˜acetic+h2s+lactic)
> summary(fit)
Call:
lm(formula = taste ˜ acetic + h2s + lactic)
Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
h2s           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

The F statistic is used to test

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$
$$\text{versus}$$
$$H_a: \text{at least one of the } \beta_j \text{ is nonzero.}$$

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. **Interpretation:** We conclude that at least one of the independent variables (`ACETIC`, `H2S`, `LACTIC`) is important in describing taste.

The coefficient of determination R^2 is 0.6518; i.e., 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.

6.3.5 Inference for individual regression parameters

IMPORTANCE: Consider our multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim N(0, \sigma^2)$. Confidence intervals and hypothesis tests for β_j can help us assess the importance of using the independent variable x_j in a model with the other independent variables. That is, inference regarding β_j is always **conditional** on the other variables being included in the model.

A $100(1-\alpha)$ **percent confidence interval** for β_j, for $j = 0, 1, 2, ..., k$,

$$[\widehat{\beta}_j \pm t_{n-p,\alpha/2}\widehat{\text{se}}(\widehat{\beta}_j)]$$

and hypothesis tests for

$$H_0: \beta_j = 0$$
$$\text{versus}$$
$$H_a: \beta_j \neq 0,$$

can be performed by examining the p-value output provided in R.

- If $H_0: \beta_j = 0$ is not rejected, then x_j is not important in describing Y in the presence of the other independent variables.

- If $H_0: \beta_j = 0$ is rejected, this means that x_j is important in describing Y even after including the effects of the other independent variables.

Hypothesis test for β_1: If our interest was to test, where $j = 0, 1, \ldots, k$

$$H_0 : \beta_j = 0$$

versus one of the following

$$H_a : \beta_j \neq 0$$
$$H_a : \beta_j > 0$$
$$H_a : \beta_j < 0$$

where β_{10} is a fixed value (often, $\beta_{10} = 0$), we would focus our attention on

$$t_0 = \frac{\widehat{\beta}_j}{\widehat{\text{se}}(\widehat{\beta}_j)}.$$

Critical value approach

Alternative hypothesis	Rejection Criterion
$H_a : \beta_j \neq 0$	$t_0 > t_{n-p,\alpha/2}$ or $t_0 < -t_{n-p,\alpha/2}$
$H_a : \beta_j > 0$	$t_0 > t_{n-p,\alpha}$
$H_a : \beta_j < 0$	$t_0 < -t_{n-p,\alpha}$

P-value approach

Alternative hypothesis	Rejection Criterion
$H_a : \beta_j \neq 0$	P-value $< \alpha$
$H_a : \beta_j > 0$	P-value $< \alpha$
$H_a : \beta_j < 0$	P-value $< \alpha$

Confidence interval approach

Alternative hypothesis	Reject Criterion
$H_a : \beta_j \neq 0$	$0 \notin \left[\widehat{\beta}_j \pm t_{n-p,\alpha/2}\widehat{\text{se}}(\widehat{\beta}_j)\right]$
$H_a : \beta_j > 0$	$0 < \widehat{\beta}_j - t_{n-p,\alpha}\widehat{\text{se}}(\widehat{\beta}_j)$
$H_a : \beta_j < 0$	$0 > \widehat{\beta}_j + t_{n-p,\alpha}\widehat{\text{se}}(\widehat{\beta}_j)$

