Elementary Statistics

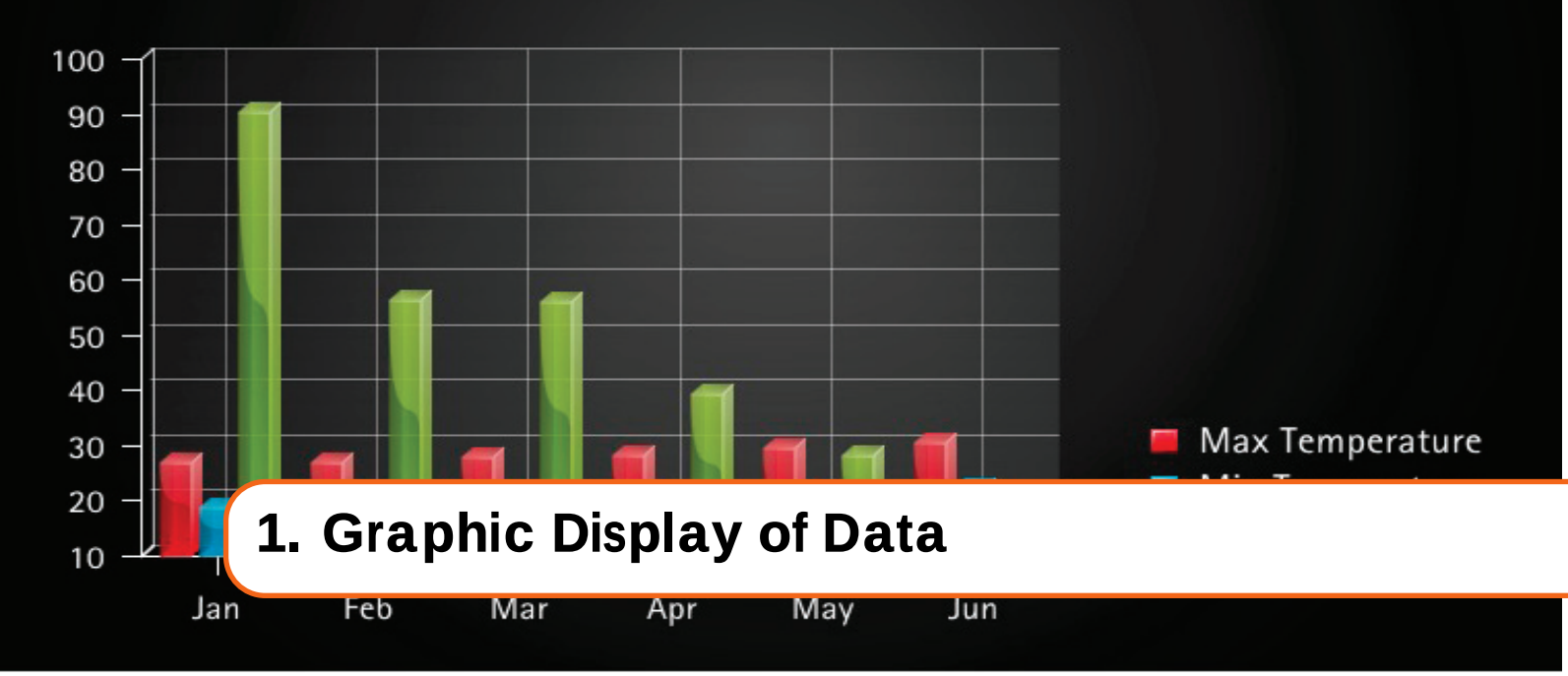
K. Krishnamoorthy

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

First printing, March 2013

Contents



1. Graphic Display of Data

Outline

1. Some Basic Terminologies
2. Types of Data
3. Displaying and Describing Categorical Data
4. Displaying and Summarizing Quantitative Data

1.1 Some Basic Terminologies

Population is a collection of objects in the context of a study, survey or an experiment.

Examples

1. Suppose we want to estimate the percentage of voters who supports a presidential candidate; here the population is all registered voters in the USA.
2. A drug manufacturer wants to study the effectiveness of his new drug for treating jaundice. Here the population consists of all patients with jaundice.
3. Suppose a light bulb manufacturer wants to estimate the percentage of defective bulbs that were manufactured in a day. Here, the population consists of all bulbs that were produced in that day.

Sample is a subset of the population. Usually a sample is selected so that it represents the entire population.

As an example, you want to estimate the chances that a candidate will win an election in a constituency. The sample should consist of set of randomly selected registered voters in that constituency.

Variable is a property or characteristic of individuals in the population; called variable because it varies from individuals to individuals.

Examples: height, weight, daily average temperature, salary, age, etc.

Constant is something that does not change over individuals; something that is fixed.
Examples: No. of days in a week; no. of hours in a day; the value of

$$\pi = \frac{\text{circumference of a circle}}{\text{diameter}}.$$

This value does not change from one circle to another.

1.2 Types of Data

Definition 1.1 — Data are collection of measurements or observations of a characteristic from a set of individuals in a population.

Examples

1. Eye colors of 10 students:
   Brown, Black, Brown, Blue, Cyan, Gray, Blue, Black, Black, Brown
2. Gender of 5 students: F, M, M, F, M
3. Heights of 7 students; 62, 66, 64, 72, 60, 65, 69 (in inches)
4. Salaries of 20 men: 62, 63, 49, 65, 32, et. (in $1,000)
5. Religion of people: C, C, M, C, C, M, O, C, H, etc. (C = Christian; M = Muslim; H = Hindu; O = others).

Types of Data: In general, there are two types of data, namely, qualitative data (also called categorical) and quantitative data.

Definition 1.2 — Qualitative data are the measurements of a characteristic or property of individuals that enable us to categorize individuals into different groups or categories.

Examples: Eye color, religion, gender, race, political affiliation, colors of cars, etc.
Another example of qualitative data. A sample of 10 alternators was tested and the following data were recorded.

1	2	3	4	5	6	7	8	9	10
D	N	R	R	D	D	N	N	N	R

N = nondefective; D = defective; R = repairable.
Qualitative data come in two types, namely, nominal and ordinal.

Definition 1.3 — Nominal data involves names, labels and categories. For example, race, religion, gender, brand names, etc. Nominal data can not be ordered or compared.

Definition 1.4 — Ordinal data can be arranged in an order according to the characteristic or quality of interest; however, the difference between any two data can not be quantified and is meaningless.

As an example, five different wines were rated on a 10-point scale as follows.

$$9, \quad 8.5, \quad 8, \quad 8, \quad 7.5.$$

The wine with rating of 9 may be better than the one with rating of 8, but the difference $9 - 8$ is meaningless. We can simply say, the wines with higher ratings are better than the ones with the lower ratings.

■ **Example 1.1** Types of data:

1. Makes of 10 different cars - nominal

2. Ranks of college teachers - ordinal (professor, associate professor, assistant professor, instructor)
3. Political affiliation of 20 voters – nominal
4. Ranks of army personnel - ordinal
5. Race of people – nominal
6. Nationality of 100 students – nominal

■

Definition 1.5 — Quantitative variable is a variable that can be measured using scales or other devices, or some methods. Quantitative data can be compared with one another, and a meaningful comparison between two data points can be made.

Examples: height, weight, age, salary, temperature, time, etc.
Quantitative variable is further classified into two types, namely, continuous or discrete.

Definition 1.6 — Discrete data are collected by counting something. Examples for discrete variables are,

1. the number of days a sample of 15 students missed a STAT class in a semester;
2. the number of auto accidents per week in the past 20 weeks;
3. the number of people per household in a sample of 100 households.

Definition 1.7 — Continuous variable assumes any value in some intervals. For example, time, height, weight, salary, area, volume, etc.

■ **Example 1.2** Identify the following variables as qualitative or quantitative.

a. The number of siblings of a person.
b. The distance a person commutes to work.
c. The number of rainy days a month.
d. The number of patients admitted in a hospital on a given day.
e. The names of street in a town.
f. The colors of cars parked on a street.
g. The amount of water used by a household in a month.

■

■ **Example 1.3** Classify the following data as nominal, ordinal, continuous or discrete.

a. The number of people per household in a town.
b. The PIN numbers of account holders in a bank.
c. The zip codes in a city.
d. The sizes of T-shirts.
e. The monthly salaries of workers in a factory.
f. The average monthly rainfall in a city for the past 10 years.
g. Religious affiliation of people.
h. The ranking of university teachers.
i. Volumes of refrigerators.

■

1.3 Displaying and Describing Categorical Data

Graphs such as bar chart and pie chart help us to understand the distribution of qualitative data.

■ **Example 1.4** A sample of 25 army inductees were given a blood test to determine their blood types. The results are

A	O	B	A	AB	B	O	B	O	A
B	B	O	O	O	AB	AB	A	O	B
O	B	O	AB	A					

a. What is the variable of interest and what type of data does it yield?
b. What is the population?
c. Summarize the data numerically.
d. Summarize the data graphically.
e. What does the bar chart indicate?

Solution:

a. blood type; qualitative or categorical.
b. All army inductees from which the sample was selected.
c. The data can be summarized numerically as follows.

Blood Type	Count	Percent
A	5	$\frac{5}{25} \times 100 = 20\%$
B	7	$\frac{7}{25} \times 100 = 28\%$
AB	4	$\frac{4}{25} \times 100 = 16\%$
O	9	$\frac{9}{25} \times 100 = 36\%$

d. The bar chart (left) and the pie chart (right) are given below.

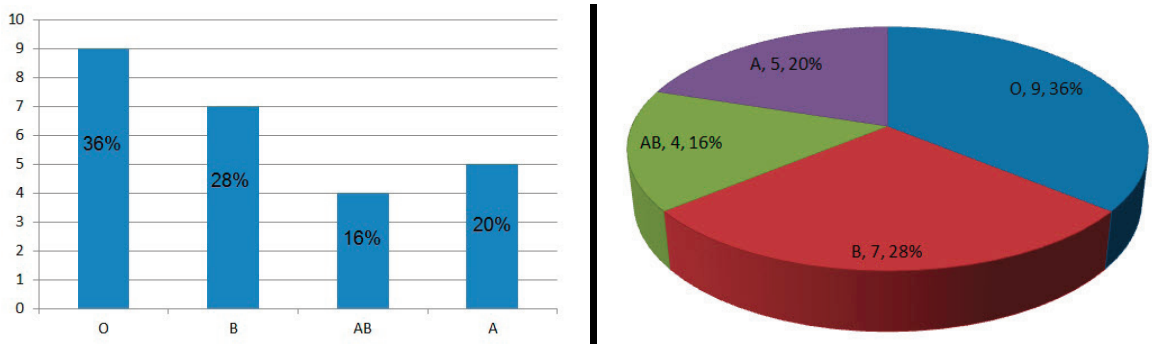


Figure 1.1: Bar chart and Pie chart for blood groups

e. The bar chart indicates that most of the inductees are with blood type O, followed by B then A and last AB.

■

■ **Example 1.5** Table 1.1 provides blood types for four ethnic groups in the US population. Notice that O positive is the most common blood type, and not all ethnic groups have the same mix of these blood types. Hispanic people, for example, have a relatively high number of O's, while Asian people have a relatively high number of B's.

a. What is the variable of interest and what type of data does it yield?
b. What is the population?
c. Summarize the data for the Hispanic group graphically.
d. What does the bar chart indicate?

Solution:

a. blood type; qualitative or categorical.

Table 1.1: Blood types of US population for four ethnic groups

	O	A	B	AB
Caucasians	45%	40%	11%	4%
African-American	51%	26%	19%	4%
Hispanic	57%	31%	10%	2%
Asian	40%	27.5%	25.4%	7.1%

b. All people in the US who belong to one of the four ethnic groups.
c. The bar chart is given below.

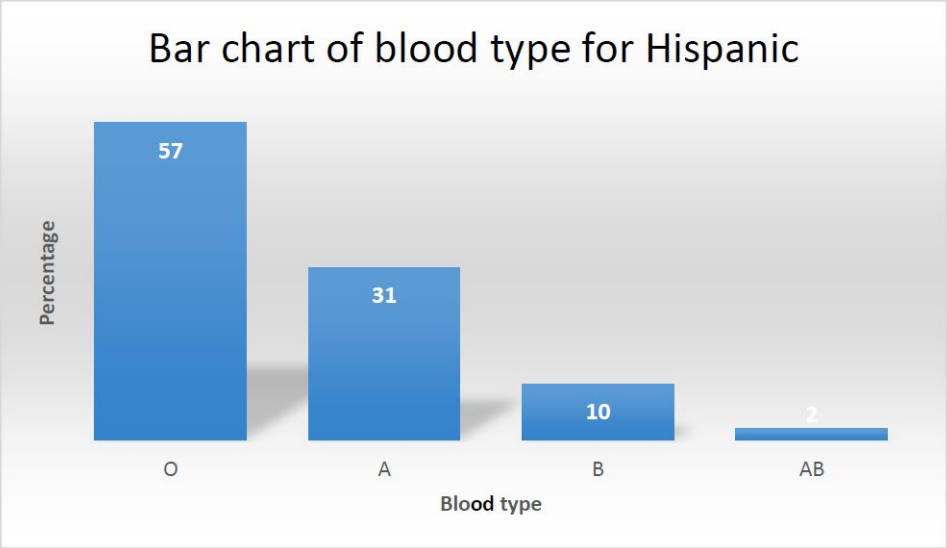


Figure 1.2: Bar chart of blood types among Hispanics

d. The bar chart indicates that most of the Hispanics are with blood type O, followed by B then A and last AB.

■

■ **Example 1.6** For blood types data in Table 1.1, construct histograms including all four groups, and describe.

Solution: The histogram for all four groups is given in Figure 1.3. The histograms for all four ethnic categories are similar in the sense that most of the people with blood type O, followed by A and B, and the last AB. The percentage of Hispanics with blood type O is the largest among all groups. Among all four groups, the percentage of Asians with blood type AB is the largest.

■

1.4 Displaying and Summarizing Quantitative Data

We here see two graphical methods of displaying quantitative data; stem-leaf and histogram. Stem-leaf display used for small data sets, and histogram is used for large data sets.

Stem-Leaf display is an arrangement of numbers in an easily readable form.

■ **Example 1.7** The following data represent the pulse rates of 24 women.

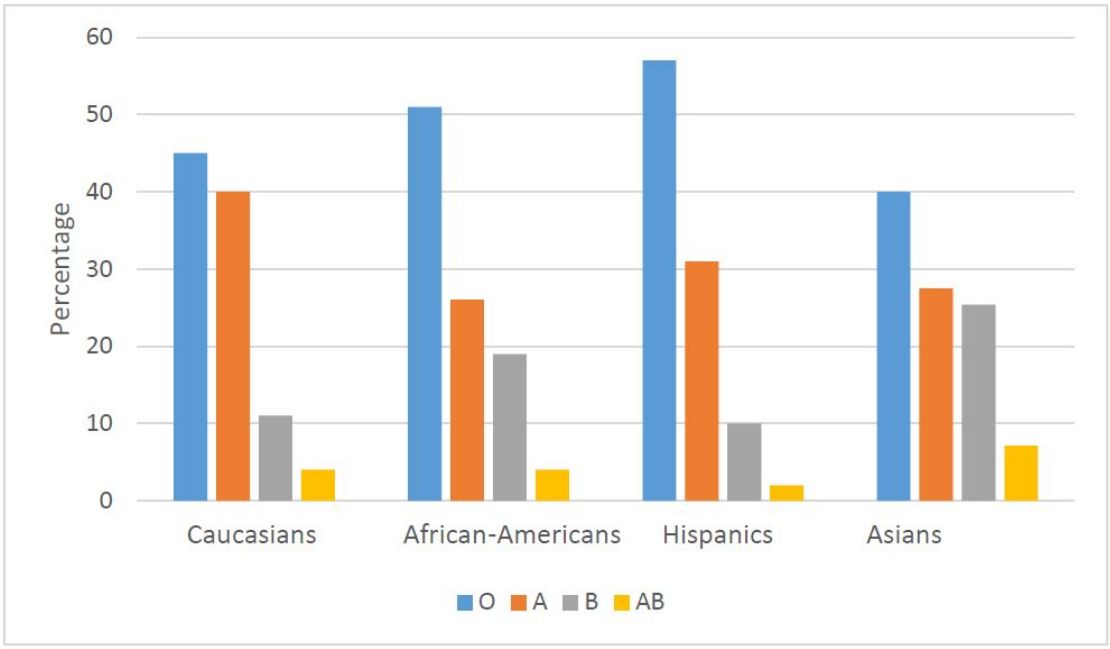


Figure 1.3: Bar chart of blood types among Hispanics

56	88	80	60	68	72	76	64	80	84	76	72	80	64	68
76	72	68	64	72	76	80	64	84	68					

a. Construct stem-leaf display for the above data.

stem-leaf of pulse rate data

5	6
6	0 4 4 4 4
6	8 8 8 8
7	2 2 2 2
7	6 6 6 6
8	0 0 0 0 4 4
8	8

b. What % of women have pulse rates no more than 70?

c. On what stem most of the pulse rates occur?

d. Find the percentage women with pulse rate at least 80.

■

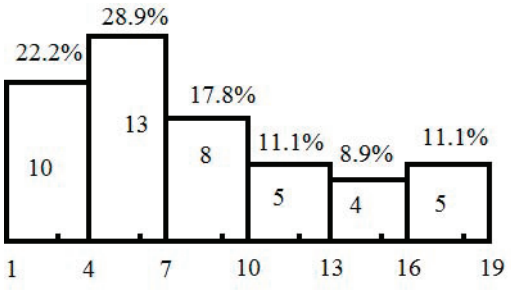
■ **Example 1.8** **(travel to work data)** The following data give the number of miles traveled to work by 45 employees of a large department store each day.

1	18	4	9	9	2	7	16	11	18	6	3	4	1	8
12	15	8	9	4	13	4	6	10	7	6	1	18	11	3
9	14	5	4	2	5	5	2	10	6	2	14	2	17	5

The following table gives percentage of data in each of six classes. Note that the percent $= \frac{\text{counts}}{\text{total}} \times 100$

a. Draw the histogram for the travel to work data.

classes	counts	percent	mid-point
$[1,4)$	10	22.2%	2.5
$[4,7)$	13	28.9%	5.5
$[7,10)$	8	17.8%	8.5
$[10,13)$	5	11.1%	11.5
$[13,16)$	4	8.9%	14.5
$[16,19)$	5	11.1%	17.5
Total	45	100	



b. What percentage of workers travel 4 to 7 miles? [Ans. 28.9%]
c. What percentage workers travel 10 or more miles? [Ans. 31.1%]
d. What is the modal class? [Ans. 4 – 7]
e. Describe the shape of the histogram [Ans. unimodal and right skewed]

■

TI Calc

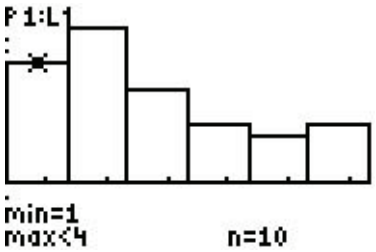
Constructing Histogram Using TI 83/84:

Entering data

1. Select [STAT]
2. Select [EDIT] and then [1]; this is default, and so press [ENTER]
3. Enter the data one by one by pressing [ENTER] after each data value.
4. After entering the last datum, press [2nd] and [QUIT]

To draw histogram

1. Select [STAT PLOT] by pressing [2nd] and [STAT PLOT]
2. Select Plot1 [ON] by moving the cursor over [ON] and then pressing [Enter]
3. Select histogram icon by moving the cursor over the icon, and pressing [Enter]
4. Enter the List where the data are stored. For example, if the data are in *L*1, press [2nd] and L1.
5. Select [WINDOW]
6. Set Xmin = 1; Xmax = 19; Xscl = 3; Ymin = -5; Ymax = 15; Yscl = 1; Xres = 1
7. Press [GRAPH]
8. Press [TRACE] to find the frequency in the first bar, and then use the navigation button to find the frequencies in other bars.



Note: Xscal determines the number of classes (bars). This number must be chosen so that the number of bars is 5 to 9; too few bars or too many bars makes histogram less informative.

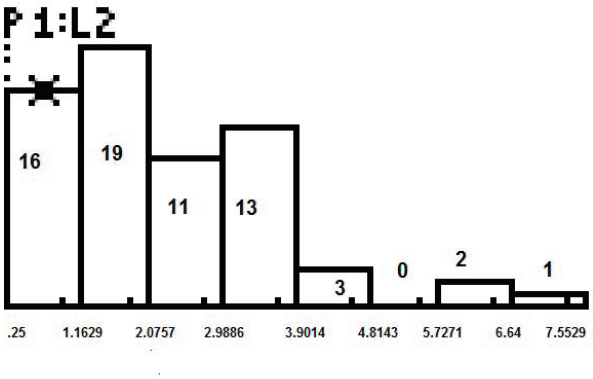
An alternative way to draw histogram

1. Select [STAT PLOT] by pressing [2nd] and [STAT PLOT]
2. Select Plot1 [ON] by moving the cursor over [ON] and then pressing [Enter]
3. Select histogram icon by moving the cursor over the icon, and pressing [Enter]
4. Enter the List where the data are stored.
5. Select [ZOOM] and then [ZOOMSTAT]
6. Press [TRACE] to find the frequency in the first bar, and then use the navigation button to find the frequencies in other bars.

■ **Example 1.9** The data below (page 48 in the book) represent particulate matter emissions (gram per gallon of fuel) for 65 vehicles. Particulate matter is a form of pollution that has been associated with respiratory disease.

```
1.5  0.87  1.12  1.25  3.46  1.11  1.12  0.88  1.29  0.94  0.64  1.31
2.49 1.48  1.06  1.11  2.15  0.86  1.81  1.47  1.24  1.63  2.14  6.64
4.04 2.48  1.4   1.37  1.81  1.14  1.63  3.67  0.55  2.67  2.63  3.03
1.23 1.04  1.63  3.12  2.37  2.12  2.68  1.17  3.34  3.79  1.28  2.1
6.55 1.18  3.06  0.48  0.25  0.53  3.36  3.47  2.74  1.88  5.94  4.24
3.52 3.59  3.1   3.33  4.5
```

a. Draw the histogram.
Using [ZOOM] and [ZOOMSTAT], we get



b. What percentage of particulate matter emissions is greater than 2 gram/gallon?
c. What percentage of particulate matter emissions fall in [.25, 4.82]?
d. What is the modal class?
e. Describe the shape of the histogram.

■

■ **Example 1.10** The following bar chart in Figure 1.4 gives the household size distributions in the UK for the year 2103. This bar chart represents real frequency for each class (here, the number of people), and is referred to as the frequency histogram.

a. Construct the relative frequency distribution for the household size data.

b. What is the modal class?

c. Describe the shape of the histogram.

d. What percentage of households having three or more people?

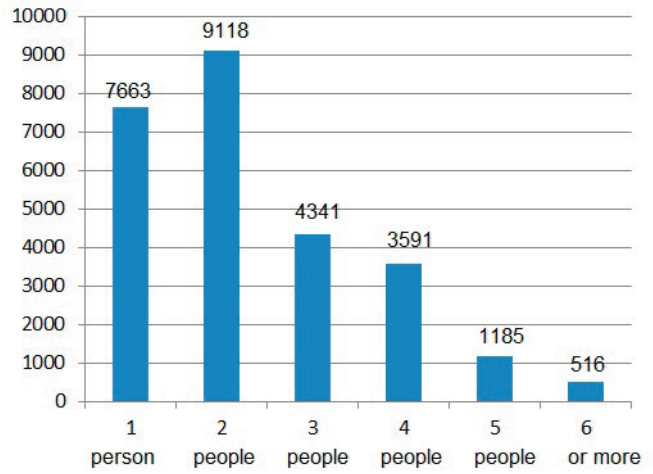


Figure 1.4: Histogram of frequency distribution of household size (in 1,000)

Solution:

a. For each class the relative frequency is

$$\frac{\text{frequency in the class}}{\text{total frequency}}.$$

These relative frequencies multiplied by 100 are:

$$\frac{7663}{26414} \times 100 = 29 \text{ for the first bar.}$$

For other bars, they are 34.5, 16.4, 13.6, 4.5, 2. The histogram of these relative frequencies is given in Figure 1.5.

b. modal class is 2.

c. unimodal and right-skewed

d. $\frac{4341+3591+1185+516}{26414} \approx 36.5$. This also means that in about $100 - 36.5 = 63.5$ percent of households only one or two people are living.

■

■ **Example 1.11** The data in Table 1.2 represent the number of earthquakes with magnitude range 6 - 7.9 in the USA from 1970 - 2012[1].

a. Construct a histogram for the data in Table 1.2.

b. Describe the shape of the histogram.

c. What is the modal class?

d. Find the percentage of years with no more than 10 earthquakes.

[1] Source: USGS

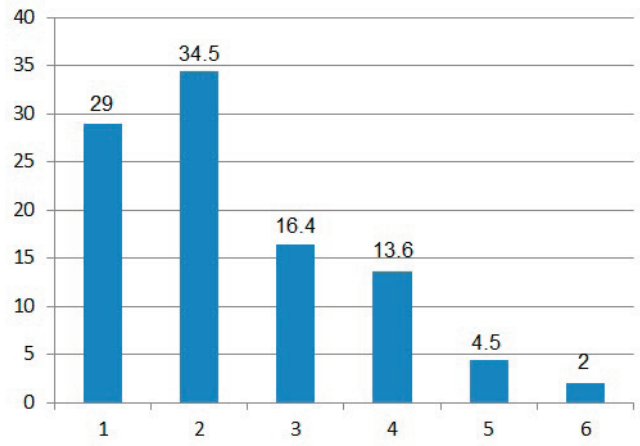


Figure 1.5: Histogram of frequency distribution of household size

Table 1.2: Number of earthquakes with magnitude range 6 - 7.9 in the USA, 1970-2012

Year	Number of earthquakes	Year	Number of earthquakes	Year	Number of earthquakes
1970	4	1985	5	2000	6
1971	8	1986	11	2001	6
1972	5	1987	13	2002	5
1973	7	1988	3	2003	9
1974	5	1989	5	2004	2
1975	12	1990	2	2005	5
1976	7	1991	6	2006	7
1977	7	1992	17	2007	10
1978	2	1993	9	2008	9
1979	9	1994	5	2009	4
1980	10	1995	6	2010	9
1981	3	1996	6	2011	4
1982	4	1997	6	2012	5
1983	9	1998	3		
1984	3	1999	8		

Solution:

a. The histogram for the number of earthquakes over the period 1970 - 2012 is shown in Figure 1.6.

b. The histogram is unimodal and right-skewed.

c. The modal class is 4 - 6.

d. The number of years with 10 or less earthquakes is

$$11 + 14 + 6 + 8 = 39.$$

So the percentage of years with no more than 10 earthquakes is $\frac{39}{43} \times 100 = 90.7$. This means that in about 91% of years the number of earthquakes is 10 or less.

■

■ **Example 1.12** The following chart presents yearly highway toll for 2010–2014. The first one is not a proper bar chart for describing road toll. In fact, it was constructed using improper scaling. For example, 12 men icons are used for 280, which implies 280/12 = 23.333 per icon whereas 7

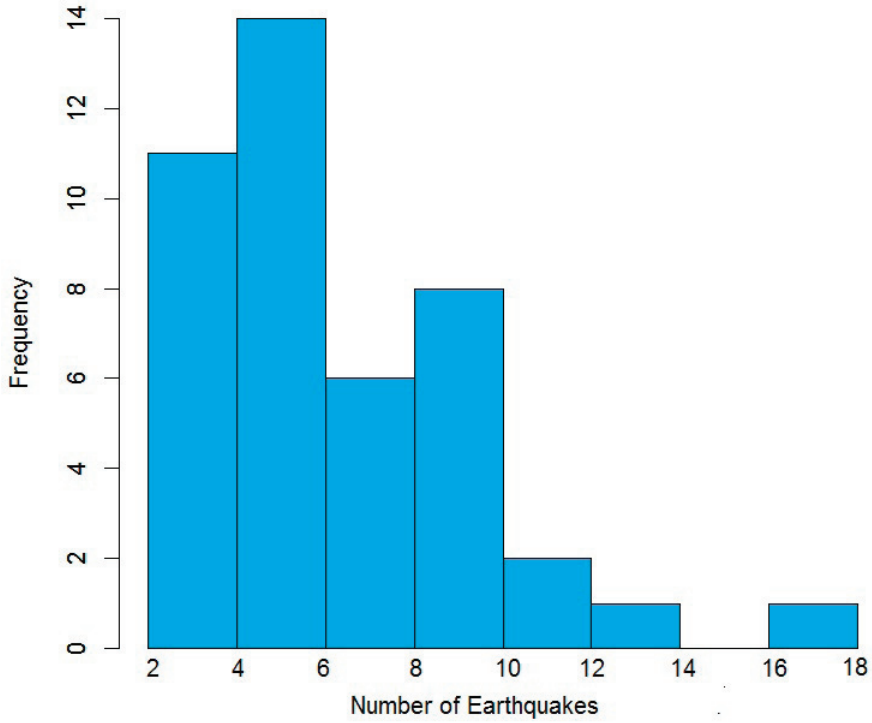


Figure 1.6: Histogram of earthquakes

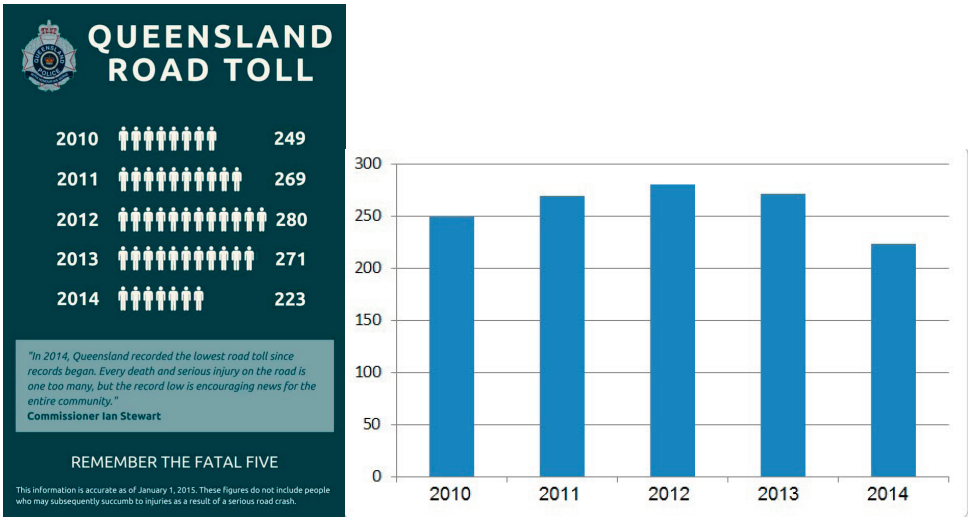


Figure 1.7: Scaling in a bar chart; Left – Improper; Right – Proper Scaling

icons are used for 223, which gives 223/7 = 31.857 per icon. The 2nd chart is the correct bar chart. The road toll was quite high in 2012, and is record low in 2014.

Another graph with improper scaling is the enrollment chart for Obama care given in Figure 1.8. ■

Figure 1.8: Graphs with improper (left) and proper (right) scaling

■ **Example 1.13** The following histograms show the distribution of the population of male and female according to age. The age structure of a population affects a nation's key socioeconomic issues. Countries with young populations (high percentage under age 15) need to invest more in schools, while countries with older populations (high percentage ages 65 and over) need to invest more in the health sector. On the basis of the histograms in Figure 1.9, answer the following.

a. Describe and compare the histograms of male and female populations.

b. Do the histograms indicate that women live longer than men? Explain.

c. Find the percentage of women who are 70+.

Solution:

a. Both histograms are bimodal and right-skewed. The first mode is in the age group 5 – 19 and the second one is in the age group 35 – 44. Thus, histograms for male and female are very similar.

b. The percentages of females in the age groups 40+ are higher than the corresponding percentages for males. This indicates that, in general, women live longer than men.

c. The percentage of women who are 70+ is

$$1.75 + 1.50 + 1 + 1 = 5.25.$$

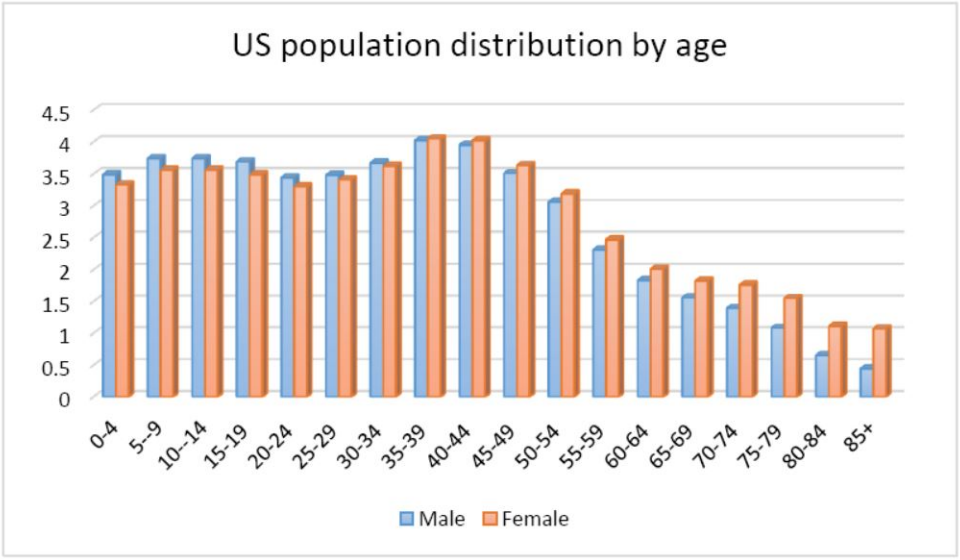


Figure 1.9: Histograms for population distribution by age

■

Chapter 1 Summary

1.1 Terminologies
- Population, samples and variables

1.2 Types of Data
- Quantitative $\rightarrow \begin{cases} \text{discrete} \\ \text{continuous} \end{cases}$
- Qualitative $\rightarrow \begin{cases} \text{nominal} \\ \text{ordinal} \end{cases}$

1.3 Summarizing and Displaying Categorical Data
- Bar chart and pie chart

1.4 Summarizing and Displaying Quantitative Data
- Bar chart and histogram
- Shape of histograms: unimodal or multimodal; left-skewed, right-skewed or symmetric

1.5 Exercises

1. The following data[2] are murder rates (per 100,000 people) in Louisiana and Iowa over years 1996 – 2013. By 2013 rank, the highest murder rates occurred in Louisiana and the lowest in Iowa.

Table 1.3: Murder Rates in Louisiana and Iowa

year	1996	1997	1998	1999	2000	2001	2002	2003	2004
Louisiana	17.5	15.7	12.8	10.7	12.5	11.2	13.2	13	12.7
Iowa	1.9	1.8	1.9	1.5	1.6	1.7	1.5	1.6	1.6
year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Louisiana	9.9	12.4	14.2	11.9	11.8	11.0	11.1	10.6	10.8
Iowa	1.3	1.8	1.2	2.5	1.3	1.2	1.4	1.6	1.4

2. The data in the following table represent the elevation of 50 highest mounts in the USA[3].
3. The data in the following table represent the lengths of the longest rivers in the USA[4].
4. In the following[5]

[2] http://www.deathpenaltyinfo.org/murder-rates-nationally- and-state#MRalpha
[3] Wikipedia
[4] Wikipedia
[5] Source: Wikipedia

Table 1.4: The 50 highest summits of the United States

Mountain Peak	Elevation	Mountain Peak	Elevation	Mountain Peak	Elevation
McKinley	20,236	Williamson	14,379	Maroon Peak	14,163
Saint Elias	18,009	La Plata Peak	14,368	Wrangell	14,163
Foraker	17,400	Blanca Peak	14,357	Sneffels	14,159
Bona	16,550	Uncompahgre Peak	14,321	Capitol Peak	14,137
Blackburn	16,390	Crestone Peak	14,300	Pikes Peak	14,115
Sanford	16,237	Lincoln	14,293	Eolus	14,090
Fairweather	15,299	Castle Peak	14,279	Augusta	14,072
Hubbard	15,016	Grays Peak	14,278	Handies Peak	14,058
Bear	14,831	Antero	14,276	Culebra Peak	14,053
Hunter	14,573	Evans	14,271	San Luis Peak	14,022
Alverstone	14,564	Longs Peak	14,259	of the Holy Cross	14,009
Whitney	14,505	Wilson	14,252	Grizzly Peak	13,995
University Peak	14,470	White ain Peak	14,252	Humphreys	13,992
Elbert	14,440	North Palisade	14,248	Keith	13,982
Massive	14,428	Princeton	14,204	Ouray	13,961
Harvard	14,421	Yale	14,200	Vermilion Peak	13,900
Rainier	14,417	Shasta	14,179		

Table 1.5: The lengths (in miles) of longest rivers in United States

Name	Length	Name	Length	Name	Length
Missouri River	2,341	Brazos River	860	Gila River	600
Mississippi River	2,202	Green River	760	Sheyenne River	591
Yukon River	1,979	Pecos River	730	Tanana River	584
Rio Grande Gulf of Mexico	1,759	White River	720	Smoky Hill River	576
Colorado River	1,450	James River	710	Niobrara River	568
Arkansas River	1,443	Kuskokwim River	702	Little Missouri River	560
Columbia River	1,243	Cimarron River	698	Sabine River	553
Red River	1,125	Cumberland River	696	Red River of the North	550
Snake River	1,040	Yellowstone River	678	Des Moines River	525
Ohio River	979	North Platte River	665	White River	506
Colorado River of Texas	970	Milk River	625	Trinity River	506
Tennessee River	935	Ouachita River	605	Wabash River	503
Canadian River	906	Saint Lawrence River	600		

Table 1.6: Number of hurricanes per year from 1850-2009

Period	Number of hurricanes	Period	Number of hurricanes
1850-59	3,3,5,4,3,4,4,3,6,7	1930-39	2,3,6,11,7,5,7,4,4,3
1860-69	5,6,3,5,3,3,5,6,3,6	1940-49	6,4,4,5,8,5,3,5,6,7
1870-79	10,6,4,3,4,5,4,3,10,6	1950-59	11,8,6,6,8,9,4,3,7,7
1880-89	9,4,5,3,4,6,10,11,6,6	1960-69	4,8,3,7,6,4,7,6,4,12
1890-99	2,7,5,10,5,2,6,3,5,5	1970-79	5,6,3,4,4,6,6,5,5,5
1900-09	3,5,3,7,3,1,6,0,6,6	1980-89	11,12,6,4,13,11,6,7,12,11
1910-19	3,3,4,4,0,5,10,2,4,2	1990-99	14,8,7,8,7,19,13,8,14,12
1920-29	4,5,3,4,5,2,8,4,4,3	2000-09	15,15,12,16,15,28,15,10,16,9

2. Summary Statistics

Outline

1. Measures of Central Tendency
2. Measures of Relative Standing
3. Measures of Spread or Variability
4. Applications of the Standard Deviation and Z-scores

Some functions of data that provide useful information on the entire data set are referred to as the **summary statistics**. Commonly used summary statistics are the mean, median, percentiles, quartiles, range, and standard deviation. These basic statistics are useful to extract key information from a data set. There are three types of summary statistics as shown below.

Measure of Central Tendency	Measure of Relative Standing	Measure of Variability
Mean, median, and mode	Percentiles, quartiles, and z-score	Range, Inter Quartile Range, and standard deviation

2.1 Measures of Central Tendency

Mean, median, and mode are referred to as measure of central tendency because these are points around which majority of the data tend to cluster.

Definition 2.1 — Mean is the usual average of all the numbers in a data set.

For example, the GPA is commonly used to judge the academic performance of a student. Suppose that there are n numbers, say,

$$x_1, x_2, \ldots, x_n$$

The mean is defined by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\text{total}}{\text{the number of data}}.$$

Read $\bar{x}$ as "x-bar." Notice that

$$\text{the number of data} \times \text{mean} = n\bar{x} = \sum_{i=1}^{n} x_i = \text{total}.$$

Definition 2.2 — Median is the middle most numbers when *the data are arranged in increasing order.* If the data set contains odd number of data points (such as 7) then the median is the middle most number when the data are arranged in increasing order. If there are even number of data points, then the median is the average of the two middle most numbers in the ordered data set.

TO FIND THE MEDIAN:

1. Arrange the data in increasing order.
2. If the number of data points is odd, then there is only one middle most datum in the above arrangement, which is the median.
3. If the number of data points is even, then there are two middle most data points in the arrangement in step 1, and the average of these two data is the median.

If the number of data is n, and is odd, then the median is the data point at the $\frac{n+1}{2}$ position in the ordered data set.

If n is even, then the median is the average of the numbers in $\frac{n}{2}$ and $\frac{n}{2}+1$ positions in the ordered data.

Mean vs. Median: If the data set is skewed or containing outliers (a data point which is either too small or too large) then the median is a better measure to describe the center of the data. If the histogram is symmetric or approximately symmetric then the mean or median can be used to summarize the data. Another important feature of the median is that it is not much affected by "extreme values," whereas the mean is affected by extreme values. Another important difference is, for any data set

$$n \times \bar{x} = \text{number of data} \times \text{mean} = \text{total},$$

where as $n\times$ median is not necessarily equal to the total.

■ **Example 2.1** Consider the following data:

3 12 11 5 9 10 1 7 4

To find the median, we arrange the data in ascending order:

There are 9 data, and so the median is the number in the $\frac{9+1}{2} = 5$th position, which is 7. The

data:	1	3	4	5	7	9	10	11	12
position :	1	2	3	4	5	6	7	8	9

mean is

$$\frac{3+12+11+5+9+10+1+7+4}{9} = \frac{62}{9} = 6.89$$

Suppose we change the data by adding two extreme numbers as shown below.

3 12 11 5 9 10 1 7 4 55 60

By sorting the above data, we obtain

data:	1	3	4	5	7	9	10	11	12	55	60
position :	1	2	3	4	5	6	7	8	9	10	11

There are now 11 data, so the median is the number in the $\frac{11+1}{2} = 6$th position, which is 9. However, the new mean is

$$\frac{3+12+11+5+9+10+1+7+4+55+60}{11} = \frac{177}{11} = 16.09.$$

The mean has increased from 6.89 to 16.09 because of the two new data, whereas the median has increased from 7 to 9. Thus, the median is little affected by the extreme values whereas the mean is much affected. ■

When do we use median instead of mean? The median is an appropriate statistic to describe the "center" of the data that are skewed. Furthermore, if a data set includes some extreme values,

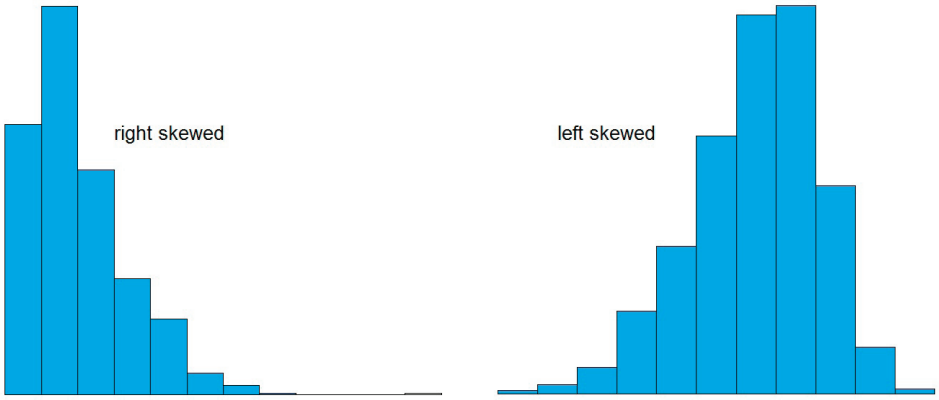


Figure 2.1: Skewed histograms

then median is preferred to the mean as a measure of central tendency. Median is commonly used to describe typical price of a house a neighborhood, and to describe the typical household income in a neighborhood or town.

Some Examples for Skewed Distributions

1. Scores of an easy test (Ans: left skewed)
2. Scores of a hard test (Ans: right skewed)
3. Scores of a fair test (Ans: symmetric or bell shaped)
4. Household incomes in a town (Ans: right skewed)
5. Salaries of employees in a university (Ans: right skewed)
6. Salaries of army personnel (Ans: right skewed)

Comparison between mean and median based on histograms

- If the histogram of a data set is skewed to left, then the mean of the data is expected to be less than the median.

- If the histogram of a data set is skewed to right, then the mean of the data is expected to be larger than the median.
- If the histogram is symmetric, then the mean and median are approximately equal. See Figure 2.2.

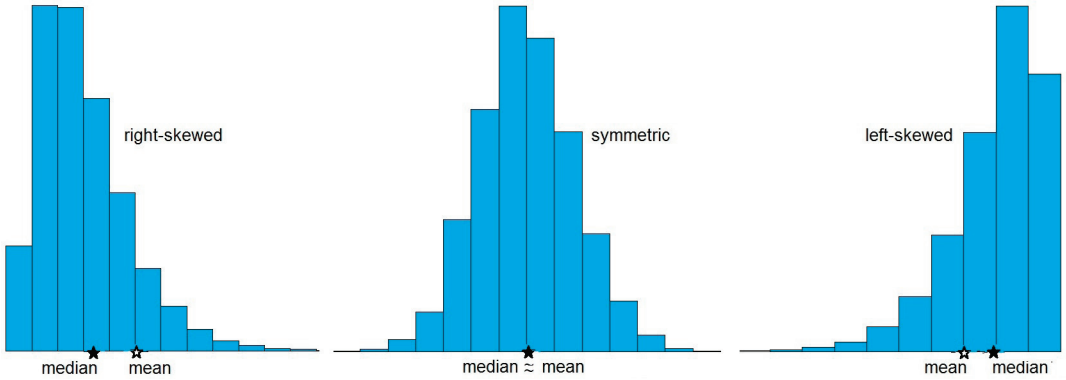


Figure 2.2: Shapes of histograms: median $<$ mean for right-skewed; mean $\approx$ median for symmetric; median $>$ mean for left-skewed

Definition 2.3 — Mode of a data set is the datum that occurs most frequently in the data.

2.2 Measures of Relative Standing

Percentiles and quartiles are measures of relative standing in the data set. For example, a student scored 80 out of 100 in a math test. This score alone does not provide much information on his/her performance in the class. It may be the case that the test was easy and majority of the students scored 80 or higher. On the other hand, if the sore is known to be in the 90th percentile then we can judge that the student's performance is very satisfactory, because "90th percentile" implies that 90 percent of the students in the class scored 80 or less. Suppose the student scored 88 in a chemistry test, and this is in the 60th percentile in the class. This means that only 60% of students scored 88 or below, and so the student's performance in the chemistry may be judged as "average."

Definition 2.4 — Percentile is a measure of comparative standing, and is the value below which a given percentage of data fall. For example, the 10th percentile is the value (or score) below which 10 percent of the data fall.

Definition 2.5 — Quartiles The 25th percentile is called the first quartile, and is denoted by Q_1. The 75th percentile is called the third quartile, and is denoted by Q_3. In this terminology, median is the 50th percentile or second quartile.

To find the quartile,

1. arrange the data in ascending order.
2. Median is the datum in the middle most position.
3. The first quartile is the median of the data below the median.
4. The third quartile Q_3 is the median of the data above the median.

Definition 2.6 — The five-number statistics: The set of statistics

$$\{\text{min}, Q_1, \text{median}, Q_3, \text{max}\}$$

is called the 5-number statistics.

■ **Example 2.2** The "travel to work data" in Example 1.8 are arranged in increasing order as shown below. For these data, find the 5-number statistics and interpret the meanings of the quartiles.

miles	1	1	1	2	2	2	2	2	3	3	4	4	4	4	4
position no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
miles	5	5	5	5	6	6	6	6	7	7	8	8	9	9	9
position no.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
miles	9	10	10	11	11	12	13	14	14	15	16	17	18	18	18
position no.	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45

Solution:

1. We note first that the min $= 1$ and the max $= 18$.
2. To find the median, we note that the number of data 45 is an odd number, and so the median is the number at the (45+1)/2 = 23rd position, which is 6. So the median is 6.
3. The first quartile Q_1 is the median of the data below the median of all data. There are 22 data below the median 6. So the median is the average of the numbers in the

$$\frac{22}{2} = 11\text{th} \quad \text{and} \quad \frac{22}{2}+1 = 12\text{th}$$

positions. That is, $Q_1 = \frac{4+4}{2} = 4$.

4. The third quartile Q_3 is the median of the data above the median. The data above the median are in positions

$$24, 25, 26, ..., 43, 44, 45.$$

The middle position is $(24+45)/2 = 34.5$. So the the 3rd quartile is the average of the numbers in 34th and 35th positions, which is

$$(11+11)/2 = 11.$$

Alternatively, we can find the third quartile as follows. Note that the first quartile is the average of the data in the 11th and 12th positions, so the third quartile should be the average of the data at the 11th and 12th positions from the top, which are

$$45-11+1 = 35 \quad \text{and} \quad 45-12+1 = 34.$$

So the 5-number statistics are

min	Q_1	med	Q_3	max
1	4	6	11	18

■

TI Calc

Calculation of 5-number statistics using the TI-84:

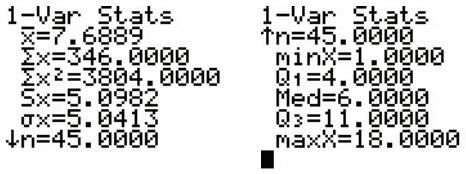
Entering data

1. Select [STAT]
2. Select [EDIT] and then [1]; this is default, and so press [ENTER]
3. Enter the data one by one by pressing [ENTER] after each data value.
4. After entering the last datum, press [2nd] and [QUIT]

Calculating statistics

1. Select [STAT] $\rightarrow$ [CALC] $\rightarrow$ [1 - Var Stats]
2. Select [2nd] [L_1], and press [ENTER]
3. Move the cursor down, to see all the statistics.

For the travel data in Example 2.2, the statistics are

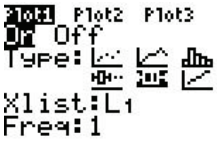


Box-plot: A graphical display of the 5-number statistics is referred to as the box-plot. Box plots also have lines extending horizontally (whiskers) from the boxes indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot or box-and-whisker diagram.

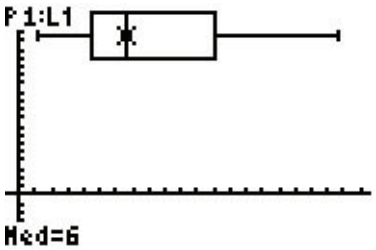
TI Calc

To draw box-plot using TI-84:

1. Select [STAT PLOT] by pressing [2nd] and [STAT PLOT]
2. Select Plot1 [ON] by moving the cursor over [ON] and then pressing [Enter]
3. Select box-plot icon by moving the cursor over the icon, and pressing [Enter]
4. Enter the List where the data are stored. For example, if the data are in $L1$, press [2nd] and L1.
5. Select [Zoom], and then [ZoomStat]
6. Press [TRACE] to find the median. Use navigation to find min, max and quartiles.



■ **Example 2.3** For travel to work data in Example 1.8, the box-plot is



Notice that the box on the right of the center is larger than the one on the left, which indicates that the data are right skewed. Larger box on the left of the center indicates that the data are left skewed. ■

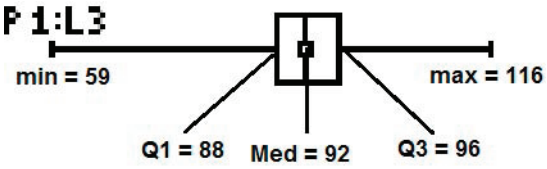
■ **Example 2.4 (Page 63, Problem 26).** The data represent durations of 60 dormancy periods (time between eruptions) of the Old Faithful geyser in Yellowstone National Park. The times are in minutes.

91	88	82	90	89	94	99	88	91	93	94	92
99	92	89	88	90	94	83	116	89	92	95	86
99	59	94	86	93	88	85	101	94	93	105	99
90	90	61	95	96	90	96	71	96	83	92	99
88	103	66	90	101	84	93	97	105	99	91	92

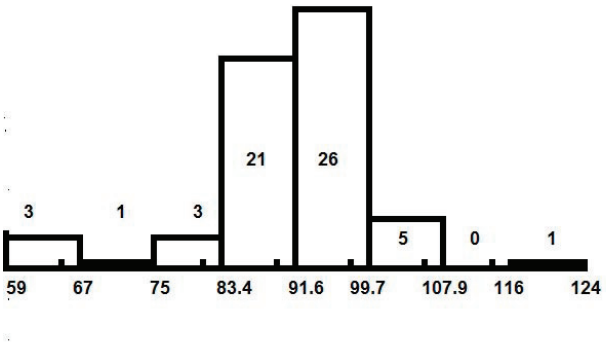
a. Draw the box-plot, and 5-number statistics.
b. Describe the shape of the plot, and interpret the 5-number statistics.
c. Draw the histogram.
d. Describe some features of the histogram.
e. What measure of central tendency would you use to summarize the center?