Example 6.3.4. For the cheese data, Example 6.3.1 continued,

```
> fit = lm(taste˜acetic+h2s+lactic)
> summary(fit)

Call:
lm(formula = taste ˜ acetic + h2s + lactic)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
h2s           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

OUTPUT: The `Estimate` output gives the values of the least squares estimates:

$$\widehat{\beta}_0 \approx -28.877, \quad \widehat{\beta}_1 \approx 0.328, \quad \widehat{\beta}_2 \approx 3.912, \quad \widehat{\beta}_3 \approx 19.670.$$

The `Std.Error` output gives

$$\begin{aligned} \widehat{\text{se}}(\widehat{\beta}_0) &= 19.735 \\ \widehat{\text{se}}(\widehat{\beta}_0) &= 4.460 \\ \widehat{\text{se}}(\widehat{\beta}_0) &= 1.248 \\ \widehat{\text{se}}(\widehat{\beta}_0) &= 8.629 \end{aligned}$$

To test

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_0 : \beta_j \neq 0,$$

for $j = 0, 1, 2, 3,$ via the confidence interval approach, we have 95% confidence intervals for the regression parameters β_0, β_1, β_2, and β_3, respectively, are, (since $t_{n-p,\alpha/2} = t_{26,.025} = 2.056$),

$$\begin{aligned}
\widehat{\beta}_0 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_0) &\implies -28.877 \pm 2.056(19.735) \implies (-69.45, 11.70) \\
\widehat{\beta}_1 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_1) &\implies 0.328 \pm 2.056(4.460) \implies (-8.84, 9.50) \\
\widehat{\beta}_2 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_2) &\implies 3.912 \pm 2.056(1.248) \implies (1.35, 6.48) \\
\widehat{\beta}_3 \pm t_{26,0.025}\widehat{\text{se}}(\widehat{\beta}_3) &\implies 19.670 \pm 2.056(8.629) \implies (1.93, 37.41).
\end{aligned}$$

The conclusions reached from interpreting these intervals are the same as those reached using the hypothesis test p-values. Note that the β_2 and β_3 intervals do not include zero. Those for β_0 and β_1 do.

The `t value` output gives the t statistics

$$\begin{aligned}
t_0 &= -1.463 \\
t_0 &= 0.074 \\
t_0 &= 3.133 \\
t_0 &= 2.279
\end{aligned}$$

To test $H_0 : \beta_i = 0$ versus $H_0 : \beta_i \neq 0$, for $i = 0, 1, 2, 3$, critical value approach provides the rejection region as

$$t > t_{n-p,\alpha/2} = t_{26,.025} = 2.056 \quad \text{or} \quad t < -t_{n-p,\alpha/2} = t_{26,.025} = -2.056$$

These tests can also be done use the P-values in `Pr(>|t|)` output. At the $\alpha = 0.05$ level,

- we do not reject $H_0 : \beta_0 = 0$ (p-value = 0.155). **Interpretation:** In the model which includes all three independent variables, the intercept term β_0 is not statistically different from zero.