Solution:

a. The box-plot using TI calculator:



b. The box-plot is approximately symmetric indicating that the mean is approximately same as the median. On the basis of 5-umber statistics, we see that the dormancy periods last between 59 and 116 minutes, with median 92 minutes. Furthermore, 25% of the dormancy periods are 96 minutes or more, and 25% of them are 88 minutes or less.
c. The histogram is unimodal with modal calss 91.6–99.7, and approximately symmetric.
d. The histogram using TI calculator:



e. Since the both histogram and the box-plot are symmetric, either mean or median can be used to summarize the center. In fact for these data, the mean is 91.13 which is very close to the median 92.

■

2.3 Measures of Spread or Variability

Definition 2.7 — Range The difference (Max – Min) is referred to as the range. This is a measure of the spread of the data, and is easy to calculate.

Range may provide some information about the spread, but this is not a good measure of spread. For example, consider the test scores (out of 100) of 50 students which are arranged in ascending order:

```
12, 14, 80, 80, 82, 82, ..., 97, 97, 98, 99
```

For this data, the range is 99 - 12 = 87 which indicates high variability in the scores. However, the data are not much spread out as the most of the scores are between 80 and 99.

Definition 2.8 — Inter Quartile Range (IQR) The difference $Q_3 - Q_1$ is referred to as the IQR. This is also called mid-range, and is a measure of the spread of the middle 50% of the data.

IQR is also useful to identify outliers in a data set. The outliers may be defined as follows.

Definition 2.9 — Outliers A datum is considered outlier if it falls below the lower boundary or fall above the upper boundary, where

$$\text{Lower Boundary} = Q_1 - 1.5 \times IQR \quad \text{and} \quad \text{Upper Boundary} = Q_3 + 1.5 \times IQR.$$

■ **Example 2.5** For the geyser data in Example 2.4, let us check if there are any outliers. For these data

$$\min = 59, \quad Q_1 = 88, \quad \text{median} = 92, \quad Q_3 = 96, \quad \max = 116.$$

$$\text{Lower Boundary} = Q_1 - 1.5IQR = 88 - 1.5(96 - 88) = 76$$

and

$$\text{Upper Boundary} = Q_3 + 1.5IQR = 96 + 1.5(96 - 88) = 108.$$

There are three data 61, 71 and 66 are less than 76, and one datum 116 is greater than 108. So, 61, 71, 66 and 116 are outliers.

■

Definition 2.10 — Standard Deviation Standard Deviation is a commonly used measure of spread, and is defined by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The quantity

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is called **variance**. As an example, let us find the variance of the data

```
3, 5, 6, 2, -4, 6
```

The mean $\bar{x} = \frac{3+5+6+2+(-4)+6}{6} = \frac{18}{6} = 3$. The variance could be obtained using the following table: So the

$$\text{variance} = s^2 = \frac{72}{6-1} = 14.4,$$

and the standard deviation

$$s = \sqrt{14.4} = 3.79.$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
3	0	0
5	2	4
6	3	9
2	-1	1
-4	-7	49
6	3	9
18	0	72

TI Calc

Calculating Mean and Std Deviation Using TI-84:

1. Enter the data in a list, say, L2
2. Press [Stat], select [Calc] and [1-Var Stats]
3. Press 2nd and then L2

For the data in the above table, you see $\bar{X} = 3$ (mean) and $S_X = 3.7947$ (std deviation).

■ **Example 2.6** Consider the following two sets of data.

```
data x:  3.06 2.58 1.47 1.47 2.42 1.92 0.88 3.35 0.41 1.26
         2.59 3.13 1.82 4.35 1.45 2.51 2.92 4.03 3.82 1.58

data y:  10.58 18.21 -4.35 -3.60 -0.46 16.52 10.10 9.85 9.25 -4.44
         -1.30 -1.96 -5.99 -2.56  9.53  1.62 -7.80 2.00 0.11  1.72
```

It appears that data x is less spread out than the data y. For these data sets, mean of x = 2.35 and the std deviation of x is 1.078. The mean of y = 2.85 and std deviation of y is 7.606. ■

■ **Example 2.7** For the travel to work data in EXAMPLE 1.8, compute the range, standard deviation and the IQR.

Range = $18 - 1 = 17$; $s = 5.098$, $IQR = Q_3 - Q_1 = 11 - 4 = 7$ ■

■ **Example 2.8** The following are the average gas mileage of 60 compact cars produced in year 2012.

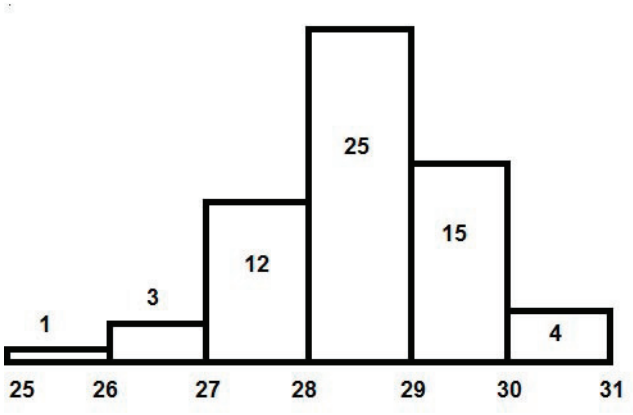
26	26	28	28	30	28	29	29	29	29	27	28
28	25	29	30	27	29	29	29	28	28	28	28
27	27	27	28	28	27	29	28	28	28	28	26
28	28	30	27	27	28	28	27	28	27	27	28
30	28	29	29	28	29	28	28	27	29	29	29

a. Draw histogram of the data using Xmin = 25, Xmax = 30, Xscl = 1, Ymin = 0, Ymax = 30, Yscl = 1, Xres = 1.
b. Describe some features of the histogram.
c. Based on the shape of the histogram, compare the mean and median.
d. Draw the box-plot with 5-number statistics, and interpret their meanings.
e. Find the mean, range, IQR and s.
f. Between the mean and median, which one would you like to use to summarize center? Explain.

Solution:

a. The histogram of the data using the given scales can be constructed as follows:

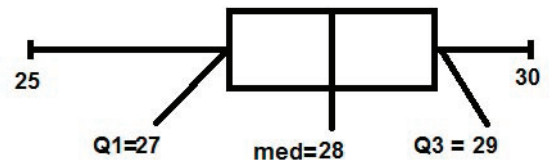
1. Select [STAT PLOT] by pressing [2nd] and [STAT PLOT]
2. Select Plot1 [ON] by moving the cursor over [ON] and then pressing [Enter]
3. Select histogram icon by moving the cursor over the icon, and pressing [Enter]
4. Enter the List where the data are stored. For example, if the data are in $L1$, press [2nd] and L1.
5. Select [WINDOW]
6. Set Xmin = 25; Xmax = 30; Xscl = 1; Ymin = 0; Ymax = 30; Yscl = 1; Xres = 1
7. Press [GRAPH]
8. Press [TRACE] to find the frequency in the first bar, and then use the navigation button to find the frequencies in other bars.



b. The histogram appears to be slightly left skewed with modal class 28 – 29.

c. The histogram is **slightly left skewed**, so we expect the mean to be less than but close to the median.

d. The box-plot with 5-number statistics is



Median gas mileage is 28, which means that 50% of the 2012 compact cars give gas mileage less than 28, and the remaining 50% cars give more than 28. The middle 50% of gas mileage range from 27 to 29. The box-plot implies that the gas mileage data are nearly symmetric.

e. mean = 28.03, range = 30 - 25 = 5, IQR = $Q_3 - Q_1 = 29 - 27 = 2$, and $s = 1.0410$.

f. Since the histogram is slightly left skewed and the box-plot implies that the data are nearly symmetric, both mean and median (they are practically the same) can be used to summarize the center.

■

2.4 Applications of the Standard Deviation and Z-scores

The mean and standard deviation are quite useful to summarize data that are approximately symmetric or bell shaped. Let $\bar{x}$ and s denote the mean and standard deviation of a data set that is approximately symmetric. Then the interval

$$(\bar{x}-s,\ \bar{x}+s) = \bar{x}\mp s$$

includes nearly 68% of the data. That is, about 68% of data fall within one standard deviation of the mean. The general rule is referred to as the **empirical rule**, which is as follows.

Result 2.1 — Empirical Rules for Symmetric Data (or bell shaped data),

1. about 68% of data fall within one std deviation of the mean;
2. about 95% of data fall within 2 std deviation of the mean;
3. about 99.7% of data fall within 3 std deviation of the mean.

■ **Example 2.9** It is known that the IQ scores are symmetrically distributed with mean 100 and $s = 16$.

a. What % of IQ scores fall within 1 std deviations of the mean? Write the interval.
about 68%; $100 \pm 16 = (84, 116)$.

b. What % of IQ scores fall within 2 std deviation of the mean? Write the interval.
about 95%; $100 \pm 2 \times 16 = (68, 132)$.

c. Find the interval that would include 99.7% of the IQ scores.
$\bar{x} \pm 3 \times s = 100 \pm 3 \times 16 = 100 \pm 48 = (52, 148)$.

■

2.5 z-Scores

Let $\bar{x}$ and s denote the mean and standard deviation of a data set. Then the z-score of a data point x is given by

$$z = \frac{x-\bar{x}}{s}.$$

The z-score is also called the **standardized score**. The original data value x is called "raw score." For a given z-score, the corresponding raw score x can be calculated using the relation

$$x = \bar{x} + z \times s.$$

Note that the mean of the z-scores are always zero and the standard deviation is 1.

■ **Example 2.10** Consider the data

$$3, 4, 6, 12, 22, 13, 42$$

The mean is $\bar{x} = 14.57$ and $s = 13.76$. The z-scores of the data are calculated in the following table.

x	3	4	6	12	22	13	42
$z = \frac{x-\bar{x}}{s}$	-0.8408	-0.7682	-0.6228	-0.1868	0.5400	-0.1141	1.993

It can be readily checked that the mean of the z-scores $\bar{z} = 0$ and the standard deviation of the z-scores $s = 1$. ■

In fact, for any data, the mean of z-scores is 0 and the standard deviation of z-scores is 1, and so we have the following empirical rules for z-scores.

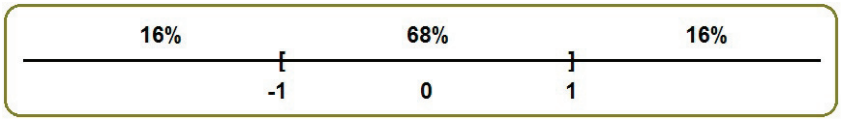
Result 2.2 — Empirical Rules for z-scores If z-scores are approximately symmetric then
1. about 68% of z-scores are in $[-1,1]$
2. about 95% of z-scores are in $[-2,2]$
3. about 99.7% of z-scores are in $[-3,3]$

The original data values are called raw scores or the actual measurements. The z-score of a datum is more informative than the datum itself. Specifically, the z-score of a datum is useful to understand the relative standing of the datum within the data set.

■ **Example 2.11** The mean and standard deviation of the scores in a math test are 72 and 9, respectively. Suppose John scored 87 in the test. What is his z-score? Suppose the instructor has decided to give *A* to the top 20% of students in the class. Will John receive an *A*? Explain.
Solution: John's z-score is $\frac{87-72}{9} = \frac{15}{9} = 1.67$. Also, if the scores are symmetric, then about 68% of the z-scores are in

$$[-1,1]$$

That means, of the remaining $100-68=32\%$, $\frac{32}{2} = 16\%$ of z-scores fall below -1 and 16% of z-scores fall above 1. So John's score is in the top 20th percentile, and he will receive an *A*.



■

■ **Example 2.12** Kevin's ACT score is 28 (max 36), and the corresponding z-score is 1. Assume that the histogram of ACT scores is approximately symmetric (bell-shaped). Can we place Kevin's score in the top 10th percentile? Explain.
Solution: Notice that about 68% of the z-scores are in $[-1,1]$. That means, of the remaining $100-68=32\%$, $\frac{32}{2}$=16% of z-scores fall below -1 and 16% of z-scores fall above 1. Thus, Kevin's score can be placed in the top 16th percentile, not on the top 10th percentile.

■

■ **Example 2.13** The z-score of the annual income of a family in a town is 2.1. Explain the comparative standing of this family income within this town.
Solution: According to the empirical rule, 95% of z-scores of the family incomes in the town is in the interval $[-2,2]$. Of the 5% of z-scores outside this interval, only 2.5% of z-scores fall above 2. So this family makes annual income which is in the top 2.5th percentile. ■

2.6 Exercises

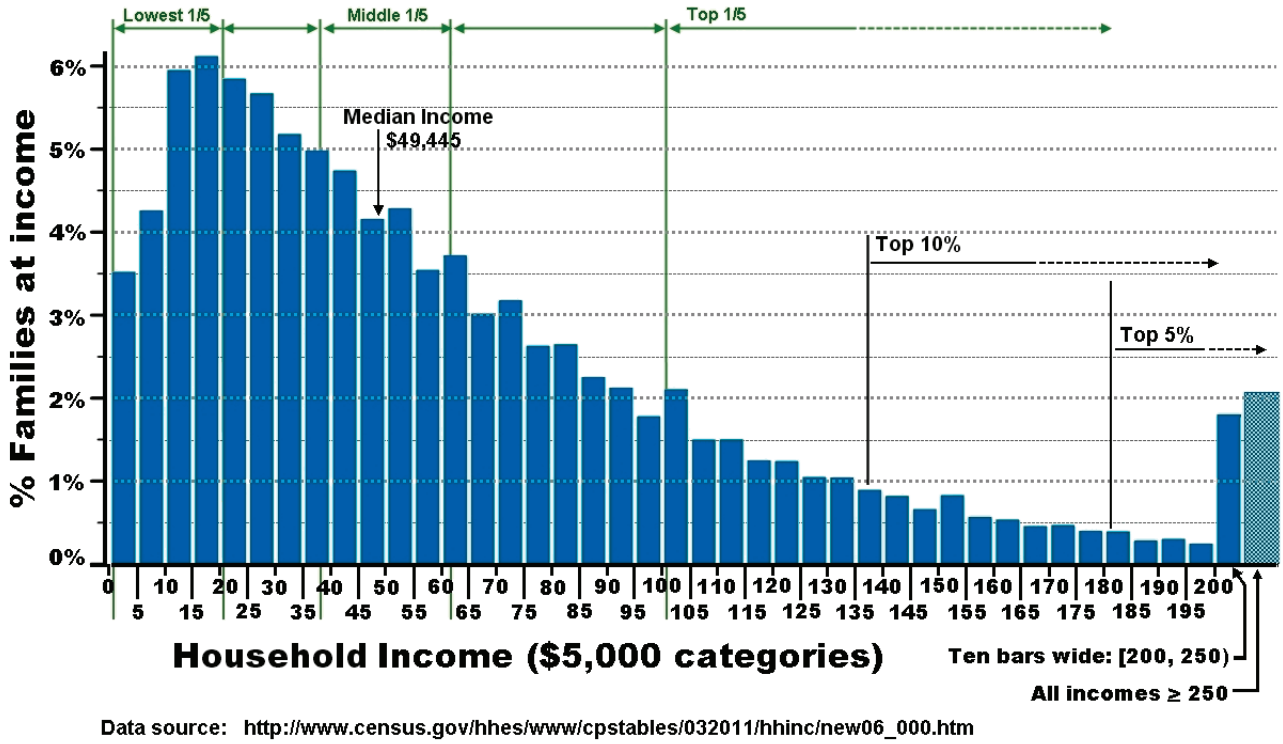
1. The following data in Table 2.1 are the average gas prices by state, as of August, 2015.

   a. Find the 5-number statistics.
   b. Construct the box-plot, and comment.
   c. Construct the histogram of the data, and give some features of the histogram.
   d. Find the mean and standard deviation of the average gas prices.
   e. Find the z-score for the average gas price in Illinois and interpret its meanings.
   f. Between the mean and median, which is a good measure of the center? Explain.
   g. Using the empirical rule, what can we say about the percentage of average gas prices fall in the interval $\bar{x} \pm s$, and find the interval.
   h. What is the actual percentage of data fall in the interval in part **g**.

Table 2.1: Average gas prices in states as of August, 2015

No.	State	Aver. Price	No.	State	Aver. Price	No.	State	Aver. Price
1	South Carolina	2.06	18	Massachusetts	2.44	35	New Mexico	2.64
2	Mississippi	2.12	19	Maine	2.45	36	Michigan	2.66
3	Alabama	2.12	20	Maryland	2.45	37	New York	2.67
4	Louisiana	2.20	21	Vermont	2.45	38	Washington DC	2.68
5	Tennessee	2.20	22	Kansas	2.45	39	Montana	2.71
6	Arkansas	2.24	23	Rhode Island	2.46	40	Wisconsin	2.74
7	New Jersey	2.24	24	West Virginia	2.52	41	Utah	2.77
8	Virginia	2.26	25	Ohio	2.53	42	Wyoming	2.78
9	Texas	2.28	26	Minnesota	2.55	43	Colorado	2.80
10	North Carolina	2.29	27	Pennsylvania	2.55	44	Idaho	2.89
11	Missouri	2.34	28	Arizona	2.60	45	Oregon	2.90
12	Florida	2.33	29	South Dakota	2.60	46	Illinois	2.91
13	Delaware	2.33	30	Iowa	2.61	47	Washington	2.99
14	Georgia	2.35	31	North Dakota	2.61	48	Nevada	3.15
15	Oklahoma	2.36	32	Connecticut	2.63	49	Hawaii	3.16
16	New Hampshire	2.39	33	Indiana	2.63	50	Alaska	3.39
17	Kentucky	2.44	34	Nebraska	2.63	51	California	3.41

2. The following is the histogram representing annual household incomes in the USA.



a. What percentage of household incomes that are $\leq \$49,445$?
b. How do you compare the mean and median on the basis of the above histogram.
c. What is the modal class?
d. Find the 90th percentile, and interpret its meanings.

3. Probability

3.1 Probability and Statistics

The word probability is commonly used in everyday life. The meanings of probability in different contexts may be the same, but the probability in different contexts is evaluated by different ways. For example, the probability of obtaining a head in a single flip of a coin is 1/2, which is calculated assuming that the coin is balanced or the outcomes of a head and tail are equally likely to occur. On the other hand, the probability in the statement "80% chance of precipitation" in a weather report is calculated using a different method. The probability forecast reflects the relative frequency of occurrence of a weather condition in the past circumstances similar to the current situation. This relative frequency of occurrence is also used to assess the probability of classifying a driver as low, medium or high risk for insurance purpose. Yet another form of probability is known as the subjective probability which is based on an individual's viewpoint. It does not involve precise computation but is often a reasonable assessment by an expert in the area of interest.

There are two popular approaches to define probability: The **classical approach**, and the **relative frequency approach**. The classical approach assumes that the outcomes are the results of some random process or experiment, and all the outcomes are equally likely to occur. In relative frequency approach, we assign probabilities to events on the basis of their frequency of occurrences. In this chapter, we shall consider only these two approaches of assigning probabilities to events.

3.2 Events and Sample Space

To define the probability using the classical approach, we shall first define some basic terminologies.

Definition 3.1 — Random Experiment An experiment or process whose outcomes are determined only by chance factors is called random experiment. The outcome that occurs cannot be predicted with certainty. The outcomes are also called **simple events** or **elementary events**.

Definition 3.2 — Sample Space The set of all possible outcomes of a random experiment is called sample space.

Definition 3.3 — Event A collection of none, one or more than one outcomes from a sample space is called event.

Occurrence of Event A: If one of the outcomes in event A has occurred, then we say that the event A has occurred.

■ **Example 3.1** Consider flipping a coin two times. This is a random experiment because the outcome of each flip is determined by only chance factors. The possible outcomes can be obtained using the following **tree diagram**.

At the end of the tree diagram, we see four outcomes. Thus, the sample space includes four

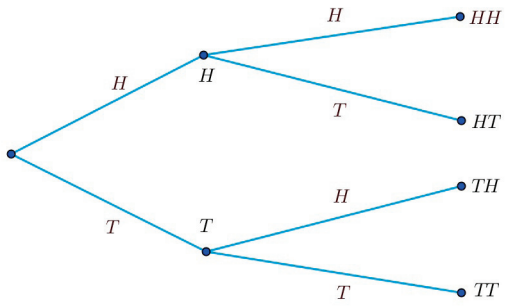


Figure 3.1: Tree diagram illustrating the sample space

possible outcomes, namely,

$$\mathbf{S} = \{HH, HT, TH, TT\}.$$

For example, here HT means the outcome of the first flip is head and the outcome of the second is tail. Let A denote the event of observing exactly one head out of these two flips. Then the event A includes two outcomes, namely,

$$\{HT, TH\}.$$

Let B denote the event of observing at least a tail. Then the event B includes three outcomes

$$\{HT, TH, TT\}.$$

The event B occurs if one of the simple events in B occurs. In other words, while flipping a coin twice, if we observe a head and tail, tail and head or head both times, then we say that the event B has occurred. ■

3.3 Calculation of Probabilities

Definition 3.4 — Classical Approach Assuming that all outcomes of a random experiment are equally likely, the classical probability of an event A is defined as

$$P(A) = \frac{\text{number of outcomes in the event } A}{\text{number of outcomes in the sample space}}.$$

The notation $P(A)$ means the probability that A occurs or probability of observing the event A.

We shall now consider two examples where we can assume the outcomes of a random experiment are equally likely to occur, and assign probabilities using the classical approach.

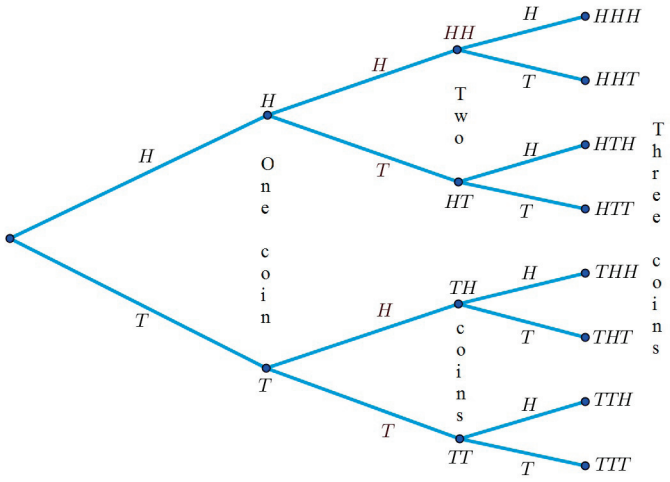


Figure 3.2: Sample spaces for one coin, two coins and for three coins

■ **Example 3.2** Consider a roll of a fair die. Since it is a fair die, it is reasonable to assume that the outcomes 1, 2, 3, 4, 5, and 6 are all equally likely to occur. The sample space is given by

$$\mathbf{S} = \{1,2,3,4,5,6\},$$

and so the probability of observing a 1 in a single roll is 1/6; we write this as $P(1) = \frac{1}{6}$. Similarly, we assign probabilities to other outcomes as $P(2) = \frac{1}{6}, \ldots, P(6) = \frac{1}{6}$.

Suppose that A is the event of observing an even number. The event A includes three outcomes, namely, 2, 4 and 6. We write $A = \{2,4,6\}$, and

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in the smple space}} = \frac{3}{6} = \frac{1}{2}.$$

Notice that $P(\mathbf{S}) = \frac{6}{6} = 1$. This means that if a die is rolled, then one of the six outcomes will certainly occur. ■

Axioms of Probability

1. For any event A, $0 \leq P(A) \leq 1$.
2. Probability of an impossible event is zero.
3. $P(\mathbf{S}) = 1$.

■ **Example 3.3** Suppose three coins are to be flipped simultaneously. What are the probabilities of observing

a. exactly 2 heads?
b. at least 2 tails?
c. at most one head?
d. all tails?

■

Solution: The sample space can be obtained by extending the tree diagram in Figure3.1 for one more coin as shown in Figure 3.2.

coin 1	coins 1 & 2	coins 1, 2 & 3
H	HH	HHH
T	TH	THH
	HT	HTH
	TT	TTH
		HHT
		THT
		HTT
		TTT

The set of eight simple events, namely,

$$\mathbf{S} = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

is the sample space.

a. There are three outcomes THH, HTH, HHT that have exactly two heads. So

$$P(\text{exactly two heads}) = \frac{3}{8}.$$

b. There are four outcomes, namely,

$$TTH,\ THT,\ HTT,\ TTT$$

that have two or more tails and so

$$P(\text{at least two tails}) = \frac{4}{8} = \frac{1}{2}.$$

c. This event is the same as the event of observing two or more tails, and so the

$$P(\text{one or no head}) = P(\text{at least two heads}) = \frac{4}{8}.$$

d. Only one outcome with all tails, and so the probability is $\frac{1}{8}$.

■ **Example 3.4** Consider rolling a pair of dice. The sample space consists of 36 outcomes as shown in Table 3.1. Find the probabilities of the following events.

a. observing a seven;

b. observing a 7 or more;

c. observing a double;

d. observing an even number.

Solution: Since there are two dice, and for each die there are six possible outcomes, the sample space includes $6^2 = 36$ outcomes as shown in Table 3.1. For example, the entry $(2,3)$ in the table means that Die 1 shows up 2 and Die 2 shows up 3.

a. Of 36 outcomes, there are seven outcomes (antidiagonal entries) add up to 7, and so the probability is $\frac{6}{36} = \frac{1}{6}$.

b. All six entries in the antidiagonal, and 15 entries below the antidiagonal are adding up to 7 or more. So the probability of observing a 7 or more is $\frac{21}{36} = \frac{7}{12}$.

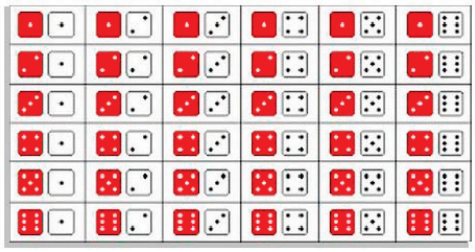
c. The six entries in the diagonal are doubles, and so the probability is $\frac{6}{36} = \frac{1}{6}$.

d. Half of the 36 entries add up to even numbers, and so the probability is $\frac{18}{36} = \frac{1}{2}$.

■

Table 3.1: Sample space for a pair of dice

Die 1	Die 2 1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)



Definition 3.5 — Statistical Odds Sometimes probability is expressed in terms of odds so that a commoner can understand. The odds in favor of an event is defined as

$$(\text{the number of outcomes in favor of } A : \text{number of outcomes against } A)$$

The odds against A is expressed by reversing the order of the pair.

■ **Example 3.5** Consider rolling a fair die. The odds for rolling a 6 is 1:5 (read 1 to 5) and the odds against rolling a six are 5:1. Suppose A denote the event of observing a 1 or 6, and B denotes the event of observing an even number. The odds for A are 2:6 (or 1:3), and odds against A are 6:2 (or 3:1). The odds in favor of B are 3:3 (or 1:1), and odds against B are also 3:3. ■

Definition 3.6 — Relative Frequency Approach The probability of an event A is estimated by the proportion of times the event A occurs in n repetition of a random experiment. If k denotes the number of times the event occurred in n repetition of the experiment, then

$$\text{proportion} = \frac{k}{n} \quad \text{approaches} \quad P(A) \quad \text{as } n \text{ getting large.}$$

Suppose all outcomes of a random experiment are really equally likely, then the **Law of Large Numbers**[1] asserts that the classical approach and the frequency approach with large number of repetitions will produce the same results. The following example illustrates this relation between the classical and frequency approaches.

■ **Example 3.6** In Example 3.2, we assumed that the die was fair, and all six outcomes are equally likely to occur. The probabilities are all equal to 1/6. Suppose that the die is not known to be fair or we are not sure on it, then we could roll the die for a large number of times and on the basis of number of occurrences of each number we can assign the probability. Suppose that the die was rolled 1,000 times and the outcomes were recorded as shown in the following table.

outcomes	1	2	3	4	5	6	Total
number of occurrences	167	191	141	177	147	177	1000
probability	.167	.191	.141	.177	.147	.177	1

We see in the above table that the number 1 occurred 167 times, number 2 occurred 191 times, and

[1] According to the law, the results obtained from a large number of trials should be close to the true value, and will tend to become closer as more trials are performed.

so on. The probability of observing a 1 is estimated by

$$P(1) = \frac{167}{1000} = .167.$$

We can estimate other probabilities $P(2), ..., P(6)$ similarly as shown in the above table. ■

Remark

In the above example, we assigned probabilities on the basis of frequencies of occurrences of each outcome out of 1,000 rolls. We do not know if the die is fair or not, but the probabilities are not far from 1/6 = .1667, which is the probability that we assign if we use the classical approach. Indeed, $P(1)$ is practically 1/6 whereas $P(2) = .191$, which is little more than 1/6. We can estimate these probabilities more accurately by increasing the number of rolls to, say, 100,000. If the die is actually a fair one, then all the probabilities determined by the frequency approach should be very close to 1/6.

There are situations where the events are determined by nature, and we have no control on them. For example, tornadoes, tropical storms, eruption of volcanoes, hurricanes, etc., happening periodically in a random manner. The probabilities for such events can be determined on the basis of frequencies of occurrences in the past, as shown in the following example.

■ **Example 3.7** Consider the hurricane data in Table 1.6, which lists the number of hurricanes per year for 160 years starting from 1850 to 2009. Recall that a total of 1,013 hurricanes occurred during these 160 years. We shall assign probability of observing a specified number of hurricanes in a year using the relative frequency approach. Note that the first column lists the number of

Table 3.2: Frequency table for the hurricane data in Table 1.6

													Total
hurricanes, x_i	0	1	2	3	4	5	6	7	8	9	10	≥ 11	
Number of years, f_i	2	1	6	25	26	22	25	13	8	3	6	23	160
$P(x_i) = f_i / \sum f_j$	.013	.006	.038	.156	.163	.138	.156	.081	.050	.019	.038	.144	1

years with no hurricane, the second lists the number of years with one hurricane, and so on. Using the relative frequency approach, we estimate

$$P(\text{observing } x_i \text{ number of huuricanes in a year}) = P(x_i) = \frac{f_i}{\sum f_j}.$$

For example,

$$P(\text{four hurricanes in a year}) = \frac{26}{160} = 0.1625.$$

Suppose it is desired to estimate the probability of having 10 or more hurricanes in a year. An estimate of this probability is

$$\frac{6}{160} + \frac{23}{160} = 0.1813.$$

■

Exercise 3.1-3.3

3.3.1 A sample space include three elementary events A, B and C. If $P(A) = .3$, $P(B) = .4$, what is $P(C)$?

3.3.2 Consider an experiment with the sample space

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}.$$

Suppose $P(E_1) = .1$, $P(E_2) = .05$, $P(E_3) = P(E_4) = .15$, $P(E_5) = .2$ and $P(E_6) = P(E_7)$. Let $A = \{E_1, E_3, E_7\}$ and $B = \{E_2, E_4, E_5, E_6\}$

a. What is $P(E_6)$?
b. Find $P(A)$.
c. Find $P(B)$.
d. Between A and B, which event is more likely to occur?

3.3.3 Three identical chips, each with number 1 on one side and number 2 on the other side, are tossed simultaneously. Assume that all possible outcomes are equally likely to occur.

a. Write the sample space by listing all elementary events.
b. What is the probability of observing the sum of the numbers is equal to 4?
c. Find the probability of observing the sum of the numbers is 5 or more?

3.3.4 A four-sided die and a six-sided die are rolled simultaneously. Assume that all elementary events are equally likely to occur.

a. Write the sample space by listing all elementary events.
b. Find the probability of observing a double.
c. What is the probability of observing sum of the numbers greater than or equal to 5?

3.3.5 There are three identical chips in an urn. Two of the chips are red and the other one is blue. Two chips are drawn from the urn.

a. Find the sample space using a tree diagram.
b. Find the probability that both chips are red.
c. Find the probability that one is red and another is blue.

3.4 Some Counting Rules to Calculate Probabilities

Calculation of probabilities by listing all the elementary events in a sample space is not often feasible. For instance, what is the probability of observing exactly five heads in 10 flips of a coin? In this case, the sample space includes (as will be seen later) 1,024 outcomes, and it is not wise to write all 1,024 outcomes. In fact, to find the probability, what we need is the total number outcomes in the sample space, and the total number of outcomes with exactly five heads. These numbers can be obtained without listing all possible outcomes by using some counting formulas.

Definition 3.7 — Mutually Exclusive Two events are mutually exclusive if occurrence of one prevents the occurrence of the other. Mutually exclusive events cannot occur together at the same time.

Definition 3.8 — Exhaustive A set of events is collectively exhaustive if they exhaust all possibilities. The collectively exhaustive events cover the entire sample space.

Some examples for mutually exclusive and collectively exhaustive events

1. Consider flipping a coin once. The outcomes are mutually exclusive because head and tail cannot occur together. These outcomes are also exhaustive because these are the only possible outcomes, and $\{H, T\}$ is the sample space.
2. Consider tossing two coins simultaneously. Let A be the event that one of the coin shows up head, and B denote the event that one of the coins shows up tail. These two events are not mutually exclusive, because these two events can occur together as HT or TH.

3. The outcomes of a roll of a die, namely,

$$\{1,2,3,4,5,6\}$$

are mutually exclusive and exhaustive. The event that the die roll to 1 (say), and the event that the die roll to 4 (say) are mutually exclusive, but not exhaustive, because there are other possible outcomes. The event that the die roll to

$$1,2 \text{ or } 5$$

and the event that the die roll to

$$2,3,4 \text{ or } 6$$

are collectively exhaustive (because their union is the sample space), but not mutually exclusive because both occur together if the die rolls to 2.

Result 3.1 — Multiplication Rule Suppose an experiment is performed at two stages. The experiment results into one of n_1 mutually exclusive events at the stage 1, and for each event at the first stage, the experiment at the second stage results into one of n_2 mutually exclusive events. Then the sample space includes $n_1 n_2$ simple events.

■ **Example 3.8** Suppose we roll a 4-sided die first and then we roll a 6-sided die. In the first stage we have 4 simple events and in the second stage there are 6 simple events, and so the number of simple events in the sample space is $4 \times 6 = 24$. ■

Multiplication Rule - General case: The multiplication rule can be generalized as follows. Suppose an experiment is performed at k stages. At the first stage, the experiment will result into one of n_1 mutually exclusive events, at the second stage n_2 mutually exclusive events, and so on. Then the sample space of the k-stage experiment includes

$$n_1 \times n_2 \times n_3 \times \cdots \times n_k$$

simple events.

■ **Example 3.9** In a men clothing store, a particular brand shirt is available in different types, sizes, and colors as follows: There are two types, namely, half-sleeve and full-sleeve; four sizes, namely, small, medium, large and X-large; four different colors, namely, light yellow, white, gray and light blue. Then the total number of available options for this brand of shirt is

$$2 \times 4 \times 4 = 32.$$

■

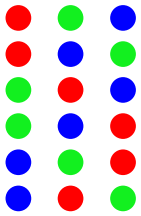
Definition 3.9 — Permutation is an ordered arrangement of the objects from a list of objects. The number of permutations on a set of n objects is

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1.$$

We read $n!$ as "n factorial." Thus, permutations are arrangements of objects where order matters.

As an example, consider the following arrangement of three color stickers:

All possible rearrangements of these three stickers are shown in the right. To find the number of possible different arrangements, we can apply the multiplication rule as follows. There are three possible stickers for the first position. Once one of the stickers is placed in the first position, there are two possible stickers for the second position, and only one for the third position. So using the multiplication rule, we have



$$3 \times 2 \times 1 = 3! = 6$$

different arrangements of three stickers as shown on the right.

Result 3.2 — ${}_nP_k$ The number of ways of obtaining an ordered subset of k objects from a set of n different objects is given by

$$_nP_k = \frac{n!}{(n-k)!}.$$

■ **Example 3.10** Consider the number lock on the right. This lock has 4 wheels, each is numbered 0 through 9. Suppose the sequence of numbers that unlocks is a set of four different numbers. If the sequence of unlocking numbers is unknown, what is the maximum number of tries required to unlock? ■

Solution: The unlocking numbers are four different numbers chosen from 0 through 9. For example, the set of numbers 4,3,1,3 can not unlock because wheels 2 and 4 show the same number. So the maximum number of tries should be equal to the number of arrangements of four different numbers selected from the set $\{0,1,2,...,9\}$, and is

$$_{10}P_4 = \frac{10!}{(10-4)!} = \frac{10!}{6!} = 5040.$$

TI Calc

Calculating ${}_nP_r$:

- Enter the value of n, then press [math], select [PRB] and [nPr], and then enter the value of r, and press [ENTER].
- For example, 10 [math], [PRB], [nPr], 4 [ENTER] = 5040
  That is, ${}_{10}P_4 = 5040$.

Result 3.3 — Total Count from n Repetitions of an Experiment Suppose an experiment may result into one of k mutually exclusive and exhaustive outcomes, and the experiment is repeated n times, then

the total number of outcomes in the sample space $= k^n$.

Note that there are n repetitions, and at each time there are k outcomes. So applying the multiplication rule, we obtain

$$\overbrace{k \times k \times \cdots \times k}^{n \text{ times}} = k^n$$

For example, if a coin is flipped 5 times (or 5 coins are tossed simultaneously), then there are 2 outcomes at each flip, and 5 repetitions, so the total number of possible simple events is $2^5 = 32$.

■ **Example 3.11** Suppose a coin is to be flipped 10 times.

a. What is the number of simple events in the sample space?

b. What are the chances of observing head in all 10 flips?

c. What are the chances of observing at least one tail?

Solution:

a. Two outcomes for each coin, and there are 10 coins. So $2^{10} = 1,024$ outcomes.

b. This event includes only one outcome, namely, $(HHHHHHHHHH)$. So the

$$\text{probability} = \frac{1}{1024}.$$

c. This event includes all outcomes except the one $(HHHHHHHHHH)$. So the probability is

$$P(\text{at least one head}) = \frac{1023}{1024}.$$

■

Combination Permutations are for arrangements of objects where order matters, and a **combination** is a set of objects where order does not matter. For example, the following arrangements of three stickers

are two different permutations, but same combination. There are situations where one is interested in just a combination of objects ignoring the order in which the objects were drawn. For example, to assess the quality of a shipment of items, one may inspect a sample of selected items. Here, the order in which the items were selected is irrelevant for assessing the quality. In the following, we shall see the number of different combinations of objects one can select from a set of objects.

Result 3.4 — Total Number of Combinations The number of ways one can select combination of k different objects from a set of n different objects is given by

$$_nC_k = \frac{n!}{k!(n-k)!}.$$

■ **Example 3.12** How many combinations of two different numbers one can select from the set $\{1,2,3,4,5\}$?

Solution: We can enumerate all possible combinations of two different numbers as shown on the right. We see that there are 10 combinations of two different numbers. Instead of enumerating all possible two different numbers, we can just evaluate the possible number of two different numbers as

1,2	1,3	1,4	1,5
	2,3	2,4	2,5
		3,4	3,5
			4,5

$$_5C_2 = \frac{5!}{2!(5-2)!} = \frac{120}{2\times 6} = 10.$$

Note that $_5C_3$ is also 10, because for every combination of two different numbers selected, there is one combination of three different numbers is left. ■

TI Calc

Calculating $_nC_k$: Enter the value of n, then press [math], select [PRB] and [nCr], and then enter the value of k, and press [ENTER]. For example,

10 [math], [PRB], [nCr], 2 [ENTER] = 45

That is, ${}_{10}C_2 = 45$.

■ **Example 3.13** What is the total number of possible samples of 10 students from a class of 40 students?

Solution: It is not feasible to list all samples of size 10 from a class of 40 students, and such list is also not necessary. The total number of possible samples of size 10 can be directly calculated as

$$ {}_{40}C_{10} = 847,660,528. $$

■

■ **Example 3.14 — Pick 3** is a Louisiana lotto game that is described as follows. Pick any 3 different numbers 0 through 9 in a play slip by paying $1. If all three numbers picked match the winning numbers in exact order, then the player wins "straight" and if the drawn numbers match the winning numbers in any order then the player wins "box." If a player buys one play slip,

a. what is the probability that he will win straight?

b. What is the probability that he will win box?

Solution:

a. The number of different arrangements of 3 numbers from 0, 1, 2, ..., 9 is

$$ {}_{10}P_3 = \frac{10!}{7!} = \frac{10 \times 9 \times \times ... \times 1}{7 \times 6 \times \times ... \times 1} = 10 \times 9 \times 8 = 720. $$

So if only one play slip is bought, then the probability of winning straight is $\frac{1}{720}$.

b. To win box, the combination of drawn numbers should match the winning numbers. The number of combinations of 3 numbers from 10 numbers is

$$ {}_{10}C_3 = \frac{10!}{3!(10-3)!} = 120. $$

So the probability of winning box is $\frac{1}{120}$.

■

■ **Example 3.15 — Lotto** Consider a lotto where a player selects a set six different numbers from

$$ 1,2,3,4....,42,43,44. $$

If the all the six numbers match the winning numbers that will be drawn randomly, the player win the jackpot. If the player buys only one ticket, what are the chances of winning the jackpot?

Solution: The number of possible combinations of six different numbers that can be selected from 44 different numbers is

$$ {}_{44}C_6 = 7,059,052. $$

If the player buys only one ticket, then the chances of winning the jackpot is

$$ \frac{1}{7059052} = .00000014. $$

If a buyer buys all possible combinations by spending $7,059,052 (assuming that each ticket costs one dollar), then she/he will certainly win the jackpot. ■

■ **Example 3.16 — Powerball** is a lotto game where a person can choose 5 white ball numbers out of

$$1, 2, 3, ..., 68, 69$$

white ball numbers, and choose one red ball (powerball) numbers from

$$1, 2, ..., 26$$

red ball numbers. Every Wednesday and Saturday night at 10:59 p.m. (Eastern Time), five white balls are drawn out of a drum with 69 white balls and one red ball out of a drum with 26 red balls. The jackpot is won by matching all five white balls in any order and the red Powerball. If you buy only one ticket, what are the chances of winning the jackpot? What are the chances of matching four white ball numbers?

Figure 3.3: One of the past winning numbers

Solution: The number of possible combinations of five different numbers that can be drawn from 69 numbers is

$$_{69}C_5 = 11,238,513.$$

This, the probability of matching white ball numbers is

$$\frac{1}{11,238,513}$$

The probability of matching the red ball number is $\frac{1}{26}$. So the probability of matching all five white ball numbers and the red ball number is

$$\frac{1}{11,238,513} \times \frac{1}{26} = \frac{1}{292,201,338} = 0.000000003422$$

Thus, the probability of winning the jackpot is 1 in 292,201,338.

To find the probability of matching four white ball numbers, we calculate

$$_{69}C_4 = 864,501$$

and the probability is 1 in 864,501. ■

■ **Example 3.17** Suppose that any child in a family is male or female, independently of others, with probability 0.5. In a family of five children, what are the chances of observing all boys? What are the chances of observing at least one girl?

Solution: Each child is either boy or girl, and there are five children, so the total number of possible outcomes in the sample space is $2^5 = 32$. To find the probabilities, we do not need to write the sample space. Among these 32 elementary events, only one elementary event includes all boys, namely, $(BBBBB)$. Therefore, the probability of observing all boys is

$$\frac{1}{32} \text{ or the chances are } \frac{1}{32} \times 100 = 3.125\%.$$

That is, of 100,000 families with 5 children, 3,125 families are expected to have all 5 boys.

Now to find the probability of observing at least one girl, we note that all outcomes, except the one $(BBBBB)$ include at least one girl. Therefore, the probability of observing at least one girl in a family of five children is $\frac{31}{32}$ or 96.875%. ■

Exercise 3.4

3.4.1 In a group 7 students, each student is classified based on her/his eye color, namely, Black (B), Brown (Br), Cyan (C), Hazel (H) and other (O). For example, an elementary event could be (Br,C,H,B,C,B,O). What is the total number of possible elementary events?

3.4.2 What is the probability of drawing a number card less than or equal to 4 from a standard deck of 52 cards?

3.4.3 There are 10 red marbles, 5 green marbles and 15 yellow marbles in a bag. One marble is selected randomly. Find the probabilities of following.

- **a.** The selected marble is yellow.
- **b.** The selected marble is not yellow.
- **c.** The selected marble is neither red nor yellow.

3.4.4 A royal flush in a poker game is the sequence of five cards A, K, Q, J, Q, 10 in the same suit. For example,

$$A\clubsuit\ K\clubsuit\ Q\clubsuit\ J\clubsuit\ 10\clubsuit$$

is the royal flush club. A set of 5 cards was drawn from a standard deck.

- **a.** What is the probability that it is royal flush club?
- **b.** What is the probability that it is a royal flush?

3.4.5 How many different ways can one answer all the questions of a true-false test consist of 10 questions?

3.4.6 How many ways can 5 marbles with different colors (say, red, yellow, green, orange, blue) be arranged in a row?

3.4.7 How many 3-digit numbers can be formed with the 9 digits $1,2,3,4,5,6,7,8,9$ if

- **a.** repetitions are not allowed, and
- **b.** the last digit must be 9 and repetitions are not allowed?

3.4.8 The number lock on the right has 5 wheels, each has numbers 0 through 9. Suppose the owner of the lock forgot the combination, and remembers that the unlock numbers is a sequence of 5 different numbers.

- **a.** What is the maximum number of tries required to unlock?
- **b.** What is the probability of unlocking by a random sequence of 5 numbers?

3.4.9 A team of 4 students is selected from a class of students consisting of 15 boys and 25 girls.

- **a.** How many ways one can select a team from the class?
- **b.** Suppose the team should include two boys and two girls. How many ways such team can be selected?

3.5 Combination of Events and Probability Rules

Definition 3.10 — Compound Event If an event is a combination of two or more events, then it is referred to as the compound event.

Definition 3.11 — Union and Intersection Suppose event C occurs if an event A occurs or event B occurs or both occur on a single performance of an experiment, then C is a compound event (union of A and B) and is expressed as

$$C = (A \cup B).$$

That is, C includes all simple events either in A or in B or in both. Furthermore,

$$P(A \cup B) = P(\text{A occurs or B occurs or both occur together}).$$

Suppose a compound event C occurs if both events A and B occur together in a single performance of an experiment, then the event C is expressed as

$$C = (A \cap B), \text{ or equivalently } (A \text{ and } B).$$

Result 3.5 For any two events A and B,

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

For mutually exclusive events C and D,

$$P(C \cup D) = P(C) + P(D).$$

Recall that

$$P(A \cup B) = \frac{\text{number of outcomes in } A \cup B}{\text{number of outcomes in the sample space}}.$$

From venn diagrams in Figure 3.4, we see that the number of outcomes in $A \cup B$ is

no. of outcomes exclusively in A + no. of outcomes exclusively in B + no. of outcomes in $A \cap B$

which is the same as

$$\text{no. of outcomes in } A + \text{no. of outcomes in } B - \text{no. of outcomes in } A \cap B.$$

The number of outcomes in $A \cap B$ is subtracted because A includes the outcomes in $A \cap B$ and B also includes the outcomes in $A \cap B$.

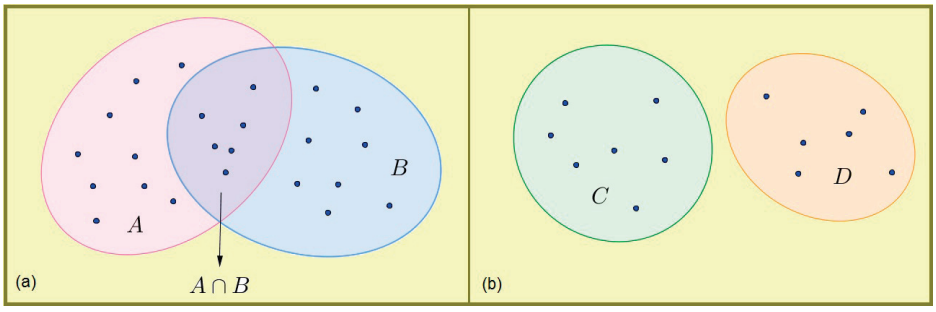


Figure 3.4: Venn diagrams of (a) joint (not mutually exclusive) events and (b) mutually exclusive events

■ **Example 3.18** US Census Bureau reported the following data (numbers are in thousands) on population characteristics and health insurance coverage status for the year 2013. A total of 313.396 millions people were classified according to their health insurance coverage status and their citizenship status as shown in the following table:[2]

[2]Source: U.S. Census Bureau, Current Population Survey, 2014 Annual Social and Economic Supplement.

Characteristic	Not covered at any time in 2013	Covered by some type of insurance	Total
Native	30,587	241,800	272,387
Naturalized citizen	3,043	16,107	19,150
Non-citizen	8,324	13,535	21,859
Total	41,954	271442	313,396

Assume that similar classification holds for the current year. Compute the following probabilities for a randomly selected person from the USA.

a. The person is native not covered by any health insurance.

b. The person is not covered by any insurance.

c. The person is covered by some type of insurance.

d. The person is either a naturalized citizen or covered by some type of insurance.

Solution:

a.

$$\begin{aligned} P(\text{the person is native and not covered}) &= \frac{\text{no. of natives not covered by insurance}}{\text{total no. persons in the US}} \\ &= \frac{30587}{313396} \\ &= .098 \end{aligned}$$

b.

$$\begin{aligned} P(\text{the person is not covered }) &= \frac{\text{no. of persons not covered by insurance}}{\text{total no. persons in the US}} \\ &= \frac{41954}{313396} \\ &= .134 \end{aligned}$$

c.

$$\begin{aligned} P(\text{the person is covered }) &= \frac{\text{no. of persons covered by some insurance}}{\text{total no. persons in the US}} \\ &= \frac{271442}{313396} \\ &= .866 \end{aligned}$$

d. Let A denote the event that the person is a naturalized citizen, and let B denote the event that the person is covered by some type of insurance. Notice that these two events are not mutually exclusive. In other words, a person can be a naturalized citizen with some type of

health insurance. Therefore,

$$\begin{aligned}
& P(\text{naturalized citizen or with some insurance}) \\
= \; & P(A \cup B) \\
= \; & P(A) + P(B) - P(A \cap B) \\
= \; & \frac{\text{no. naturalized citizens}}{\text{total no. persons in the US}} + \frac{\text{no. persons with some insurance}}{\text{total no. persons in the US}} \\
& - \frac{\text{no. naturalized citizens with some insurance}}{\text{total no. persons in the US}} \\
= \; & \frac{19150}{313396} + \frac{271442}{313396} - \frac{16107}{313396} \\
= \; & .061 + .866 - .051 \\
= \; & .876.
\end{aligned}$$

■

Definition 3.12 Let A be an event. The event that A does not occur is called the complement of A, and is denoted by A^c. Notice that A and A^c are mutually exclusive events, and

$$P(A) + P(A^c) = 1 \quad \text{or} \quad P(A^c) = 1 - P(A).$$

For some cases, it is easier to find $P(A)$ from $P(A^c)$. The following example is one such cases.

■ **Example 3.19** A pair of dice is rolled. Let A denote the event of observing two distinct numbers. Then A^c is the event of observing a double. Since there are only six doubles in the sample space, $P(A^c) = \frac{6}{36}$ and

$$P(A) = 1 - P(A^c) = 1 - \frac{6}{36} = 1 - \frac{1}{6} = \frac{5}{6}.$$

■

Definition 3.13 — De Morgan's Laws Let A and B be any two events. Then

$$(A \cap B)^c = A^c \cup B^c$$

which is "the negation of a conjunction is the disjunction of the negations." Further,

$$(A \cup B)^c = A^c \cap B^c$$

which is "the negation of a disjunction is the conjunction of the negations."

The relations stated in De Morgan's Laws are easy to understand directly. For example, the statements

"I don't like coke or pepsi $[C^c \cup P^c]$"

and

"I do not like coke and I do not like pepsi $[C^c \cap P^c]$"

are the same.

To understand the relation by a mathematical example, let

A denote the event that a variable $x \geq 2$,

and let

B denote the event that $x \leq 4$.

This means that

$$A \cap B = 2 \leq x \quad \text{and} \quad x \leq 4 = \text{“}2 \leq x \leq 4,\text{”}$$

that is, x is in the interval $[2,4]$. So $(A \cap B)^c$ means that x is not in the interval $[2,4]$ or equivalently,

$$(A \cap B)^c = \text{“}x < 2\text{”} \quad \text{or} \quad \text{“}x > 4\text{”}$$

Also, $A^c = \text{“}x < 2\text{”}$ and $B^c = \text{“}x > 4\text{”}$. So $A^c \cup B^c$ is “$x < 2$” or “$x > 4$”.

■ **Example 3.20** Nationality and eye color for a group of 4,848 people from European countries are given in the following table.