- we do not reject $H_0 : \beta_1 = 0$ (p-value = 0.942). **Interpretation:** `ACETIC` does not significantly add to a model that includes `H2S` and `LACTIC`.

- we reject $H_0 : \beta_2 = 0$ (p-value = 0.004). **Interpretation:** `H2S` does significantly add to a model that includes `ACETIC` and `LACTIC`.

- we reject $H_0 : \beta_3 = 0$ (p-value = 0.031). **Interpretation:** `LACTIC` does significantly add to a model that includes `ACETIC` and `H2S`.

6.3.6 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 the simple linear regression case, the former is called a **confidence interval** (since it is for a mean response) and the latter is called a **prediction interval** (since it is for a new random variable).

CHEESE DATA: Suppose that 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 $\mathbf{x}_0 = (5.5, 6.0, 1.4)$. 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

$$\begin{aligned}\widehat{Y}(\mathbf{x}_0) &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{10} + \widehat{\beta}_2 x_{20} + \widehat{\beta}_3 x_{30} \\ &= -28.877 + 0.328(5.5) + 3.912(6.0) + 19.670(1.4) \approx 23.936.\end{aligned}$$

- 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 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 cheese specimen will be between 2.75 and 45.12.

6.4 Model diagnostics (residual analysis)

IMPORTANCE: We now discuss certain diagnostic techniques for linear regression. 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.

MODEL ASSUMPTIONS: 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, from the fit of the model, we can calculate the residuals

$$e_i = Y_i - \widehat{Y}_i,$$

where the ith fitted value

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}.$$

We can think of the residuals $e_1, e_2, ..., e_n$ as "proxies" for the error terms $\epsilon_1, \epsilon_2, ..., \epsilon_n$, and, therefore, we can use the residuals to check our model assumptions instead.

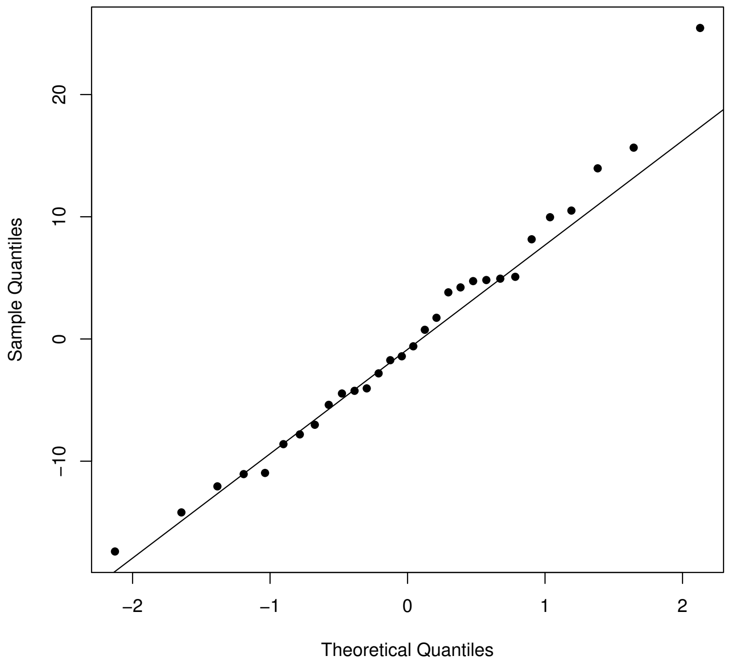


Figure 6.4.4: Cheese data. Normal qq-plot of the least squares residuals.

QQ PLOT FOR NORMALITY: To check the normality assumption (for the errors) in linear regression, it is common to display the qq-plot of the 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 6.4.4 supports the normality assumption for the errors in the multiple linear regression model for the cheese data.

RESIDUAL PLOT: By the phrase "residual plot," I mean the plot of the residuals (on the vertical axis) versus the predicted values (on the horizontal axis). This plot is simply the scatterplot of the residuals and the predicted values.

- Advanced linear model arguments show that if the model does a good job at describing the data, then the residuals and fitted values are independent.