	Nationality					
Eye Color	English	German	Irish	Italian	French	Total
Blue	709	539	185	22	145	1600
Brown	661	462	145	145	229	1642
Green	335	222	86	19	55	717
Hazel	346	256	66	39	94	801
Other	45	13	13	6	11	88
Total	2096	1492	495	231	534	4848

let us assume that this sample people represents the all people in these five nationalities. Suppose we select a person randomly from these five nationalities. Find the following probabilities.

a. The person is an English;

b. eye color of the person is hazel;

c. the person is a German or with brown eye;

d. neither English nor with green eye.

Solution:

a. The sample space consists of 4848 persons, of which 2096 are English. So the probability is

$$P(\text{the person is an English}) = \frac{2096}{4848} = 0.432.$$

b. There are 801 persons with hazel eye. So

$$P(\text{hazel eye person}) = \frac{801}{4848} = 0.165.$$

c. Let G denote the event of selecting a German, and let B denote the event of selecting a person with brown eye. Then

$$\begin{aligned} P(G \cup B) &= P(G) + P(B) - P(G \cap B) \\ &= \frac{1492}{4848} + \frac{1642}{4848} - \frac{462}{4848} \\ &= .308 + .339 - .095 \\ &= .552 \end{aligned}$$

d. Let E denote the event of selecting an English, and let N denote the event of selecting a

person with green eye.

$$\begin{aligned} P(E^c \cap N^c) &= P((E \cup N)^c) \qquad \text{[De Morgan's Laws]} \\ &= 1 - P(E \cup N) \\ &= 1 - [P(E) + P(N) - P(A \cap N)] \\ &= 1 - \left[\frac{2096}{4848} + \frac{717}{4848} - \frac{335}{4848}\right] \\ &= 1 - [.432 + .148 - .069] \\ &= 1 - .511 \\ &= .489 \end{aligned}$$

■

Exercise 3.5

3.5.1 Consider rolling a six-sided die. Let A denote the event that includes the elementary evens $\{1,3,5\}$ and B denote the event that includes the elementary events $\{2,3,4,6\}$.

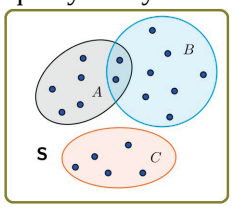
a. Write the elementary events in $A \cap B$, and find $P(A \cap B)$.
b. Write the elementary events in $A \cup B$, and find $P(A \cup B)$.
c. Find A^c, B^c and $A^c \cup B^c$.
d. Verify that $(A \cap B)^c = A^c \cup B^c$.

3.5.2 Consider two events A and B with $P(A) = .4$, $P(B) = .2$ and $P(A \cap B) = .15$.

a. Find $P(A \cup B)$.
b. Find $P(A^c \cap B^c)$.
c. Find $P(A^c \cup B^c)$.

3.5.3 Consider the sample space with three events shown in the venn diagram below. The blue dots are the elementary events, and assume that all elementary events are equally likely to occur.

a. Are the events A, B and C mutually exclusive? Explain.
b. Are the events A, B and C exhaustive? Explain.
c. Find $P(A \cap B)$.
d. Find $P(A \cap B \cap C)$.
e. Find $P(A \cup B)$.



3.5.4 Consider the sample space

$$\mathbf{S} = \{E_1, E_2, E_3, E_4, E_5\}.$$

Let $A = \{E_1, E_3\}$ and $B = \{E_2, E_5\}$.

a. Are these events A and B mutually exclusive? Explain.
b. Are these events A and B exhaustive? Explain.
c. Find the elementary events in $A^c \cap B^c$.
d. Verify that $(A \cup B)^c = A^c \cap B^c$.

3.5.5 A foundry manufactured 500 cast aluminum parts in three shifts of a day. Some of these parts are defective, some are good, and some of them can be repaired so that they meet the standard. The following table presents the results. A part is selected at random.

shifts	defective	non-defective	repairable	Total
morning	10	125	15	150
evening	5	160	10	175
night	15	155	5	175
Total	30	440	30	500

a. Find the probability that the chosen part is defective.

b. Find the probability that the chosen part is defective and was produced during the night-shift.

c. What is the probability that the chosen part is defective or produced in the night-shift?

d. What percentage of parts neither produced in the morning-shift nor defective?

e. Find the probability that a randomly selected part either non-defective or produced in the morning shift.

3.6 Conditional Probability

We shall now see a method of computing probability of an event A, given that another related event B has already occurred. For example, consider flipping two fair coins simultaneously. The sample space is

$$S = \{HH, TH, HT, HH\}.$$

Let A denote the event of observing at least a tail. Then

$$A = \{HT, TH, TT\},$$

and $P(A) = \frac{3}{4}$. Now suppose that after flipping coins, we are told that at least one of the coins turned up head. Given this information, what is the probability of A? If B denotes the event that at least one of the coins is head, then

$$B = \{HT, TH, HH\}.$$

Given that the event B has occurred, our sample space changes to the set of outcomes in B, and the event A now includes only HT and TH. So the probability of A given that B has occurred is $\frac{2}{3}$. This probability is referred to as the conditional probability, denoted by $P(A|B)$. Notice that

$$\begin{aligned} P(A|B) &= \frac{\text{no. of outcomes in } \{HT, TH\}}{\text{no. of outcomes in } \{HT, TH, HH\}} \\ &= \frac{\text{no. of outcomes in } A \cap B}{\text{no. of outcomes in } B} \\ &= \frac{\text{no. of outcomes in } A \cap B / [\text{no. of outcomes in } S]}{\text{no. of outcomes in } B / [\text{no. of outcomes in } S]} \\ &= \frac{P(A \cap B)}{P(B)}. \end{aligned}$$

Definition 3.14 — Conditional Probability The conditional probability of an event A given that B has already occurred is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) > 0.$$

Conditional probabilities are often expressed by by words like "given that," "if," or by words implying a subset of the population. Here are some examples of statements that refer to a conditional probability:

- "90% of students who attended class regularly pass a math test" is translated to $P(\text{pass} \,|\, \text{attended class regularly}) = .9$.
- "the diagnostic test shows positive in 95% of those with the disease" is translated to $P\left(\text{test positive} \,|\, \text{person has disease}\right) = .95$.

- "African-Americans are three times more likely to die of asthma than white Americans" is translated to
  $P\left(\text{an African-American die}\middle|\text{asthma}\right) = 3P\left(\text{a white American die}\middle|\text{asthma}\right)$
- "Asian Americans are 60% more likely to develop diabetes in comparison to European Americans" is translated to
  $P\left(\text{developing diabetes}\middle|\text{Asian Amercian}\right) = 1.6P\left(\text{developing diabetes}\middle|\text{European American}\right)$.

To calculate the conditional probability of an event A given that a related event B has already occurred, we change our sample space that includes only the outcomes contained in B as shown in the following examples.

■ **Example 3.21** For this problem, let us assume that any child in a family is male or female, independently of others, with probability 0.5.

a. In a family of three children, what is the probability of oberving at least one girl?

b. In a family of three children with the youngest one is girl, what is the probability that all three children are girls?

c. In a family of three children with at least one girl, what is the probability that all three children are girls?

Solution: To find the probabilities and the conditional probabilities, we shall first find the sample space which include all possible combinations of boys (B) and girls (G) in a family with three children.

$$\{BBB, BGB, GBB, GGB, BBG, BGG, GBG, GGG\}$$

For example, the outcome BGB means that the first one boy, second girl and the third boy.

a. All simple outcomes except BBB. So the probability is $\frac{7}{8}$

b. Given that the youngest one is girl, the conditional sample space includes the last four outcomes of the sample space, namely,

$$\{BBG, BGG, GBG, GGG\}$$

and only one outcome represents all girls, so the probability is $\frac{1}{4}$.

c. Given the information that there is at lest one girl in the family, the sample space includes all outcomes except the outcome BBB, which is

$$\{BGB, GBB, GGB, BBG, BGG, GBG, GGG\}.$$

Since the only one outcome in the conditional sample space represent three girls, the probability is $\frac{1}{7}$.

■

■ **Example 3.22** A pair of dice is rolled.

a. What is the probability of observing a 7?

b. What is probability of observing a seven given that an odd number has occurred?

c. What is the probability of observing an 8 given that a pair has occurred?

d. What is the probability of observing a 4 given that a number less than or equal to 6 has occurred?

Solution: First, we shall write the sample space:

a. $P(7) = \dfrac{\text{number of outcomes that give 7}}{\text{number of outcomes in the sample sapce}} = \frac{6}{36} = \frac{1}{6}$.

Table 3.3: Sample space for a pair of dice

	Die 1					
Die 2	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

b.

$$\begin{aligned} P(7|\text{odd number}) &= \frac{P(\text{observing 7 and odd number})}{P(\text{observing a odd number})} \\ &= \frac{(\text{no. of outcomes that are odd and add to 7})/(\text{total no. of outcomes})}{(\text{no. of outcomes that are odd})/(\text{total no. of outcomes})} \\ &= \frac{6/36}{18/36} \\ &= \frac{1}{3}. \end{aligned}$$

c.

$$\begin{aligned} P(8|\text{pair}) &= \frac{P(\text{observing 8 and pair})}{P(\text{observing a pair})} \\ &= \frac{(\text{no. of outcomes that are pair and add to 8})/(\text{total number of outcomes})}{(\text{no. of outcomes that are pair})/(\text{total number of outcomes})} \\ &= \frac{1/36}{6/36} \\ &= \frac{1}{6}. \end{aligned}$$

■

In some cases, the conditional probability can be directly evaluated without using the formula. In fact for some problems it is not wise to use the formula to find the conditional probability. We illustrate evaluating the conditional probability in a direct manner using the following two examples.

■ **Example 3.23** In a deck of playing cards, there are four suits and there are 13 cards in each suit. Suppose a card is drawn randomly. What is the probability that the drawn card is 10 heart given that a red card was selected? ■

Solution: Given that a red card was selected, there are 26 possibilities:

A♥, 2♥,...,10♥, J♥, Q♥, K♥, A♦, 2♦,...,10♦, J♦, Q♦, K♦

So the

$$P(\text{the selected card is 10 heart}|\text{red}) = \frac{1}{26}.$$

■ **Example 3.24** Suppose that a box contains 20% red chips, 40% green chips and the remaining blue chips. All chips are identical in shape and weight.
A chip is selected randomly from the box.

a. What is the probability that it is a red or green chip?
b. What is the probability that it is red given that it is not blue?

Solution:

a. As the event of observing a red chip and the event of observing a green chip are mutually exclusive, using the Result 3.5, we obtain

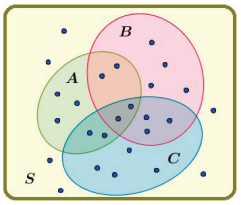
$$P(\text{red chip or green chip}) = P(\text{red chip}) + P(\text{green chip}) = .2 + .4 = .6$$

b. Given that it is not blue, it must be either green or red. The randomly selected chip must be one of those 20% + 40% = 60% chips consisting of red and green chips. So the probability that it is being red is $\frac{.2}{.2+.4} = \frac{1}{3}$.

■

Exercise – Chapter 3

3.6.1 Consider the sample space S that includes 26 equally likely simple outcomes (blue dots) as shown in the picture. The event A includes 9 outcomes, B includes 11 outcomes and C includes 11 outcomes. The sample space includes 26 outcomes.



a. Are the events A, B and C mutually exclusive?
b. Are they exhaustive?
c. Find $P(A|B)$ and $P(B|A)$.
d. Find $P((A \cap C)|B)$ and $P((A \cap B)|C)$.
e. Find $P(A^c \cup B^c)$.

3.6.2 American Red Cross reported the following table[3] on blood types in for African American population. Type O persons can donate red blood cells to anybody, type A can donate to A and AB, type B can donate to B and AB, and *AB* can donate only people with type AB blood. Find the following probabilities for a randomly selected African American.

Blood Type	O	A	B	AB
Percentage	51	26	19	4

a. The person is with blood type A or B.
b. The person is with blood type not AB.
c. The person can donate blood to people with blood type AB.

3.6.3 Suppose a sample space includes only three events A, B and C, with

$$P(A) = .3, \quad P(B) = .4, \quad P(C) = .4 \quad \text{and} \quad P(A \cap B) = .1.$$

The event C can't occur with the event A or with the event B. Find
a. $P(A|B)$
b. $P(B|A)$
c. $P(A|C)$

3.6.4 A coin is flipped 10 times.
a. What is the probability of observing head in all 10 flips?
b. What is the probability of observing at least one head out of these 10 flips?

3.6.5 A die is to to be rolled four times. What is the probability of observing 6 in all four rolls?

3.6.6 Suppose that one of your friends roll a pair of dice.

[3] http://www.redcrossblood.org/learn-about-blood/blood-types

a. What is the probability their sum is 7?

b. After rolling the pair of dice, your friend tells that the outcomes differ by 3. Given this information, what is the probability that sum is 7?

3.6.7 Suppose a high school graduate applying to a college has an 70% chance of being accepted, and that dormitory housing will be provided only for 50% of all of the accepted students. Find the probability that a student being accepted and receiving dormitory housing?

3.6.8 Translate each of the following statements to a probability statement.

a. In China, women are 40% more likely to commit suicide than men.

b. Hispanics are about 1.5 times more likely to develop Alzheimer's disease than whites.

c. The chances that both engines in an aircraft fail are .01%.

d. Sixty percent of applicants get admissions in a university, and 15% of those admitted will be awarded scholarships.

3.6.9 The receiving department of a computer manufacturing company has received 634 computers and monitors. The supervisor of the department classified the equipments and the reasons for returning as follows.

Equipment type	**hardware problem**	**poor performance**	**other reasons**	**Total**
Desktop	82	20	48	150
Laptop	145	43	90	278
Monitors	78	90	40	208
Total	305	153	178	636

a. Among the returned equipments, what percent of laptop with hardware problems?

b. Find the probability that a randomly chosen returned equipment is a monitor or an equipment returned for poor performance?

c. What is the probability that a randomly selected returned equipment is neither laptop nor returned for hardware problems?

d. Find the conditional probability that a randomly selected item has a hardware problem given that it is a monitor.

3.6.10 For this problem, let us assume that any child in a family is male or female, independently of others, with probability 0.5. In a family of three children

a. with the oldest one is boy, what is the probability that all three children are boys?

b. with at least one boy, what is the probability that all three children are boys?

3.6.11 U.S. National Highway Traffic Safety Administration reported the results on number of vehicles involved in fatal crashes by vehicle type and rollover[4] occurrences for the year 2009 as shown in the following table.[5]

	Rollover occurrence		
	Yes	No	Total
Passenger Cars	3,000	15,400	18,400
Pickup trucks	2,400	6,100	8,500
Utility	2,200	4,700	6,900
Van	400	2,100	2,500
Total	8,000	28300	36,300

Assume that the classification of vehicle crash still holds approximately.

[4] A rollover is a type of accident in which a vehicle tips over onto its side or roof

[5] Source: U.S. National Highway Traffic Safety Administration, Traffic Safety Facts, annual. http://www-nrd.nhtsa.dot.gov/CATS/index.aspx

a. What is the probability of a fatal vehicle crash involving a pickup truck with rollover occurrence?

b. Find the probability that a fatal vehicle crash involving a van or a roll over occurrence.

c. Find the probability that a fatal crash involves neither a van nor a rollover occurrence.

3.6.12 There are two urns containing red and green marbles. In urn 1, there are 75% red marbles and 25% green marbles. The urn 2 contains 25% red marbles and 75% green marbles. One urn was selected at random, and a marble was selected from the urn. Given that the selected marble is red, what is the probability that urn 1 was selected?

3.6.13 Suppose a voter poll is taken in three states. In state A, 50% of voters support the liberal candidate, in state B, 60% of the voters support the liberal candidate, and in state C, 35% of the voters support the liberal candidate. Of the total population of the three states, 40% live in state A, 25% live in state B, and 35% live in state C. Given that a voter supports the liberal candidate, what is the probability that she lives in state B?

3.6.14 If you are a man with type 1 diabetes, the probability that your child developing diabetes is 0.06. About .4% of the population in the US have type 1 diabetes. Suppose we select a person from the US randomly, and the person is type 1 diabetes. What is the probability that the person's father is a type I diabetes?

3.6.15 Center for Disease Control (CDC) estimates that about .4% of persons aged 13 years and older are living with HIV infection. The probability that a test will correctly detect HIV given that the person is infected is .95, and the probability that it will show positive for uninfected person is .15. If the test shows positive for a randomly selected person from the age group 13 or above, what is the probability that the person is HIV infected?

Review for the First Test

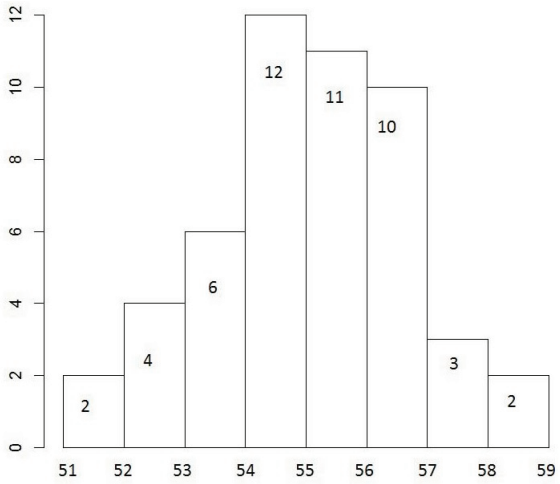
TOPICS	SUMMARY	
Types of variables	**Quantitative** Continuous: height, weight, salary, age, temperature Discrete: that can be counted; number of births in a hospital number of auto accidents in a day	**Qualitative** Nominal: race, religion, gender brand names, colors, etc. Ordinal: variables that can be ordered on a qualitative characteristic; ranks of college teachers, movie ratings, wine ratings, etc.
Summarizing Data Graphically	**bar chart**: Qualitative data; **histogram**: quantitative data; **features of histogram**: unimodal or bimodal; skewed to left or right; symmetric; modal class	
Summary Statistics	**Measures of Location**: Mean, Median and Mode median is preferred to the mean if the data (histogram) is skewed. Variance and Standard Deviation: Empirical rule; z score = $\frac{x-\bar{x}}{s}$ **Measures of Relative Standing**: Percentiles and z-scores 5-number statistics: min, Q1, median, Q3, max; IQR = Q3 – Q1 outliers: Lower boundary: $Q_1 - 1.5 \times IQR$ Upper boundary: $Q_3 + 1.5 \times IQR$	
Probability	event, sample space: probability(event) = $\frac{\text{number of outcomes in the event}}{\text{total number of ouctcomes in the sample sapce}}$	

Model Problems

1. Identify the following data as nominal, ordinal, continuous or discrete.
   - **a.** travel times to work for a sample of 10 workers;
   - **b.** political affiliation of 20 voters;
   - **c.** ranks of army personnel;
   - **d.** the number of rainy days in a month;
   - **e.** race of people;
   - **f.** the number of siblings of a person;
   - **g.** the number of patients admitted in a hospital on a given day;
   - **h.** names of streets in a town;
   - **i.** the colors of cars parked on a street;
   - **j.** amounts of water used by households in a month.
2. The following data represent annual household incomes of 50 households in a neighborhood.

```
56 54 55 53 55 57 56 56 54 52
55 54 53 58 55 52 55 54 54 55
56 55 54 56 55 56 59 57 54 55
54 53 53 59 54 56 56 57 54 56
55 55 56 52 57 56 55 57 52 57
```

   The histogram of the data:



a. Discuss the shape of the histogram.

b. What is the modal class?

c. Calculate the 5-number statistics, and interpret the meanings of the first quartile.

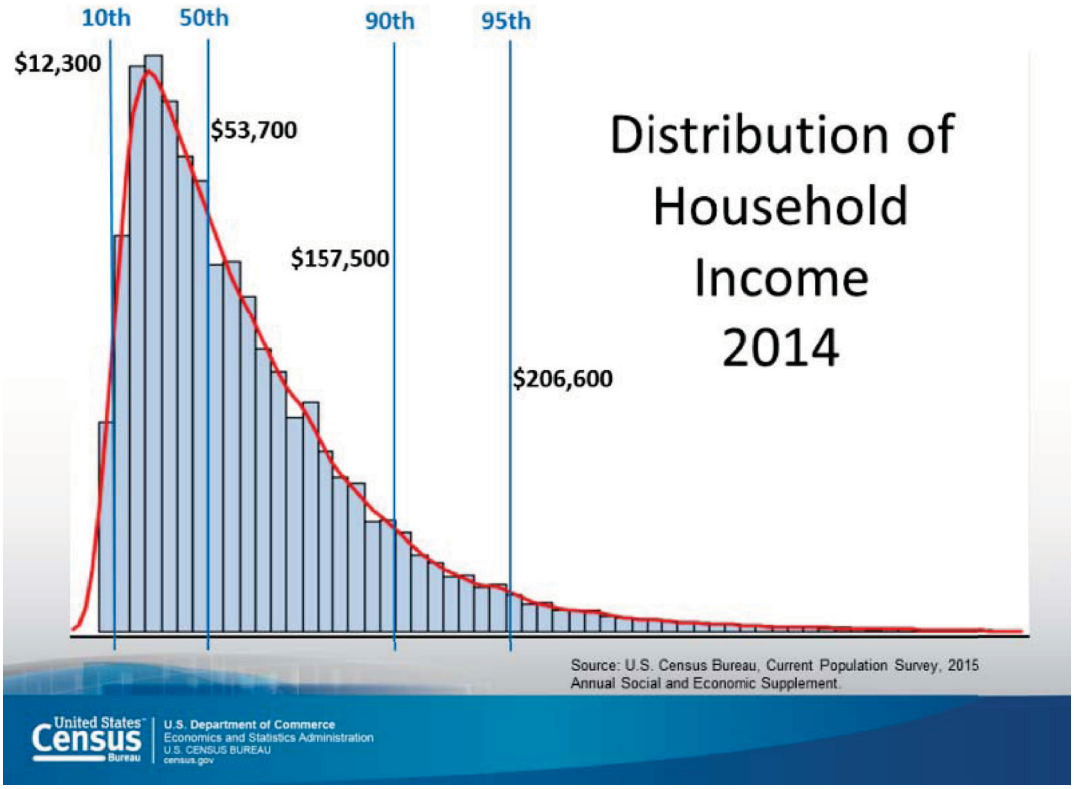
d. Find the IQR and interpret its meanings.

e. Identify the outliers if there are any.

f. Compute the mean and standard deviation.

g. Find the percentage of incomes fall in the interval (52, 56).

3. The following are the histograms of household incomes in the USA for the year 2014.



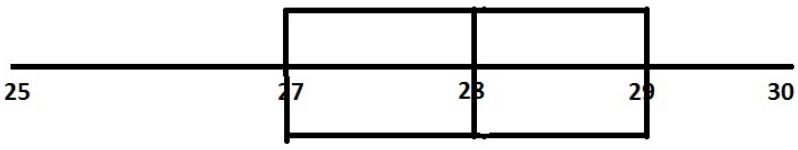
a. Describe some features of the histogram.

b. Is the mean household income expected to be greater than $53,700? Explain.

c. What percentage of household incomes greater than $206,600?

d. Can we place the household income of $160,000 in the top 10th percentile? Explain.

4. The box-plot of city mileage of compact cars is as follows.



a. What is the median mileage, and interpret its meanings.

b. Find IQR and interpret its meanings.

c. Can we say a car with mileage of 25 an outlier? Explain.

d. Can we say the mean should be close to 28? Explain.

5. The average height of trees in a forest is 33 feet with standard deviation 3 feet. Assume that the histogram of the heights of all trees in the forest is symmetric.

   a. Suppose that the z-core of the height of a tree is 1. What percentage of trees taller than this tree?

   b. Suppose that the z-score of a tree is -1.5. What is the height of the tree?

   c. What percentage of trees with height no more than 33 feet?

   d. What percentage of trees with height 30 feet or taller but no more than 36 feet?

   e. The first quartile of heights is 28 and the IQR = 4. What is the third quartile?

   f. Can we say a 20-foot tree an outlier? Explain.

6. A pair of 4-sided dice will be rolled.

   a. Write the sample space of all elementary events.

   b. What is the probability of observing eight?

   c. What is the probability of observing 5 or more?

   d. What is the most likely sum?

e. What is the probability of observing an even number?

7. Consider a lotto in which a player chooses 4 different numbers from the set

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15.$$

If all four selected numbers match the winning numbers, she/he will win the jackpot.

a. If a player buys only one ticket, what is the probability that he will win the jackpot?

b. What is the probability that he will lose the jackpot?

8. Four dice are to be rolled simultaneously.

a. What is the probability of observing the sum 24?

b. What is the probability of observing an even number?

c. What is the probability of observing a sum of 4 or more?

d. What is the probability of observing four up face numbers of same kind?

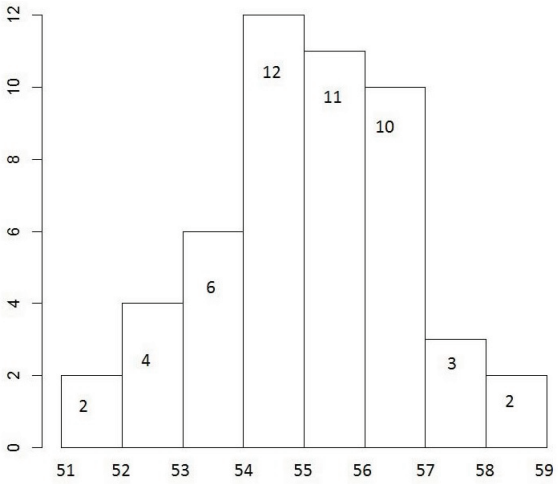
Review for the First Test

TOPICS	SUMMARY	
Types of variables	**Quantitative** Continuous: height, weight, salary, age, temperature Discrete: that can be counted; number of births in a hospital number of auto accidents in a day	**Qualitative** Nominal: race, religion, gender brand names, colors, etc. Ordinal: variables that can be ordered on a qualitative characteristic; ranks of college teachers, movie ratings, wine ratings, etc.
Summarizing Data Graphically	**bar chart**: Qualitative data; **histogram**: quantitative data; **features of histogram**: unimodal or bimodal; skewed to left or right; symmetric; modal class	
Summary Statistics	**Measures of Location**: Mean, Median and Mode median is preferred to the mean if the data (histogram) is skewed. Variance and Standard Deviation: Empirical rule; z score = $\frac{x-\bar{x}}{s}$ **Measures of Relative Standing**: Percentiles and z-scores 5-number statistics: min, Q1, median, Q3, max; IQR = Q3 – Q1 outliers: Lower boundary: $Q_1 - 1.5 \times IQR$ Upper boundary: $Q_3 + 1.5 \times IQR$	
Probability	event, sample space: probability(event) = $\frac{\text{number of outcomes in the event}}{\text{total number of ouctcomes in the sample sapce}}$	
Random Variable	Probability distribution of a random variable; $\mu = E(X)$ and $\sigma^2 = \text{var}(X)$	
Rules of Probability	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (for any two events) $P(A \cup B) = p(A) + P(B)$ (A and B are mutually exclusive) $P(A^c) = 1 - P(A)$; $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B)$ $P(A^c \cup B^c) = P((A \cap B)^c) = 1 - P(A \cap B)$	

1. The following data represent annual household incomes of 50 households in a neighborhood.

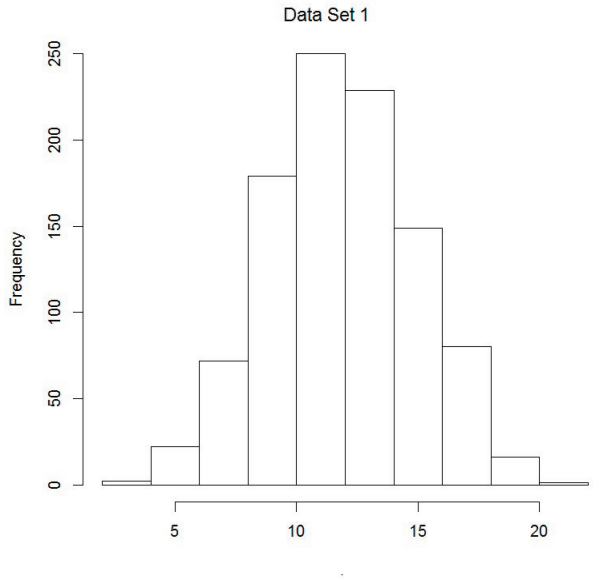
```
56 54 55 53 55 57 56 56 54 52
55 54 53 58 55 52 55 54 54 55
56 55 54 56 55 56 59 57 54 55
54 53 53 59 54 56 56 57 54 56
55 55 56 52 57 56 55 57 52 57
```

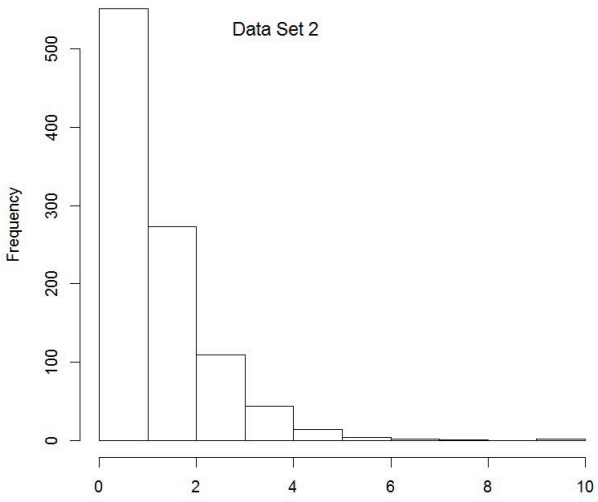
The histogram of the data:



a. Discuss the shape of the histogram.
b. What is the modal class?
c. Calculate the 5-number statistics, and interpret the meanings of the first quartile.
d. Find the IQR and interpret its meanings.
e. Identify the outliers if there are any.
f. Compute the mean and standard deviation.
g. Find the percentage of incomes fall in the interval (52, 56).

2. The following are the histograms of two data sets.





Compare the mean and median for data set 1.
Compare the mean and median for data set 2.

3. The average height of trees in a forest is 33 feet with standard deviation 4 feet. Assume that the histogram of the heights of all trees is symmetric.
   - **a.** Suppose that the z-core of the height of a tree is 1. What percentage of trees taller than this tree?
   - **b.** What percentage of trees with height no more than 33 feet?
   - **c.** What percentage of trees with height 30 feet or taller but no more than 36 feet?
   - **d.** What is the 99th percentile of the heights of all trees in the forest?
4. A pair of 4-sided dice is rolled. Let X denote the sum of the number that show up.
   - **a.** What is the total number of possible outcomes? List all the outcomes.
   - **b.** Write the probability distribution of X.
   - **c.** Find the probability $P(X = 8)$
   - **d.** Find $P(X \geq 5)$
   - **e.** Find $P(4 \leq X \leq 6)$
   - **f.** Find $E(X)$

4. Probability Distributions

4.1 Random Variable, Expectation and Variance

Definition 4.1 — Random variable A variable whose numerical values are subject to variation due to chance is called random variable. In other words, the values of a random variables are determined only by chance factors.

Examples of random variables include

- The number of heads that can be observed by tossing a coin, say, 10 times.
- The number of times a three can be observed in 5 rolls of a die.
- The height of a randomly selected student from a college.
- The annual income of a randomly selected family from a town.
- The number of children in a randomly selected household from a city.

Definition 4.2 — Probability Distribution A formula or table that assigns probability to each possible value of a random variable is called probability distribution.

■ **Example 4.1** Let X be the number of heads that can be observed by tossing 3 coins simultaneously. Here, the possible values of X are 0, 1, 2, or 3. To find the probability distribution of X, we first write the sample space as

$$\{HHH, THH, HTH, TTH, HHT, THT, HTT, TTT\}$$

For example, the outcome HTH means that the first coin shows up head, the second tail and the third head. The possible values of X are denoted by x, and the probability distribution of X simply lists the possible values and the corresponding probabilities as shown in Table4.1. Note that sum

Table 4.1: Probability distribution of the number of heads in a flip of three coins

x	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

of all probabilities is 1. ■

Definition 4.3 — Discrete random variable A random variable whose possible values are countable is called discrete random variable.

Some examples of discrete random variable are

- the number of auto accidents that may occur in a city during a day;
- the number of text messages that you receive in a day;
- the number of still births per day in a hospital;
- the number of heads that can be observed by flipping a coin 20 times. Here $X = 0, 1, 2, ..., 20$

Definition 4.4 — Mean and Standard Deviation For a discrete random variable X, the expectation (or mean) of X is denoted by μ and is defined by

$$\mu = E(X) = \sum xP(X = x)$$

and the variance of X is denoted by σ^2 and is defined as

$$\sigma^2 = \sum (x-\mu)^2 P(X = x) = \sum x^2 P(X = x) - \mu^2.$$

The square root of the variance, denoted by σ, is called the **standard deviation** of X.

■ **Example 4.2** Let X be the sum of numbers that can be observed by rolling a pair of dice.

a. Find the probability distribution of X.

b. Find the expected value μ and the standard deviation σ of X.

c. Find $P(X \leq 9)$ and $P(5 \leq X \leq 9)$.

Solution:

a. The set of all possible outcomes that will result when rolling a pair of dice is given in Table 3.1, which is the sample space. For convenience, the outcomes are given below.

```
1,1  1,2  1,3  1,4  1,5  1,6
2,1  2,2  2,3  2,4  2,5  2,6
3,1  3,2  3,3  3,4  3,5  3,6
4,1  4,2  4,3  4,4  4,5  4,6
5,1  5,2  5,3  5,4  5,5  5,6
6,1  6,2  6,3  6,4  6,5  6,6
```

The possible values of X are $x = 2, 3, 4, ..., 10, 11, 12$. The probability that X will assume, for example, 3 is given by

$$P(X = 3) = \frac{\text{number of outcomes add up to 3}}{\text{total number of outcomes}} = \frac{2}{36},$$

because only two outcomes, namely, (2, 1) and (1, 2) add up to 3. The other probabilities can be found similarly as given in Table 4.2.

b. From Table 4.2, we find the expectation of X as

$$\mu = E(X) = \sum xP(X = x) = \frac{252}{36} = 7.$$

The variance of X is given by

$$\sigma^2 = \sum x^2 P(X = x) - (E(X))^2 = \frac{1974}{36} - 7^2 = 54.83 - 49 = 5.83,$$

and the standard deviation of X is

$$\sigma = \sqrt{5.83} = 2.41.$$

Table 4.2: Probability distribution of X

x	$P(X=x)$	$xP(X=x)$	$x^2P(X=x)$
2	1/36	2/36	4/36
3	2/36	6/36	18/36
4	3/36	12/36	48/36
5	4/36	20/36	100/36
6	5/36	30/36	180/36
7	6/36	42/36	294/36
8	5/36	40/36	320/36
9	4/36	36/36	324/36
10	3/36	30/36	300/36
11	2/36	22/36	242/36
12	1/36	12/36	144/36
	1	252/36	1974/36

c. This probability is the cumulative probability up to 9, starting from the minimum possible value of X. That is,

$$\begin{aligned} P(X \leq 9) &= P(X=2)+P(X=3)+\ldots+P(X=9) \\ &= 1-[P(X=10)+P(X=11)+P(X=12)] \\ &= 1-\left(\frac{4}{36}+\frac{3}{36}+\frac{2}{36}\right) \\ &= 1-\frac{9}{36} \\ &= \frac{3}{4}. \end{aligned}$$

The probability $P(5 \leq X \leq 9)$ is the sum of all probabilities that X assumes all values between 5 and 9, including 5 and 9. That is,

$$\begin{aligned} P(5 \leq X \leq 9) &= P(X=5)+P(X=6)+P(X=7)+P(X=8)+P(X=9) \\ &= \frac{24}{36} = .667 \text{ (or } 66.7\%). \end{aligned}$$

■

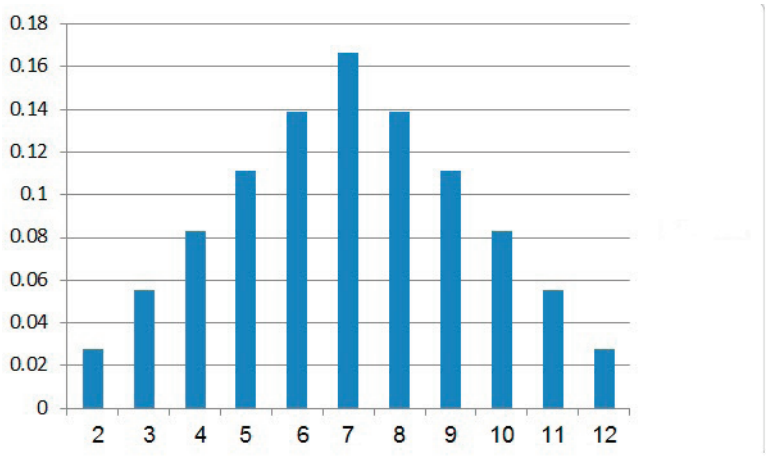


Figure 4.1: Probability distribution of the outcome in a roll of a pair of dice

■ **Example 4.3** Consider a gambling in which a pair of dice will be rolled. If a player bets a dollar, and if the outcome is 7 then she/he will win \$2, if it is 11, then the player will win \$6, else the player will lose the dollar. What is the expected amount of winning?

Solution: Let X denote the outcome of a roll. The probability distribution of X is given in Table 4.2. Let W denote the amount of winning. Then W is a random variable with the following possible values.

$$W = \begin{cases} 2 & \text{if } X \text{ is } 7, \\ 6 & \text{if } X \text{ is } 11, \\ -1 & \text{otherwise.} \end{cases}$$

So from Table 4.2, we find

$$W = \begin{cases} 2 \text{ with probability } P(X=7) = \frac{6}{36} \\ 6 \text{ with probability } P(X=11) = \frac{2}{36} \\ -1 \text{ with probability } 1 - \frac{6}{36} - \frac{2}{36} = 1 - \frac{8}{36} = \frac{28}{36} \end{cases}$$

So the expected winning amount is given by

$$\begin{aligned} E(W) &= 2 \times P(X=7) + 6 \times P(X=11) + (-1) \times P(X = \text{other numbers}) \\ &= 2 \times \frac{6}{36} + 6 \times \frac{2}{36} + (-1) \times \frac{28}{36} \\ &= \frac{12}{36} + \frac{12}{36} - \frac{28}{36} \\ &= -\frac{4}{36} \\ &= -\frac{1}{9}. \end{aligned}$$

x	$P(X=x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

That is, the expected gain is $\$-\frac{1}{9}$, or equivalently, the expected loss is $\$\frac{1}{9}$. This means that if a player plays this game 9 times, each time betting a dollar, then he/she is expected to lose \$1.

■

■ **Example 4.4** Let X denote the size of a random household in the UK in year 2013. Using the data in Example 1.10,

a. write the probability distribution of X.

b. Find the expected value $E(X)$ and the standard deviation of X.

c. Find $P(X \leq 2)$, and interpret its meanings.

d. Find $P(1 \leq X \leq 3)$, and interpret its meanings.

e. Is it unusual to find a household with six or more people? Explain.

Solution:

a. On the basis of data in Example 1.10, we write the probability distribution as shown in Table 4.3. Other quantities $xP(X=x)$ and $x^2P(X=x)$ to find the expected value and σ are also given in the following table.

Table 4.3: Probability distribution of the size of a random household in the UK

x	1	2	3	4	5	≥ 6	total
$P(X=x)$	0.290	0.345	0.164	0.136	0.045	0.020	1
$xP(X=x)$	0.290	0.690	0.492	0.544	0.225	0.120	2.36
$x^2P(X=x)$	0.290	1.380	1.476	2.176	1.125	0.120	6.567

b. From Table 4.3, we have

$$\begin{aligned} E(X) &= \sum xP(X=x) \\ &= 2.36 \end{aligned}$$

and

$$\begin{aligned} \sigma^2 &= \sum x^2 P(X=x) - [E(X)]^2 \\ &= 6.567 - (2.36)^2 \\ &= .9974. \end{aligned}$$

The std deviation $\sigma = \sqrt{.9974} = .9987$.

c. The cumulative probability is

$$\begin{aligned} P(X \le 2) &= P(X=2) + P(X=1) \\ &= .290 + .345, \quad \text{(from Table 4.3)} \\ &= .635. \end{aligned}$$

That is, the probability of observing a 1 or 2 people in a randomly selected household is 0.635.

d.

$$\begin{aligned} P(1 \le X \le 3) &= P(X=1) + P(X=2) + P(X=3) \\ &= .290 + .345 + .164 \quad \text{(from Table 4.3)} \\ &= 0.799. \end{aligned}$$

That is, the probability of observing a 1, 2 or 3 people in a randomly selected household is 0.799.

e. In Table 4.3, we see that $P(X \ge 6) = .02$. Yes, it is unusual because the probability of finding such household .02, and is less than .05.

■

■ **Example 4.5** The following table gives the probability distribution of the number of serious earthquakes X over a period of seventy five years. An earthquake is considered serious if its magnitude is at least 7.5 on Richter scale or at least 100 people were killed.

x	0	1	2	3	≥ 4
$P(X=x)$	.41	.38	.19	.01	.01

a. Find the probability of observing one or more serious earthquakes in a year. [Ans: .59]

b. What is the probability of observing at most 2 serious earthquakes in a year? [Ans: .98]

c. Find $E(X)$, and interpret its meanings.

$$E(X) = 0 \times .41 + 1 \times .38 + 2 \times .19 + 3 \times .01 + 4 \times .01 = 0.83$$

On average, 0.83 earthquake per year. This means that, on average, about 8 earthquakes per decade.

■

Properties of a Probability Distribution

1. Each probability should be nonnegative and no more than 1. That is,

$$0 \le P(X=x) \le 1.$$

2. Sum of the probabilities should be 1. That is,

$$\sum_x P(X = x) = 1.$$

■ **Example 4.6** Consider the probability distribution of a random variable X, which represents the number of auto accidents per day in a city.

No. of accidents per day x	0	1	2	3	4	5	6	≥ 7
$P(X = x)$	0.05	0.15	0.22	–	0.17	0.10	0.05	0.04

a. Find the probability of observing exactly three auto accidents in a day.

b. Find $E(X)$, and interpret its meanings.

■

4.2 Binomial Distribution

There are situations where experiments or surveys produce similar outcomes, which make it possible to develop formulas for the probability distribution of the outcomes. Many random experiments result into one of two mutually exclusive outcomes. For example, consider the following experiments: These two experiments are quite similar. The gender of a baby does not depend on

Experiment 1	Experiment 2
20 fair coins are flipped simultaneously X = number of heads	20 babies were delivered in a maternity hospital Y = number of boy babies

the genders of other babies, and similarly the outcome of a coin does not depend on the outcomes other coins. Each baby is a boy with probability approximately .5, and each coin shows up head with probability .5. So the probability distributions of X and Y must be similar. Other experiments whose outcomes can be described by similar probability distributions are

a. quality of a product (nondefective, defective).
b. marital status of a person (married or unmarried).
c. result of a medical treatment (successes or failure).
d. political affiliation of a person (Republican or Non-Republican).

Traditionally, these outcomes are labeled as "success" or "failure." The words success and failure are just labels and they do not invoke literal meanings. For example, while assessing the percentage of defective items in a batch, a quality control inspector may label a defective item as a "success" and a nondefective item as a "failure."

The binomial distribution is appropriate to model the number of outcomes that can be observed in a sequence of independent trials where each trial results into either "success" or "failure." Such

trials are called **Bernoulli trials**. For example, flip of a coin is a Bernoulli trial, and a binomial distribution can be used to model the number of heads out of, say, 10 flips of a coin.

Description of the Binomial Distribution

The binomial random variable and the binomial distribution are formally described as follows.

- Consider a sequence of n independent Bernoulli trials.
- Let p denote the probability of observing a success at each of the trial.
- Let X denote the number of successes observed out of these n trials

The random variable X is called the **binomial random variable** and its distribution is called binomial with the number of trials n and the success probability p.

Some examples of binomial random variable

1. The number of heads that can be observed by flipping a coin 10 times is a binomial random variable with $n = 10$ and $p =$ probability of observing a head at each flip = 0.5.
2. The number of female voters in a sample of 100 voters is a binomial random variable with $n = 100$ and $p =$ is the probability that a randomly selected voter is female.
3. The number of sixes that can be observed by rolling a die 25 times. This binomial random variable has $n = 25$ and $p = 1/6$, which is the probability of observing a six in a roll.
4. Suppose 5% of the items in a shipment are defective. If we select a sample of 10 items **with replacement** from the shipment, then the number of defective items in the sample is a binomial random variable with $n = 30$ and $p = .05$, which is the probability of selecting a defective item at each draw.

Remark

In Example 4 above, if the sample was selected **without replacement** then the number of defective items in the sample is not a binomial random variable, because the probabilities of selecting a defective item in the second and subsequent draws are not 5%, they depend on the goodness of the items in the preceding draws.

Result 4.1 — Binomial Probability Mass Function For the binomial random variable X with the number of trials n, and the success probability p,

$$P(X = k) = {}_nC_k \; p^k (1-p)^{n-k}, \;\; k = 0, 1, ..., n.$$

Read ${}_nC_k$ as "n choose k." For example, for a binomial random variable X with $n = 10$ and $p = .2$,

$$\begin{aligned} P(X = 3) &= {}_{10}C_3 \; (.2)^3 (1-.2)^{10-3} \\ &= 120(.2)^3(.8)^7 \\ &= 0.2013. \end{aligned}$$

The cumulative probability

$$\begin{aligned} P(X \le 3) &= P(X=3)+P(X=2)+p(X=1)+P(X=0) \\ &= {}_{10}C_3\,(.2)^3(1-.2)^{10-3}+{}_{10}C_2\,(.2)^2(1-.2)^{10-2}+{}_{10}C_1(.2)^1(1-.2)^{10-1} \\ &+ {}_{10}C_0\,(.2)^0(1-.2)^{10-0} \\ &= 120(.2)^3(.8)^7+45(.2)^2(.8)^8+10(.2)^1(.8)^9+1(.2)^0(.8)^{10} \\ &= 0.2013+.3020+.2684+.1073 \\ &= .879 \end{aligned}$$

TI Calc

Calculating Binomial Probabilities:

$$\begin{aligned} P(X=k) &= \text{binompdf}(n,p,k) \quad \text{[Press [2nd], [Distr], select binompdf()]} \\ P(X \le k) &= \text{cumulative sum of probabilties from 0 to } k \\ &= P(X=0)+P(X=1)+\ldots+P(X=k) \\ &= \text{bniomcdf}(n,p,k) \quad \text{[Press [2nd], [Distr], select binomcdf()]} \end{aligned}$$

For example, $n=10$, $p=.2$ and $k=3$, $P(X=3)=\text{binompdf}(10,.2,3)=.2103$, and $P(X\le 3)=\text{binomcdf}(10,.2,3)=.879$.

Remark

Calculation of other Probabilities: For a given n, p and k, the TI calculator calculates only

$$P(X\le k)=P(X=0)+P(X=1)+\ldots+P(X=k)=\text{binomcdf}(n,p,k).$$

To find $P(X\ge k)$ use the following relation:

$$P(X\ge k)=1-P(X\le k-1)=1-\text{binomcdf}(n,p,k-1),$$

because

$$\underbrace{P(X=0)+P(X=1)+\ldots+P(X=k-1)}_{P(X\le k-1)} + \underbrace{P(X=k)+P(X=k+1)+\ldots+P(X=n)}_{P(X\ge k)} = 1$$

If we need to find $P(X>k)$ then use the relation that

$$\begin{aligned} P(X>k) &= P(X\ge k+1) \\ &= P(X=k+1)+P(X=k+2)+\ldots+P(X=n) \\ &= 1-P(X\le k). \end{aligned}$$

Finally, note that

$$P(X<k)=P(X\le k-1)$$

because possible values of X are only integers, namely, $0,1,2,\ldots,n$.

■ **Example 4.7** Let X be the number of fives that can be observed in 14 rolls of a die.

a. Identify the random variable X.
b. Find the probability of observing exactly 4 fives;
c. no more than 6 fives;
d. at least 3 fives;
e. 3 or 4 fives.

Solution:

a. The X is a binomial random variable with

$$n = 14 \text{ rolls, and } p = \text{probability of observing 5 in a roll} = \tfrac{1}{6}.$$

b. $P(X = 4) = \text{binompdf}(14, 1/6, 3) = .1247$
c. $P(X \leq 6) = P(X = 6) + P(X = 5) + \ldots + P(X = 0) = \text{binomcdf}(14, 1/6, 6) = 0.9959$
d. $P(X \geq 3) = 1 - P(X \leq 2) = 1 - \text{binomcdf}(14, 1/6, 2) = 1 - 0.5795 = .4205$
e. $P(X = 3) + P(X = 4) = \text{binompdf}(14, 1/6, 3) + \text{binompdf}(14, 1/6, 4) = 0.2268 + 0.1247 = .3515$

■

Result 4.2 — The Mean and Standard Deviation For a binomial random variable with number of trials n, and success probability p, the mean is given by

$$\mu = \sum_{x=0}^{n} xP(X = x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np$$

and the variance is given by

$$\sigma^2 = \sum_{x=0}^{n} (x-\mu)^2 P(X = x) = \sum_{x=0}^{n} (x-\mu)^2 \binom{n}{x} p^x (1-p)^{n-x} = np(1-p).$$

The

$$\text{standard deviation } \sigma = \sqrt{\text{variance}} = \sqrt{np(1-p)}.$$

The binomial distribution is right skewed if $p < .5$, symmetric if $p = .5$ and left skewed if $p > .5$. See Figure 4.2.

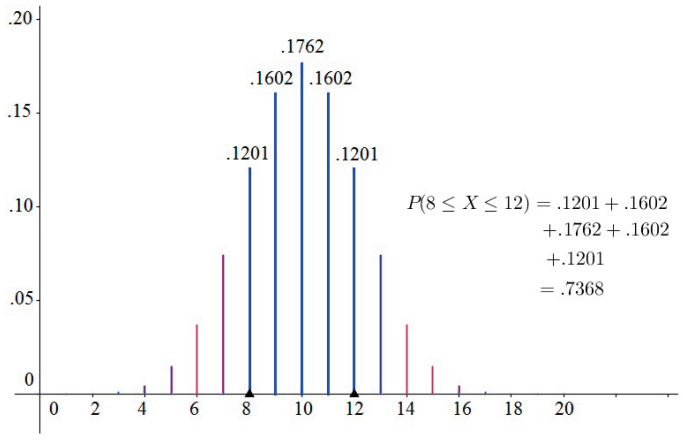


Figure 4.2: Binomial probabilities when $n = 20$, and $p = 0.5$

■ **Example 4.8** The blood type O negative is called the "universal donor" type. According to the Red Cross only 7% of U.S. residents have type O negative blood type. Suppose in a group of 50 donors recruited by a blood bank, let X denote the number of donors with blood type O negative.

a. Identify the random variable X.

b. What is the expected number of donors with O negative type blood?

c. What is the probability that 8 or fewer donors with blood type O negative?

d. What is the probability of 4 or more of them have type O negative blood?

e. Would it be unusual if none of the donors had type O negative blood?

Solution:

a. The X is a binomial random variable with $n = 50$ and $p = 7\% = .07$

b. $E(X) = np = 50 \times .07 = 3.5$

c. $P(X \leq 8) = \text{binomcdf}(n, p, k) = \text{binomcdf}(50, .07, 8) = 0.9927$

d. $P(X \geq 4) = 1 - P(X \leq 3) = 1 - \text{binomcdf}(50, .07, 3) = 1 - 0.5327 = 0.4673.$

e. $P(X = 0) = 0.0266$. Since this probability is less than .05, we can say it is unusual.

■

■ **Example 4.9** Consider a multiple-choice test consisting of 20 questions. For each question, there are 4 choices, and only one correct answer. A student, who has not prepared for the test, pickups the answer randomly for all questions. If the passing score is 13,

a. what is his expected score?

b. What are the chances that he will pass the test?

c. If 10,000 students take the test, and all of them guess the answer for each question, how many of them are expected to pass the test?

Solution: Let X be the score of the student. Since the student picks up the answers randomly, X is a binomial random variable with

$$n = 20 \quad \text{and} \quad p = 1/4 = .25,$$

which is the probability of choosing the correct answer.

a. So the expected score is $E(X) = n \times p = 20 \times .25 = 5$.

b. The chances that he will pass the test is

$$P(X \geq 13) = 1 - P(X \leq 12) = 1 - \text{binomcdf}(20, 1/4, 12) = 1 - 0.9998 = .0002.$$

c. $10000 \times .0002 = 2$.

■

■ **Example 4.10** A large shipment of items is submitted for inspection. The manufacturer assures that the percentage of defective items is no more than 0.5. A customer decided to use the following plan whether to buy or not to buy the entire shipment: He will inspect a sample of 100 items, and if he finds 2 or more defective items he will not buy the shipment. If the actual percentage of defective items is indeed 0.5, what is the probability that the buyer will buy the shipment.