- This means that a plot of the residuals versus the fitted values should reveal no noticeable patterns; that is, the plot should appear to be random in nature (e.g., "a random scatter of points").
- On the other hand, if there are definite (non-random) patterns in the residual plot, this suggests that the model is inadequate in some way or it could point to a violation in the model assumptions.
- The plot in Figure 6.4.5 does not suggest any obvious model inadequacies! It looks completely random in appearance.

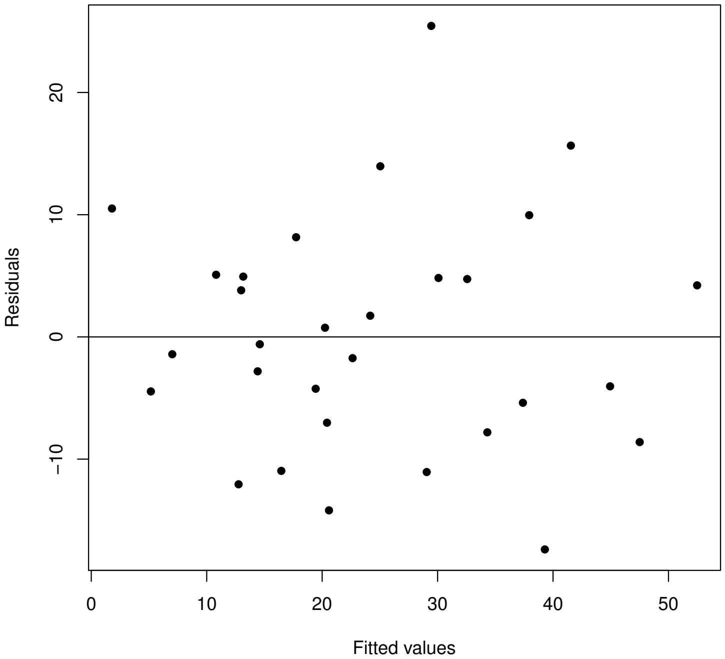


Figure 6.4.5: Cheese data. Residual plot for the multiple linear regression model fit. A horizontal line at zero has been added.

COMMON VIOLATIONS: Although there are many ways to violate the statistical assumptions associated with linear regression, the most common violations are

- non-constant variance (heteroscedasticity)

- misspecifying the true regression function
- correlated observations over time.

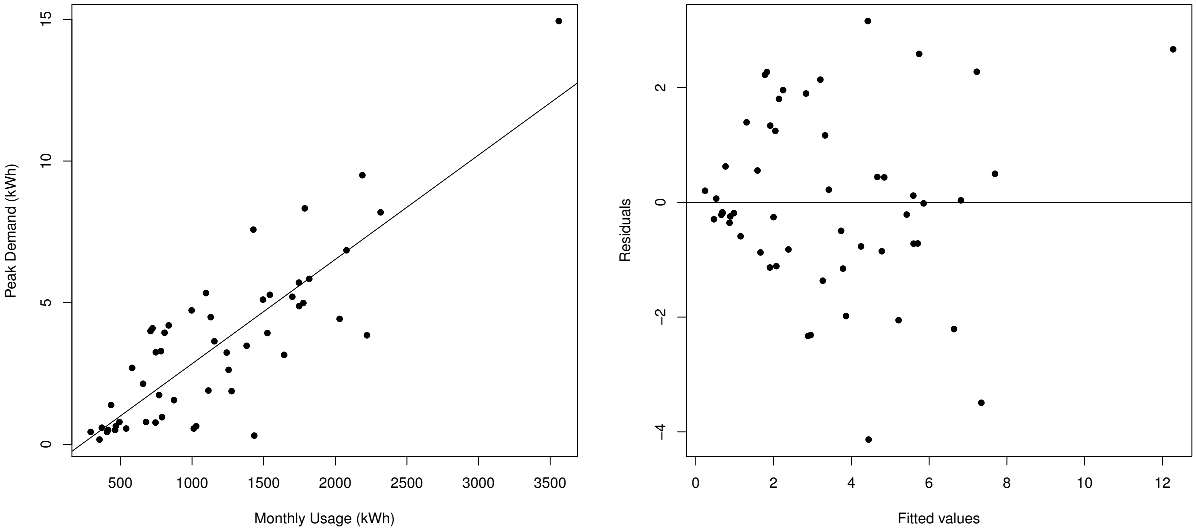


Figure 6.4.6: Electricity data. Left: Scatterplot of peak demand (Y, measured in kWh) versus monthly usage (x, measured in kWh) with least squares simple linear regression line superimposed. Right: Residual plot for the simple linear regression model fit.

Example 6.3. An electric company is interested in modeling peak hour electricity demand (Y) as a function of total monthly energy usage (x). This is important for planning purposes because the generating system must be large enough to meet the maximum demand imposed by customers. Data for $n = 53$ residential customers for a given month are shown in Figure 6.4.6.

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: A common course of action to handle non-constant variance is to apply a transformation to the response variable Y. Common transformations are logarithmic ($\ln Y$), square-root ($\sqrt{Y}$), and inverse ($1/Y$).

ELECTRICITY DATA: A square root transformation is commonly applied to address non-constant variance. Consider the simple linear regression model

$$W_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., 53$, where $W_i = \sqrt{Y_i}$. It is straightforward to fit this transformed model in R as before. We simply regress W on x (instead of regressing Y on x).

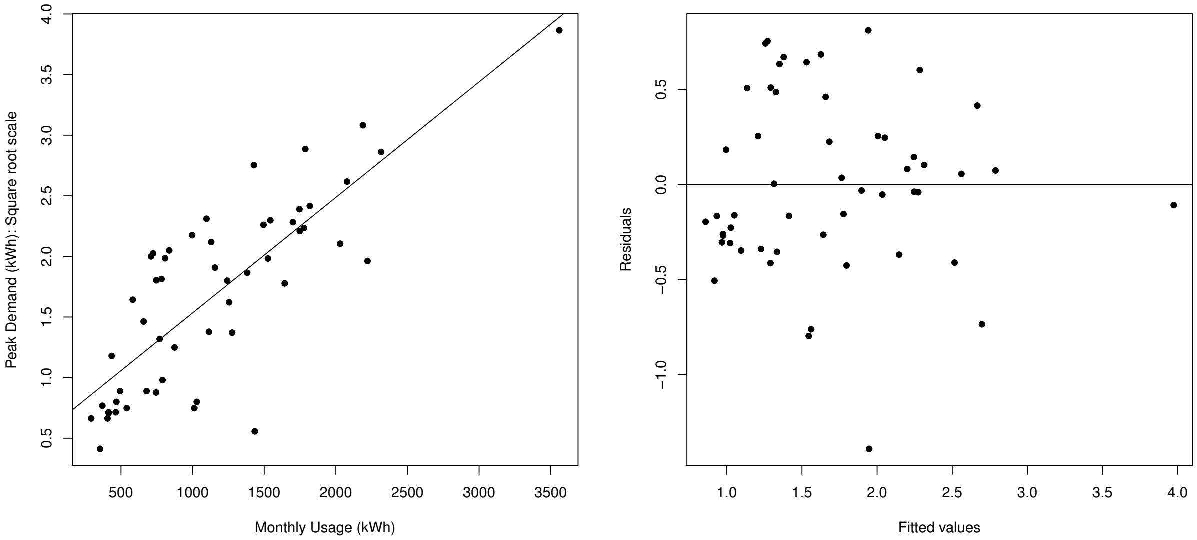


Figure 6.4.7: Electricity data. Left: Scatterplot of the square root of peak demand ($\sqrt{Y}$) versus monthly usage (x, measured in kWh) with the least squares simple linear regression line superimposed. Right: Residual plot for the simple linear regression model fit with transformed response.