Solution: Let X denote the number of defective items in a sample of 100 items. Notice that X is a binomial random variable with $n = 100$ and the probability

$$p = \text{probability of observing a defective item} = .5\% = .005.$$

If $X \geq 2$, the customer will not buy the shipment. So the probability that the customer will buy the shipment is given by

$$P(X \leq 1) = \text{binomcdf}(100, .005, 1) = 0.9102.$$

Thus, the probability that the customer will buy the shipment is 0.9102. ■

■ **Example 4.11** A poultry farm owner estimated that about 3% of egg cartons are getting damaged (at least one broken egg) during transportation to various retailers. Suppose a retailer bought 80 cartons of eggs from the poultry farm. Let X denote the number of damaged cartons among the 80 cartons.

a. Identify the random variable X.

b. How many damaged cartons does the retailer expect to have among 80?

c. Find the probability that at most 3 damaged cartons among 80, and interpret the meanings of this probability.

d. Would it be unusual if 5 or more damaged cartons are found among the 80 cartons?

Solution:

a. Here X is the binomial random variable with $n = 80$ and $p = .03$.

b. $E(X) = 80 \times .03 = 2.4$.

c. $P(X \leq 3) = P(X \leq 3) = \text{binomcdf}(80, .03, 3) = 0.7807$. This means that in 78% of such orders of 80 cartons, the retailer is expected to find three or less damaged cartons.

d. $P(X \geq 5) = 1 - P(X \leq 4) = 1 - \text{binomcdf}(80, .03, 4) = 1 - 0.9072 = 0.0928$. Since this probability is NOT less than .05, we can not say it is unusual.

■

Exercise 4.2

4.2.1 Let X be a binomial random variable with $n = 12$ and $p = 0.4$. Find the following probabilities.

a. $P(X = 5)$
b. $P(X \leq 5)$
c. $P(X \geq 6)$
d. $P(3 \leq X \leq 8)$

4.2.2 Let X denote the number of heads in 20 flips of a fair coin.

a. Find $P(X = 9)$
b. Find $P(X = 11)$
c. Explain why $P(X = 9) = P(X = 11)$
d. Find $P(X \leq 8)$ and $P(X \geq 12)$. Explain why these two probabilities are equal.

4.2.3 For year 2014, the 90th percentile of SAT combined scores for male college-bound seniors was 1,940. In a random sample of 200 college-bound seniors who took SAT in 2014,

a. find the expected number of students who scored 1,940 or less.
b. Find the probability that at most 160 students scored 1,940 or less.
c. Find the probability that at least 15 students scored 1,940 or more.

4.2.4 About 9% of people in the USA have type B positive blood. A sample of 100 people is selected at random.

a. What is the probability that exactly 9 of them in the sample have Type B positive blood?
b. Find the probability of observing 10 or more people with type B positive.
c. Is it unusual to find 3 or fewer people in the sample with type B positive blood? Explain.

4.2.5 A batch containing 20 electronic components is submitted for quality inspection. A random sample (without replacement) of 6 components was selected for inspection. Let X equal the number of defective components. Is X a binomial random variable? Explain.

4.2.6 Consider flipping a pair of balanced coins 16 times. Let X denote the number of times both coins showup head.

a. Identify the random variable X.

b. What is the expected value of X.

c. Find $P(X \leq 4)$.

d. Find $P(X \geq 5)$.

4.2.7 Consider rolling a pair of dice 30 times. Let X denote the number of times the sum of the numbers is even.

a. What is the distribution of X?

b. Find $E(X)$.

c. Find $P(X \leq 14)$

d. What is the probability that X assumes a value of 20 or more?

4.2.8 The branch manager of a departmental stores estimated that 60% of customers spent \$100 or more in the store. Let X denote the number of customers who will spend \$100 or more in a sample 50 customers who are shopping at the store.

a. Find $E(X)$ and interpret its meanings.

b. Find $P(X \leq 30)$.

c. Find $P(35 \leq X \leq 45)$.

4.2.9 It is reported that about 65% internet users use Google as the search engine. In a sample of 30 internet users, let X denote the number of users of Google.

a. Identify the random variable X.

b. What is the expected number of users of Google in the sample?

c. What is the probability that at most 20 of them will use Google?

d. What is the probability that at least 10 of them will use Google?

e. Would it be unusual if 9 or fewer use Google?

4.2.10 A credit card company estimated that about 5.6% of customers are not paying their dues on time. Let X denote the number of customers who are late in paying their dues on time in a sample of 120 customers.

a. Identify the random variable X.

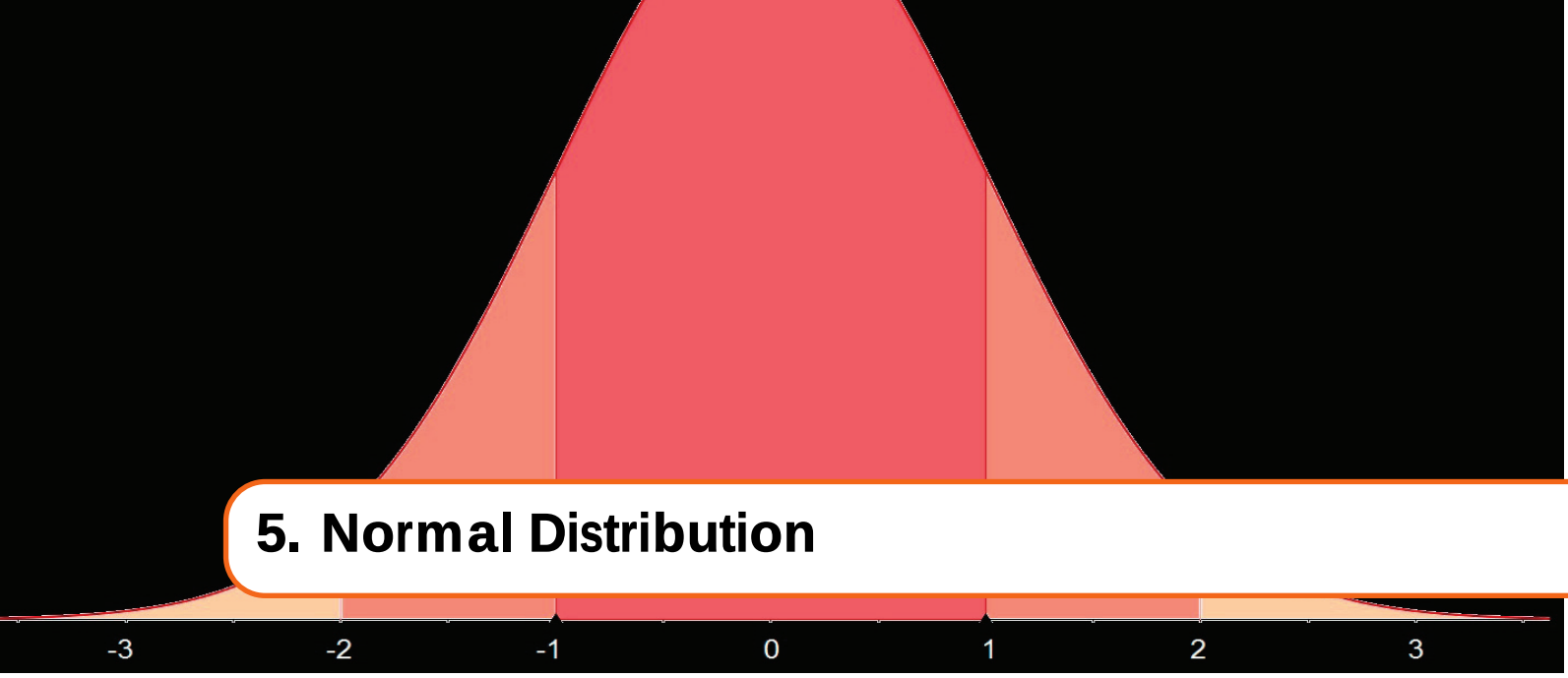
b. Find the expected number of customers who are late in paying their dues on time in the sample.

c. Let Y denote the number of customers in the sample who pay their dues on time. Identify the random variable Y.

d. Find $E(Y)$

e. What is the probability of observing at most 90 customers in the sample who pay their dues on time?

f. Find $P(110 \leq Y \leq 116)$.



5. Normal Distribution

Normal distribution is used to model the data that is approximately symmetric. This distribution is determined by only two parameters, namely, mean μ and variance σ^2. This distribution is symmetric about the mean μ, and so the median is also μ. Also, this distribution can be used to approximate other probability distributions.

If the histogram of z-scores is symmetric and unimodal (bell shaped curve), then the distribution of z-scores are approximated by a normal model with mean 0 and variance 1, referred to as the standard normal curve. The standard normal curve is determined by the function

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

The standard normal curve is illustrated in Figure 5.1.

TI Calc

Calculation of Standard Normal Probabilities

In general, for a normal random variable X with mean μ and standard deviation σ,

$$P(a \le X \le b) = \text{normcdf}(a, b, \mu, \sigma).$$

Select [DISTR] by pressing [2nd] and [DISTR]; then select **normcdf**
enter the values of a, b, mean, and standard deviation
press [ENTER].

TI Calc

Calculating Probabilities in Figure 5.2

$P(-1 \le Z \le 1) = \text{normcdf}(-1, 1, 0, 1) = .6827$
$P(-2 \le Z \le 2) = \text{normcdf}(-2, 2, 0, 1) = .9545$
$P(-3 \le Z \le 3) = \text{normcdf}(-3, 3, 0, 1) = .9973$

1. Press [2nd] and [DISTR]
2. Select [2: normalcdf( ]
3. normalcdf$(-1, 1, 0, 1)$, [Enter]
4. .6827

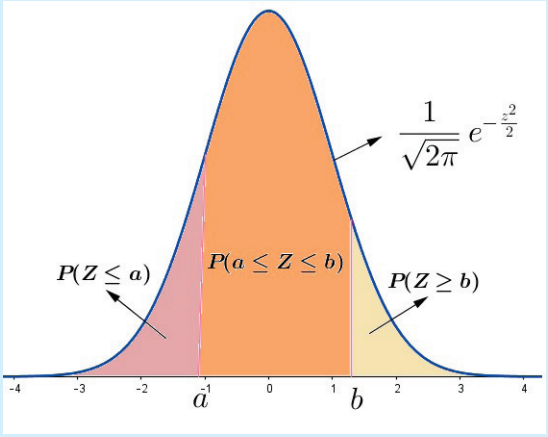


Figure 5.1: Areas under the standard normal curve between a and b

The areas under the standard normal curve bounded by the interval (a,b) is highlighted by orange color, and this area is the probability that the standard normal random variable Z falls between a and b, that is, $P(a \le Z \le b)$. The entire area under the curve is 1, so that

$$P(Z \le a) + P(a \le Z \le b) + P(Z \ge b) = 1.$$

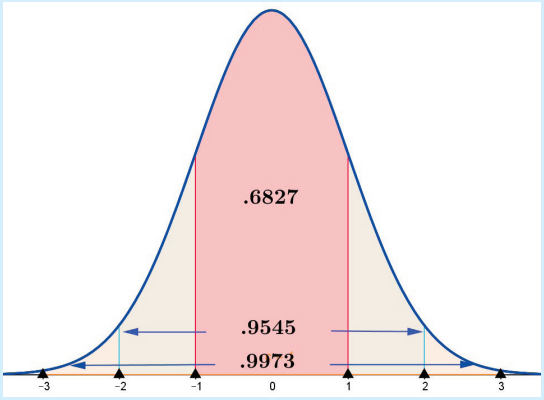


Figure 5.2: Areas under the standard normal curve

The areas under the standard normal curve bounded by the intervals $(-1,1)$, $(-2,2)$ and $(-3,3)$ are illustrated in Figure 5.2. That is,
$P(-1 \le Z \le 1) = .6827$,
$P(-2 \le Z \le 2) = .9545$, and
$P(-3 \le Z \le 3) = .9973$.
The empirical rules were obtained by rounding the above probabilities to two decimal places. Specifically, about 68% of z-scores are in $(-1,1)$, 95% of z-scores are in $(-2,2)$ and 99.7% of z-scores are in $(-3,3)$.

■ **Example 5.1** Let Z be a standard normal random variable. Find the following probabilities.

a. $P(Z \ge -1)$ [Ans: normcdf$(-1, 10^7, 0, 1) = .8413$]

b. $P(Z \le 1.4)$ [Ans: normcdf$(-10^7, 1.4, 0, 1) = .9192$]

c. $P(1 \le Z \le 2)$ [Ans: normcdf$(1, 2, 0, 1) = .1359$]

d. $P(-1 \le Z \le -2)$ [Ans: normcdf$(-1, -2, 0, 1) = .1359$]

■

Remark

In TI-84, while computing the normcdf(), the default values of the mean μ and standard deviation σ are 0 and 1, respectively. For example, normacdf$(-1, 10^7, 0, 1) = .8413$ and normcdf$(-1, 10^7) = .8413$. When the mean is different from zero or the standard deviation is different from 1, we should enter these values in normcdf(). It may be a good practice to enter the mean and standard deviation always, regardless of their values.

■ **Example 5.2** Let X be a normal random variable with mean 6 and the standard deviation 2.

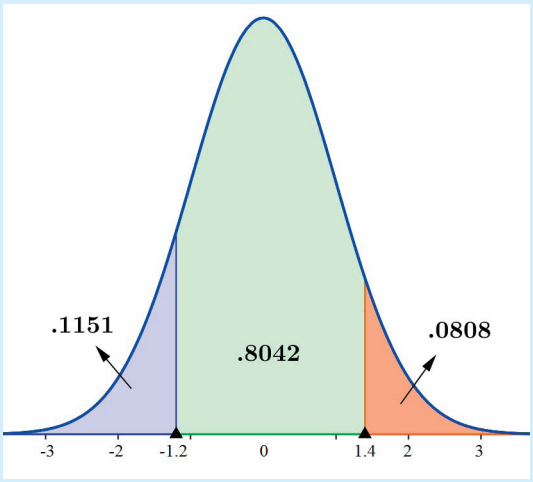


Figure 5.3: Percentage of z-scores between -1.2 and 1.4

The probabilities in the tail areas and in the center are illustrated in Figure 5.3. The area on the left of -1.2 is

$$P(Z \le -1.2) = .1151,$$

the area on the right of 1.4 is

$$P(Z \ge 1.4) = .0808,$$

and the area between -1.2 and 1.4 is

$$P(-1.2 \le Z \le 1.4) = .8042.$$

These areas can be calculated using TI-84 as follows. Left tail area below -1.2 is normcdf$(-10^7, -1.2, 0, 1)$, center area between -1.2 and 1.4 is normcdf$(-1.2, 1.4, 0, 1)$ and the right area above 1.4 is normcdf$(1.4, 10^7, 0, 1)$.

Note: We used 10^7 for ∞.

Find the following probabilities.

a. $P(X \le 8)$ [Ans: normcdf$(-10^7, 8, 6, 2) = .8413$]
b. $P(X \ge 8)$ [Ans: normcdf$(8, 10^7, 0, 1) = .1587$; this is also $1 - P(X \le 8) = 1.8413$]
c. $P(4 \le X \le 8)$ [Ans: normcdf$(4, 8, 6, 2) = .6827$]
d. $P(2 \le X \le 10)$ [Ans: normcdf$(2, 10, 6, 2) = .9547$]

Note that, in part **c**, the probability in the interval

$$(4, 8) = 6 \mp 2 = \text{mean} \pm 1 \times \text{Std Deviation}.$$

That is, the probability within one standard deviation from the mean, and is approximately .68. The probability in part **d** is within 2 standard deviation from the mean, and the probability is approximately .95. ■

■ **Example 5.3** The IQ scores are normally distributed with mean 100 and standard deviation 16. What % of IQ scores are

a. above 116? [Ans. normalcdf$(116, 10^7, 100, 16) = .1587$; about 16%]
b. above 120? [Ans. normalcdf$(120, 10^7, 100, 16) = .1057$; about 11%]
c. below 80? [Ans. normalcdf$(-10^7, 80, 100, 16) = .1057$; about 11%]
d. between 68 and 132? [Ans. normalcdf$(68, 132, 100, 16) = .9545$; about 95%]

■

Advantages of Modeling

There are several advantages of modeling data by a normal distribution. For example, if a normal model is appropriate for a data set, then what we need are only the mean and standard deviation of the data. Using the mean and the standard deviation, we can calculate

- the five-number statistics
- various percentiles
- the percentage of data fall in a given interval.

■ **Example 5.4** The following is the histogram of heights of a sample of 1,000 male students from a university. The mean height is 68 inches with the standard deviation of 2.4 inches. Let X denote

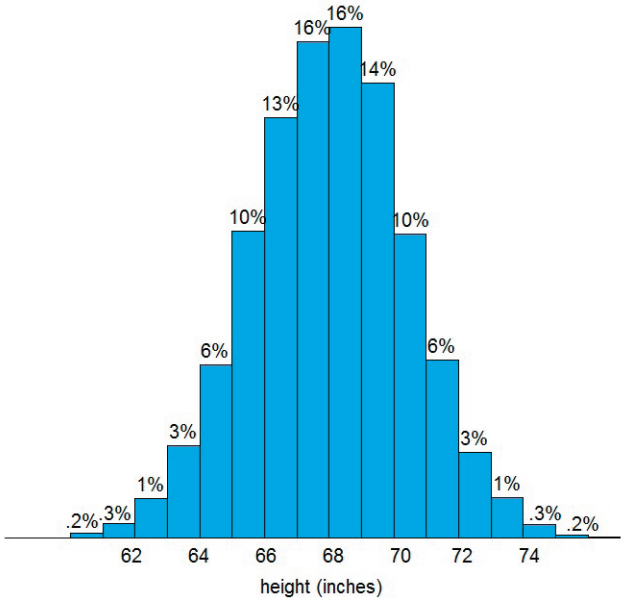


Figure 5.4: Relative frequency histogram of heights 1,000 college students

the height of a randomly selected male student from the university. As the histogram of the height distribution is very symmetric, we can assume that X is approximately normally distributed with mean 68 and standard deviation 2.4. In other words, we use the normal model with mean 68 and standard deviation 2.4 to describe the height distribution.

One of the purposes of modeling data is to use the model to estimate the percentages of students with heights in different intervals. For example, we find from the above histogram that percentage of students who are 66 inches or shorter is

$$10\% + 6\% + 3\% + 1\% + .3\% + .2\% = 20.5\%.$$

Using the assumed normal model, we find

$$\begin{aligned} P(X \le 66) &= \text{normcdf}(-10^7, 66, 68, 2.4) \\ &= .202, \end{aligned}$$

which is 20.2%, and is very close to the true percentage 20.5; see Figure 5.4. Similarly, we see that the percentage students who are 6 feet or taller is $3\% + 1\% + .3\% + .2\% = 4.5\%$, whereas the model estimate is

$$\begin{aligned} P(X \ge 72) &= \text{normcdf}(72, 10^7, 68, 2.4) \\ &= .048, \end{aligned}$$

which is 4.8% and is close to the true percentage of 4.5%; see Figure 5.4. An advantage of the modeling is that the percentages in intervals such as [68.5, 70.5] can be easily estimated as

$$\begin{aligned} P(68.5 \le X \le 70.5) &= \text{normcdf}(68.5, 70.5, 68, 2.4) \\ &= .269. \end{aligned}$$

Notice that it is not easy to estimate the above probability from the histogram in Figure5.4. Furthermore, all 1,000 data are required to construct the histogram, and then estimate the probabilities in different intervals. On the other hand, the normal approximation is quite convenient, and it requires only the mean and standard deviation of the data to estimate the probabilities, quartiles or median of the heights of all male students in the university (see Example5.7). ■

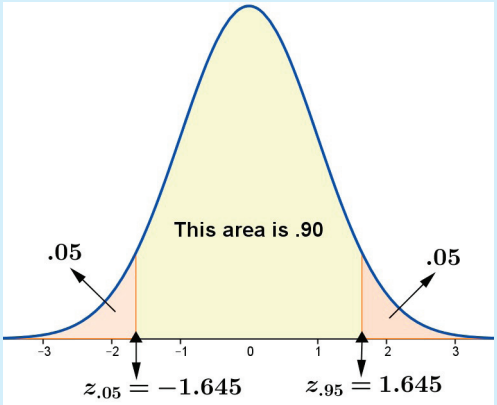


Figure 5.5: Illustration of the lower and the upper fifth percentage points

The lower and upper 5th percentiles of a standard normal distribution are illustrated in Figure 5.5. The lower 5th percentile denoted by $z_{.05}$, is determined so that $P(Z \leq z_0) = .05$, and the value of $z_{.05}$ is -1.645. The upper 5th percentile (or 95th percentile) is denoted by $z_{.95}$, and is determined so that $P(Z \leq z_{.95}) = .95$. The value of $z_{.95}$ is 1.645.

TI Calculator: invNorm(.95,0,1) = 1.645; invNorm(.05,0,1) = -1.645.

Percentiles of the Standard Normal Distribution

The $100p$th percentile of the standard normal distribution is the value z_p determined by

$$P(Z \leq z_p) = p.$$

That is, the value of z for which the left-tail probability is equal to p is called the $100p$th percentile of the standard normal distribution.

■ **Example 5.5** Find the following percentiles of the standard normal distribution.

a. The 90th percentile.
b. The top 5th percentile.
c. The lower 5th percentile.
d. The 25th percentile.
e. Find the quartiles, median and the inter quartile range (IQR).

Solution:

a. The 90th percentile is denoted by $z_{.90}$, and is the value of z for which the left-tail probability $P(Z \leq z) = .90$. Using TI, it is

$$\text{invNorm}(.90, 0, 1) = 1.282.$$

b. The top 5th percentile is simply the 95th percentile, which is

$$\text{invNorm}(.95, 0, 1) = 1.645.$$

c. The 5th percentile is

$$\text{invNorm}(.05, 0, 1) = -1.645.$$

d. The 25th percentile is

$$\text{invNorm}(.25, 0, 1) = -.674.$$

e. The quartiles and median are

$$Q_1 = \text{invNorm}(.25, 0, 1) = -.674, \quad Q_3 = \text{invNorm}(.75, 0, 1) = .674,$$

and the

$$\text{median} = \text{invNorm}(.5, 0, 1) = 0.$$

So the IQR = $Q_3 - Q_1 = .674 - (-.674) = 1.348$. The middle 50% of the standard normal distribution is in $(-.674, .674)$.

■

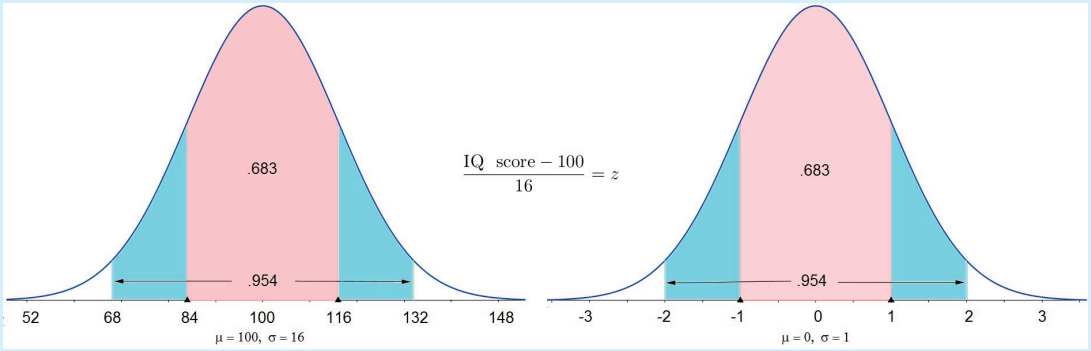


Figure 5.6: Normal distributions of IQ scores (left) and z-scores (right)

The above two graphs show that the distributions of the IQ scores and the corresponding z-scores are similar.

■ **Example 5.6** The owner of a convenient store find that the daily average demand for a particular brand bread is 36 loafs with standard deviation 2.56. In order to meet the demand for the bread for 99% of the days, how many loafs should the owner keep in the store?

Solution: Let x denote the number of loafs such that 99% of the days the demand for the bread is less than or equal to x. This means that x is the 99th percentile, and it can be calculated as

$$\text{invNorm}(.99, 36, 2.56) = 41.956 \approx 42.$$

Thus, to meet the demand for 99% of the days, the owner should keep 42 loafs. ■

■ **Example 5.7** Estimate the quartiles, median and IQR for the height distribution in Example5.4.
Solution: Recall that the mean is $\mu = 68$ inches and the standard deviation is 2.4 inches. So the first quartile is

$$Q_1 = \text{invNorm}(.25, 68, 2.4) = 66.38$$

and the third quartile is

$$Q_3 = \text{invNorm}(.75, 68, 2.4) = 69.62.$$

The

$$IQR = Q_3 - Q_1 = 69.62 - 66.38 = 3.24.$$

Recall that the median is the 50th percentile, and is calculated as

$$\text{median} = \text{invNorm}(.5, 68, 2.4) = 68.$$

Note that the median is the same as the mean, because the normal distribution is symmetric about its mean. ■

■ **Example 5.8** SAT verbal test scores are approximately symmetric with mean 500 and standard deviation 100. Suppose the administrators of a college decided to admit only students with SAT verbal scores in the top 10th percentile. How high a score does it take to be eligible for admission?

Solution: We need to find the 90th percentile of normally distributed data with mean 500 and std deviation 100. This is calculated as

$$\text{invNorm}(.90, 500, 100) = 628.155$$

Thus, to get admitted in the college, the SAT score should be 628 or more. ■

■ **Example 5.9** Heights in inches for American males aged 20 and over are approximately normally distributed (symmetric) with the mean height 69.3 inches and std deviation 2.99 inches.

a. What percentage of American males in the above age group who are 6 feet or taller?

b. Find the 99th percentile of the American males in the above age group and interpret it.

c. What is the median?

d. What is the inter quartile range (IQR)?

e. Suppose you are an American male aged 20 or more, and your height is 74 inches. What is the percentile of your height and interpret its meanings.

Solution:

a. Greater than or equal to 6 feet (72 inches); $\text{normcdf}(72, 10^7, 69.3, 2.99) = 0.1833$. This means that 18.33% of American males in the age group 20 or more are 6 feet or taller.

b. 99th percentile = invNorm(.99, 69.3, 2.99) = 76.2558. This means that 99% of all American males in the age group 20 plus are 76.3 inches or shorter. In other words, 1% of American males aged 20 or more are 76.3 inches or taller.

c. Since the data are approximately symmetric, the median should be approximately equal to 69.3 inches.

d. Recall that $IQR = Q_3 - Q_1$, that is the difference between the 75th percentile and the 25th percentile. This is equal to

$$\text{invNorm}(.75, 69.3, 2.99) - \text{invNorm}(.25, 69.3, 2.99) = 71.32 - 67.28 = 4.04.$$

e. To find the percentile, we find the percentage of males in the age group who are 74 inches or shorter. This is

$$\text{normcdf}(-10^7, 74, 69.3, 2.99) = .9420.$$

This means that your height is in the 94th percentile; that is, $.9420 \times 100 \approx 94\%$ of the American males in the age group are as tall as you or shorter.