```
> fit.2 = lm(sqrt(peak.demand) ~ monthly.usage)
> fit.2
Coefficients:
  (Intercept)  monthly.usage
     0.580831       0.000953
```

ANALYSIS: Figure 6.4.7 above shows the scatterplot (left) and the residual plot (right) from fitting the transformed model. The "fanning out" shape that we saw previously (in the untransformed model) is now largely absent. The fitted transformed model is

$$\widehat{W} = 0.580831 + 0.000953x,$$

or, in other words,

$$\sqrt{\texttt{Peak demand}} = 0.580831 + 0.000953 \ \texttt{Monthly usage}.$$

Further analyses can be carried out with the transformed model; e.g., testing whether peak demand (on the square root scale) is linearly related to monthly usage, etc.

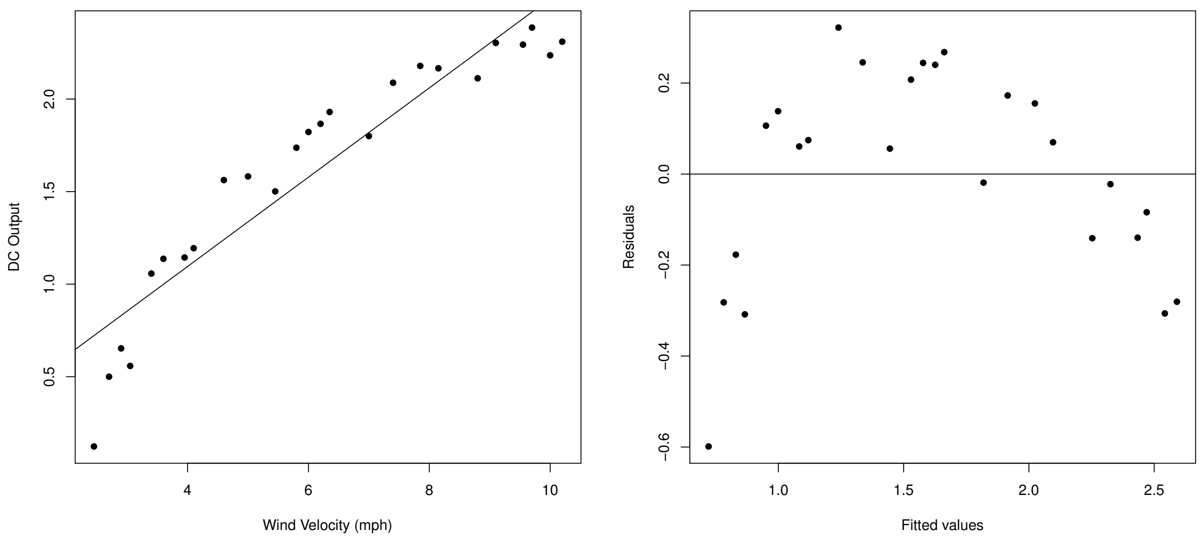


Figure 6.4.8: Windmill data. Left: Scatterplot of DC output Y versus wind velocity (x, measured in mph) with least squares simple linear regression line superimposed. Right: Residual plot for the simple linear regression model fit.

Example 6.4. A research engineer is investigating the use of a windmill to generate electricity. He has collected data on the direct current (DC) output Y from his windmill and the corresponding wind velocity (x, measured in mph). Data for $n = 25$ observation pairs are shown in Figure 6.4.8.

Problem: There is a clear quadratic relationship between DC output and wind velocity, so a simple linear regression model fit (as shown above) is inappropriate. The residual plot shows a pronounced quadratic pattern; this pattern is not accounted for in fitting a straight line model.

Remedy: Fit a multiple linear regression model with two independent variables: wind velocity x and its square x^2, that is, consider the quadratic regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

for $i = 1, 2, ..., 25$. It is straightforward to fit a quadratic model in R. We simply regress Y on 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
```

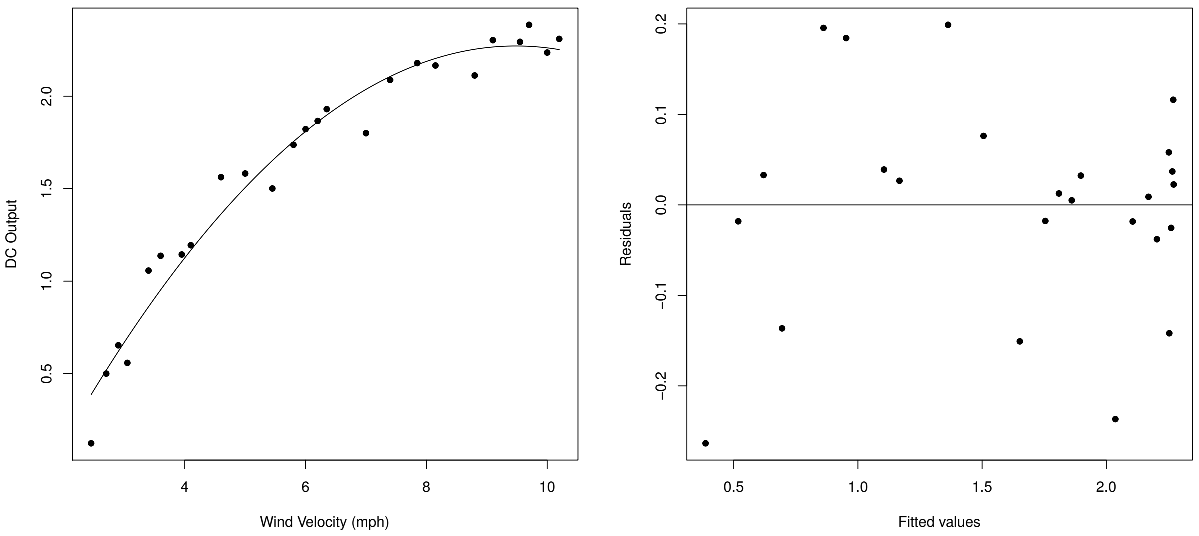


Figure 6.4.9: Windmill data. Scatterplot of DC output Y versus wind velocity (x, measured in mph) with least squares quadratic regression curve superimposed. Right: Residual plot for the quadratic regression model fit.

The fitted quadratic regression model is

$$\widehat{Y} = -1.15590 + 0.72294x - 0.03812x^2$$

or, in other words,

$$\widehat{\texttt{DC output}} = -1.15590 + 0.72294\ \texttt{Wind.velocity} - 0.03812\ (\texttt{Wind.velocity})^2.$$

Note that the residual plot from the quadratic model fit, shown above, now looks quite good. The quadratic trend has disappeared (because the model now incorporates it).

Example 6.5. The data in Figure 6.4.10 (left) are temperature readings (in deg C) on land-air average temperature anomalies, collected once per year from 1900-1997. To emphasize that the data are collected over time, I have used straight lines to connect the observations; this is called a **time series plot**.

- Unfortunately, it is all too common that people fit linear regression models to time series data and then blindly use them for prediction purposes.

- It takes neither a meteorologist nor an engineering degree to know that temperature observations collected over time are probably correlated. Not surprisingly, residuals from a simple linear regression display clear correlation over time.

- Regression techniques (as we have learned in this chapter) are generally not appropriate when analyzing time series data for this reason. More advanced modeling techniques are needed.

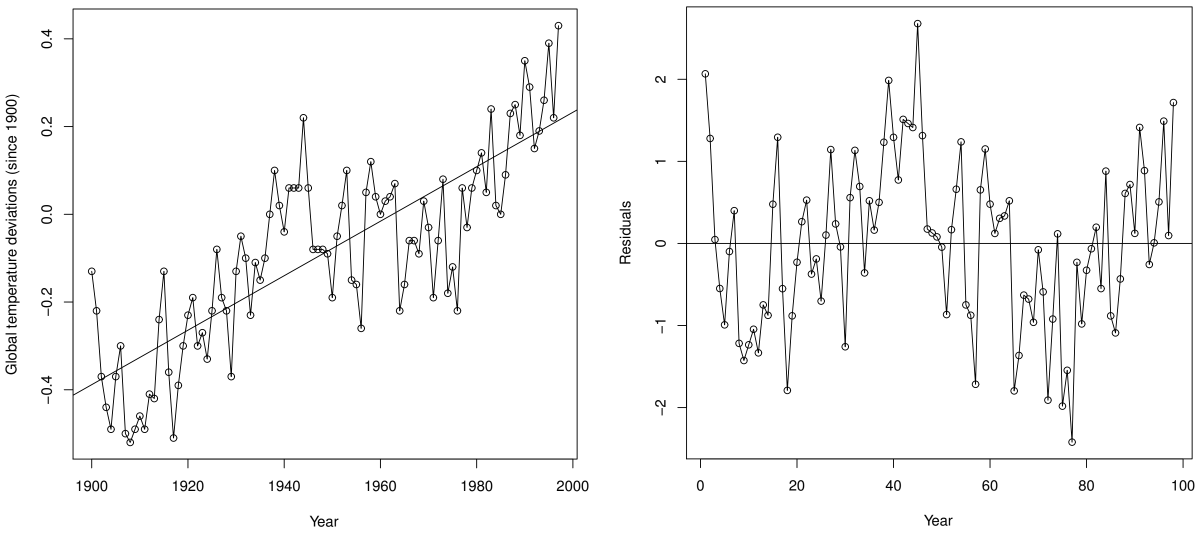


Figure 6.4.10: Global temperature data. Left: Time series plot of the temperature Y measured one time per year. The independent variable x is year, measured as 1900, 1901, ..., 1997. A simple linear regression model fit has been superimposed. Right: Residual plot from the simple linear regression model fit.