■

■ **Example 5.10** The average time for a particular desktop computer needs first service is 390 days with standard deviation of 20 days. The manufacturer gives one year (365 days) onsite service warranty for newly bought computers. Assume that the times for first service call are approximately normally distributed.

a. If 1,000 computers were sold on a particular day, how many computers are expected to have first service within the warranty period?

b. Determine the warranty period so that the manufacturer is expected to service about 5% of the computers within the warranty period.

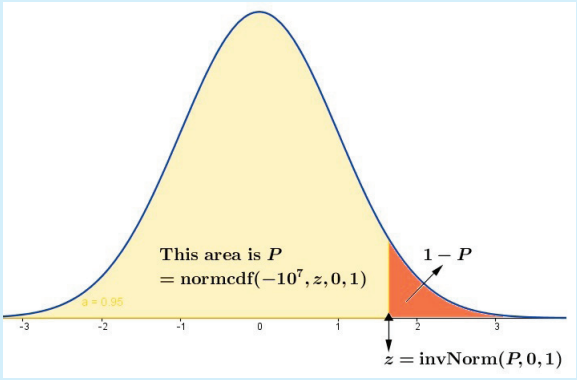


Figure 5.7: Illustration of normal percentage point

Calculation of areas under the standard normal curve and the percentage point are illustrated in Figure 5.7. Suppose the area on the left of a point is .9192, then the corresponding percentage point is 1.4. This value can be calculated using TI-83/84 as $\text{invNorm}(.9192, 0, 1)$.

Solution: Let X denote the time until the first service for a computer produced by the manufacturer. The times for the first service call are approximately normal with mean

$$\mu = 390 \quad \text{and} \quad \sigma = 20.$$

a. The probability that a computer needs a first service call is

$$P(X \le 365) = \text{normcdf}(-10^7, 365, 390, 20) = .1057.$$

That is, 10.6% of the computers sold need some types of service within the warranty period. The expected number of computers that need service within the warranty period is

$$1000 \times .1057 = 105.7 \approx 106.$$

b. Let x denote the warranty period so that only 5% of computers need service within the warranty period. This means that only 5% of "times to first service" should be less than or equal to x. That is, x should be determined so that

$$P(X \le x) = .05.$$

Thus, x is the 5th percentile of the normal distribution with mean 390 and the standard deviation 20, and is calculated as

$$x = \text{invNorm}(.05, 390, 20) = 357 \text{ days}.$$

■

Normal Approximation to the Binomial Distribution

Consider a binomial random variable X with number of trials n and success probability p. Then X has an approximate normal distribution with

$$\text{mean} = \mu = np \quad \text{and} \quad \sigma = \sqrt{np(1-p)}.$$

Furthermore,

$$P(X \le k) \approx \text{normcdf}(-10^7, k + .5, \mu, \sigma)$$

and

$$P(X \ge k) \approx \text{normcdf}(k - .5, 10^7, \mu, \sigma).$$

This approximation is quite good if p is around .5, because the binomial distribution is symmetric about the mean when $p = .5$. In general, this normal approximation is good if

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5.$$

For example, let X be a binomial random variable with $n = 20$ and $p = .7$. Then, the

$$np = 20 \times .7 = 14 \quad \text{and} \quad n(1-p) = 20 \times .3 = 6.$$

So, for this case, normal approximation is good to find binomial probabilities. The

$$\text{mean} = np = 20 \times .7 = 14 \quad \text{and} \quad \sigma = \sqrt{np(1-p)} = \sqrt{20 \times .7 \times .3} = 2.049.$$

For instance,

$$P(X \leq 12) = \text{binomcdf}(20, .7, 12) = .2277 \text{ and } P(X \geq 15) = 1 - \text{binomcdf}(20, .7, 14) = .4164.$$

Using the normal approximation, we find

$$\begin{aligned} P(X \leq 12) &= \text{normcdf}(-10^7, k + .5, \text{mean}, \sigma) \\ &= \text{normcdf}(-10^7, 12.5, 14, 2.049) \\ &= .2321, \end{aligned}$$

which is very close to the exact probability .2277. Similarly,

$$P(X \geq 15) = \text{normcdf}(14.5, 10^7, 14, 2.049) = .4036,$$

which is close to the exact probability .4164.

Chapter Exercise 5

5.1 It is known that the IQ scores are symmetrically distributed (like z-scores) with mean 100 and $s = 16$.

- **a.** Find the probability of selecting an adult with IQ score no more that 90.
- **b.** Find the probability of selecting an adult with IQ score 100 or more.
- **c.** Let X denote the number of adults with IQ scores 110 or more in a sample of 50 adults. Identify the random variable X.
- **d.** Find the expected number of adults with IQ score 110 or more in the sample.
- **e.** What is the probability of observing at most 10 adults with the IQ scores 110 or more in the sample?
- **f.** Find $P(10 \leq X \leq 16)$.

Review: Chapters 4 and 5

TOPICS	SUMMARY
Binomial Distribution	X is binomial rv with n trials and success probability p $E(X) = np$ and $\text{var}(X) = \sigma^2 = np(1-p)$ $P(X = k) = \text{binompdf}(n, p, k)$ $P(X \le k) = \text{binomcdf}(n, p, k)$; $P(X \ge k) = 1 - P(X \le k-1) = 1 - \text{binomcdf}(n, p, k-1)$
Normal Distribution Probabilities	X is normal rv with mean μ and std deviation σ $P(X \le x) = \text{normcdf}(-10^7, x, \mu, \sigma)$ $P(X \ge x) = \text{normcdf}(x, 10^7, \mu, \sigma)$ $P(a \le X \le b) = \text{normcdf}(a, b, \mu, \sigma)$
Percentiles	$x_p = \text{invNorm}(p, \mu, \sigma)$ 90th percentile = top 10th percentile $x_{.9} = \text{invNorm}(.9, \mu, \sigma)$ For std normal percentile, $\mu = 0,\ \sigma = 1$

Review Problems

1. Let X denote the number of auto accidents per day in a city. The following table gives the probability distribution of X.

No. of accidents, x	0	1	2	3	4	5	6	≥ 7
$P(X = x)$	.02	.07	.15	.20	.20	.16	.11	.09

a. Find $P(X \le 4)$.

$$.02 + .07 + .15 + .20 + .20 = .64$$

b. Find the probability of observing 3 or more accidents in a day.

$$.20 + .20 + .16 + .11 + .09 = .76$$

c. Find $E(X)$, and interpret its meanings.

$$0 \times .02 + 1 \times .07 + 2 \times .15 + 3 \times .20 + 4 \times .20 + 5 \times .16 + 6 \times .11 + 7 \times .09 = 3.86$$

On average, there are 3 to 4 auto accidents per day.

d. What are the most likely outcomes? Explain.
3 or 4 auto accidents per day, because these are the numbers with high probability of .20

e. In a year, how many days do we expect to have five or more accidents?
First, we need to find the percentage of days we expect 5 or more accidents. That is,

$$P(X \ge 5) = .16 + .11 + .09 = .36.$$

So, 36% of days in a year. Since there are 365 days in a year, we can expect

$$.36 \times 365 = 131.4 \approx 131$$

days in a year with 5 or more accidents.

2. Let X be a binomial random variable with $n = 12$ and $p = 0.4$. Find the following probabilities.
   a. $P(X = 5)$
   binompdf(n, p, k) = binompdf(12, .4, 5) = .2270
   b. $P(X \leq 5)$
   binomcdf(n, p, k) = binompdf(12, .4, 5) = .6652
   c. $P(X \geq 6)$

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - \text{binompcdf}(12, .4, 5) = 1 - .6652 = .3348$$

   d. $P(3 \leq X \leq 8)$

$$\begin{aligned} P(X \leq 8) - P(X \leq 2) &= \text{binompcdf}(12, .4, 8) - \text{binompcdf}(12, .4, 2) \\ &= .9847 - .0834 = .9013 \end{aligned}$$

3. Let X be a binomial random variable with $n = 5$ and $p = .5$. The probability distribution is shown in the following table. Find the mean and variance of X.

Table 5.1: Probability distribution of binomial$(n = 5, p = .5)$ random variable

k	0	1	2	3	4	5
$P(X = k)$	0.03125	0.15625	0.31250	0.31250	0.15625	0.03125

Note that, for a binomial random variable

$$\text{mean} = np = 5 \times .5 = 2.5 \quad \text{and} \quad \text{var}(X) = np(1-p) = 5 \times .5 \times (1 - .5) = 5 \times .5 \times .5 = 1.25.$$

4. A credit card company estimated that about 5.6% of customers are not paying their dues on time. Let X denote the number of customers who are late in paying their dues on time in a sample of 120 customers.
   a. Identify the random variable X.
   X is binomial with $n = 120$ and

$$p = P(\text{that a cutomer pay his/her due late}) = .056$$

   b. Find the expected number of customers who are late in paying their dues on time in the sample.
   $np = 120 \times .056 = 6.72$; expected number of customers in the sample who pays dues late is about 7.
   c. Let Y denote the number of customers in the sample who pay their dues on time. Identify the random variable Y.
   Y is binomial with $n = 120$ and

$$p = P(\text{that a customer pays dues on time}) = 1 - .056 = .944.$$

d. Find $E(Y)$

$$E(Y) = 120 \times .944 = 113.28; \quad \text{about 113 customers.}$$

e. What is the probability of observing at most 90 customers in the sample who pay their dues on time?

$$P(Y \leq 90) = \text{binomcdf}(120, .944, 90) = 3.193314 \times 10^{-12} = 0$$

f. Find $P(110 \leq Y \leq 116)$.
From part **c**, we know that Y is binomial with $n = 120$ and $p = .944$. So

$$\begin{aligned} P(Y \leq 116) - P(Y \leq 109) &= \text{binompcdf}(120, .944, 116) - \text{binompcdf}(120, .944, 109) \\ &= .9088 - .0740 = .8348 \end{aligned}$$

5. The percentage of uninsured motorists in Louisiana is estimated as 12. In a sample of 100 Louisiana drivers, let X denote the number of uninsured drivers.
   a. Identify the random variable X.
   X is binomial with $n = 100$ and $p = 12\% = .12$
   b. Find $E(X)$, and interpret its meanings.

$$E(X) = n \times p = 100 \times .12 = 12$$

   In a sample of 100 LA drivers, we can expect about 12 uninsured drivers.
   c. Find the probability $P(X \leq 12)$
   binomcdf(100,.12,12) = .5761
   d. Find $P(X \geq 22)$

$$1 - \text{binomcdf}(100, .12, 21) = 1 - .9966 = .0034$$

   e. Is it unusual to observe 22 or more uninsured drivers in a sample 100 LA drivers? Explain.
   Yes, because the probability in part **d** is .0034, which is far less than .05.
   f. Find $P(8 \leq X \leq 16)$, and interpret its meanings.

$$\begin{aligned} P(X \leq 16) - P(X \leq 7) &= \text{binompcdf}(100, .12, 16) - \text{binompcdf}(100, .12, 6) \\ &= .9126 - .0761 = .8365 \end{aligned}$$

   The probability of observing 8 to 16 uninsured drivers in a sample of 100 LA drivers is .8365.
6. Consider rolling a pair of dice 30 times. Let X denote the number of times the sum of the numbers is even.
   a. What is the distribution of X?
   X is binomial (odd or even) with $n = 30$ and

$$p = P(\text{observing an even number in a roll}) = \frac{18}{36} = \frac{1}{2}.$$

b. Find $E(X)$.

$$n \times p = 30 \times \frac{1}{2} = 15.$$

c. Find $P(X \leq 14)$

$$\text{binomcdf}(n, p, k) = \text{binomcdf}(30, 1/2, 14) = .4278$$

d. What is the probability that X assumes a value of 20 or more?

$$P(X \geq 20) = 1 - P(X \leq 19) = 1 - \text{binomcdf}(30, .5, 19) = 1 - .9506 = .0494.$$

7. The IQ scores of adults are normally distributed with mean 100 and standard deviation 16.

a. what is the expected IQ score of a randomly selected person?
expected value = mean = 100

b. What is the probability that the IQ score of a randomly selected person fall between 84 and 116?
Since X is normal with mean 100 and standard deviation 16,

$$P(84 \leq X \leq 116) = \text{normcdf}(84, 116, 100, 16) = .6827.$$

c. Find the 95th percentile of the IQ score, and interpret its meanings.

$$\text{invNorm}(.95, 100, 16) = 126.3$$

That is, 95% of adults have IQ scores 126 or less.

d. Is it common to find a person with IQ score 140 or more? Explain.
Since the probability $P(\text{IQ score} \geq 140) = \text{normcdf}(140, 10^7, 100, 16) = .0062$ which is much less than .05, we can say, it is unusual for a person with an IQ score of 140 or more.

e. Let X denote the number of adults with IQ scores between 84 and 116 in a sample of 100 adults. What is the distribution of X?
Here, X is number of adults (discrete) who are with IQ scores in the interval $(84, 116)$. Thus, X is binomial with $n = 100$ and

$$p = P(84 \leq \text{IQ score} \leq 116) = .6827.$$

f. In a random sample of 100 adults, what is the expected number of adults with IQ scores between 84 and 116?
From part **e**, we see that X is binomial with $n = 100$ and $p = .6827$, so $E(X) = np = 100 \times .6827 = 68.27$. That is, 68 adults.

g. In a random sample of 100 adults, find $P(63 \leq X \leq 73)$, where X is defined in part **e**.

$$\begin{aligned} P(X \leq 73) - P(X \leq 62) &= \text{binompcdf}(100, .6827, 73) - \text{binompcdf}(100, .6827, 61) \\ &= .8703 - .1086 \\ &= .7617 \end{aligned}$$

8. The average height of trees in a forest is 33 feet with standard deviation 4 feet. Assume that the histogram of the heights of all trees is symmetric.
   a. Suppose that the z-core of the height of a tree is 1.5. What percentage of trees taller than this tree?

   $$P(Z \geq 1.5) = \text{normcdf}(1.5, 10^7, 0, 1) = .0668$$

   b. What percentage of trees with height no more than 33 feet?
   Since the histogram is symmetric, mean = median = 33. So, 50% of trees are taller than 33 feet.
   c. What percentage of trees with height 30 feet or taller but no more than 36 feet?
   Since the histogram is symmetric, we can assume that the height X of a tree is normal with mean 33 and standard deviation 4. So

   $$P(X \geq 30) = \text{normcdf}(30, 10^7, 33, 4) = .7734.$$

   d. Find the 99th percentile of the heights of all trees in the forest, and interpret its meanings.
   invNorm(.99,33,4) = 42.3; about 42 feet. That is, 99% of trees are 42 feet or shorter.
9. The average time for a particular LED TV needs first service is 415 days with standard deviation of 30 days. The manufacturer gives one year onsite service warranty for newly bought TVs. Assume that the times for first service call are approximately normally distributed.
   a. Let X denote the time until the first service for an LED TV. What is the distribution of X?
   X is normal with mean = 415 days and standard deviation = 30 days.
   b. What is the probability that a TV requires the first service within the warranty period?
   One year = 365 days. So we are looking for the probability

   $$P(X \leq 365) = \text{normcdf}(-10^7, 365, 415, 30) = .0478$$

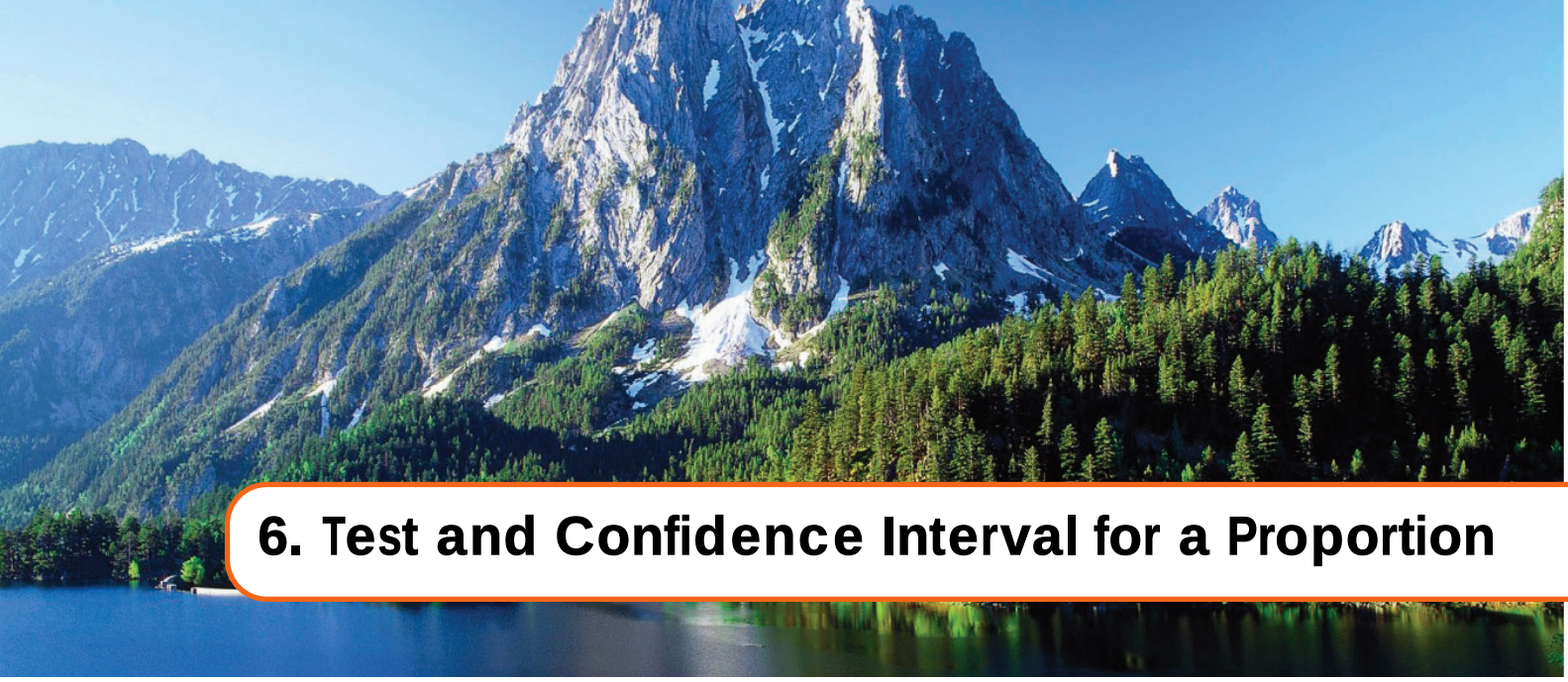
   c. Determine the warranty period so that the manufacturer is expected to service about 5% of the TVs within the warranty period.
   This means that we should take 5th percentile of the normal distribution as the warranty period. The 5th percentile is given by

   $$\text{invNorm}(.95, 415, 30) = 365.7$$

   That is, about one year. This means that if the manufacturer sets the warranty period as one year, then only 5% of the TVs some services within the warranty period.

   d. Suppose the warranty time is determined as in part **c**, and a batch of 5,000 TVs were sold. Find the probability that 300 or more TVs will require service within the warranty period.
   Let X denote the number of TVs that require services within warranty period out of the 5,000 TVs sold.
   Then X is binomial with $n = 5000$ and $p = .05$ which is the probability that a TV requires a service within one year; see part **c.** So

   $$P(X \geq 300) = 1 - P(X \leq 299) = 1 - \text{binomcdf}(5000, .05, 299) = 1 - .9991 = .0009.$$



6. Test and Confidence Interval for a Proportion

6.1 Basic Idea Behind Hypothesis Testing

We shall first describe some basic concepts to understand the method of statistical hypothesis testing.

Balanced or Unbalanced Coin: Suppose we like to test if a coin is balanced or not, on the basis of number of heads out of 10 flips of the coin. Suppose we observe 10 heads out of 10 flips. On the basis of this outcome, can we conclude the coin is unbalanced? Conventional wisdom says, yes. This is because, if the coin is really balanced then the probability of observing a head at each flip is 0.5, and the expected number of heads in a sequence of 10 flips is 5. So if we observe too many heads or too few heads, then we suspect the coin to be unbalanced. In the present case, the probability of the event of observing 10 heads out of 10 flips is

$$\frac{1}{2^{10}} = \frac{1}{1024} = .000977 < .001,$$

which indicates that the event is very unlikely to occur if the coin is balanced. Thus, the probability of observing head in a flip may be larger than that of observing tail. That is, we may conclude that the coin is unbalanced.

To find the numbers of heads that are considered to be extreme, we examine the probability distribution of

$X =$ the number of heads that can be observed out of 10 flips of a balanced coin.

Noticing that X is a binomial random variable with $n = 10$ and "success probability" $p = .5$. We

Table 6.1: Probability distribution of the number of heads X out 10 flips of a coin

x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	.00098	.00977	.044	.117	.205	.246	.205	.117	.044	.00977	.00098

see from the probability distribution in Table 6.1 that, if the coin is balanced, then probability of

observing 0, 1, 9 or 10 heads is

$$P(X=0)+P(X=1)+P(X=9)+P(X=10)=.0216.$$

Thus, if the coin is balanced, then the events of observing 0, 1, 9, or 10 heads are unlikely, (chances are less than 5%; this is a conventional value), and in case if any of these events occur we may reject the claim that the coin is balanced.

Errors: We also notice from the above approach that we may wrongly conclude that the coin is unbalanced. Even if the coin is actually balanced, the probability of observing 0, 1, 9, or 10 heads is .0216, and so 2.16% of the times our decision could be wrong. This means that if the "experiment of flipping a coin 10 times" repeated for, say, 10,000 times, we may wrongly conclude that the coin is unbalanced in about 216 times. This type of error is referred to as the **type I error**. A formal testing method is usually developed so that the maximum probability of type I error is no more than a pre specified value denoted by α, and the common values of α are .10, .05 or .01.

Some Applications of Hypothesis Tests

1. Is a new drug is more effective than an existing drug for treating a specific disease?
2. Is the percentage of defective items from a production process more than 3?
3. Does the life expectancy of U.S. woman greater than that of a man?
4. Is a drug more effective on women than on men?
5. Is average auto insurance premium for women is less than that of men?

In order to formally describe some popular testing methods, we shall see some basic terminologies in the following section.

6.2 Hypothesis Test

Hypothesis is a statement about the population parameter (such as population mean, variance, etc.)

Null Hypothesis (H_0) is a statement that the population is in a subset of the parameter space. The null hypothesis is usually denoted by H_0.

Alternative Hypothesis (H_a) is a statement which is usually in negation with the null hypothesis. The alternative hypothesis is also called the research hypothesis; in general, H_a depends on the objective of an experiment or a test. The alternative hypothesis is commonly denoted by H_a.

Test Statistic is a statistic based on the sample drawn from the population, and is used to test the hypotheses.

Errors: There are two types of errors one may commit while testing the hypotheses. These errors are referred to as the Type I and Type II errors, and they defined as follows.

Type I Error: (False Positive) Wrongly rejecting H_0 when it is actually true. For example, classifying a person as hepatitis B positive when she/he is a actually not.

Type II Error: (False Negative) Wrongly accepting H_0 when it is false. For example, classifying a person as normal when she/he is actually infected by the virus.

Power: The probability of rejecting the null hypothesis when it is false is referred to as the power.

Level of Significance: The pre specified maximum probability of making Type I error is called the level of significance. The level of significance is denoted by α, and it is usually .10, .05 or .01.

P-value: The p-value is a measure of evidence against the H_0. If the p-value is less than the level of significance α, then the H_0 is rejected. In general, smaller the p-value, stronger the evidence against H_0. The p-value is determined by the H_a, the value of the test statistic and the distribution of the test statistic under the assumption that H_0 is true.

Decision Rule: Reject the H_0 if **the p-value is less than** α. If the p-value is greater than α, we say, not enough evidence against H_0, or not enough evidence to support H_a.

	Null Hypothesis	
	TRUE	FALSE
REJECT	Type I Error	Correct Decision
DO NOT REJECT	Correct Decision	Type II Error

Statistical tests are usually developed under the following considerations:

1. The type I error rates should be very close to the level of significance, if not exactly equal to the level.
2. The power should increase with the increasing discrepancy between the parameter values under null hypothesis and under alternative hypothesis.
3. The power should be increasing with increasing sample size.

6.3 Test for a Proportion

There are several situations one wants to test if the population proportion p is equal to a specified value. Population proportion is defined as

$$p = \frac{\text{number of individuals in the population with a characteristic of interest}}{\text{total number of individuals in the population}}$$

Some Examples

1. The proportion of defective items in a shipment of items is defined as

$$p = \frac{\text{number defective items in the shipment}}{\text{total number of items in the shipment}}$$

2. The proportion of graduates from a high school who will go to college

$$p = \frac{\text{number of graduates from the school who are enrolled in colleges}}{\text{total number of graduates}}$$

3. The proportion of females in a college is defined as

$$p = \frac{\text{number of female students}}{\text{total number of students in the college}}.$$

Note that the population proportion p always lies between 0 and 1. That is,

$$0 \leq p \leq 1.$$

To test the proportion in a population, we draw a sample of n individuals from the population, and compute the sample proportion as

$$\widehat{p} = \frac{\text{number of individuals in the sample with a characteristic of interest}}{n}$$

■ **Example 6.1** Let p denote the proportion students in a college who are suffering from a flu. Suppose in a sample of 200 students, we found 64 students suffer from flu. Then, we estimate the p by

$$\widehat{p} = \frac{64}{200} = .32 \text{ or } 32\%$$

■

Test of Hypotheses on Proportion

Let p denote the population proportion, which is usually unknown. Let $\widehat{p}$ denote the sample proportion. Suppose the hypotheses are

$$H_0 : p \leq p_0 \quad \text{vs.} \quad H_a : p > p_0. \tag{6.1}$$

The test statistic is given by the z-score

$$z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\text{point estimate of the parameter} - \text{the value of the parameter in } H_0}{\text{SE of the point estimate}}.$$

Notice that the above z statistic is a random quantity, because $\widehat{p}$ is a random quantity. Let z_0 denote the numerical value of the test statistic based on a sample. The z_0 is referred to as the "observed value" of z.

Right-tailed Test

We reject the null hypothesis if z_0 is large (because $H_a : p > p_0$). For a given level α (the maximum type I error rate), the value of z_0 is considered to be large if

$$z_0 > z_{1-\alpha},$$

where $z_{1-\alpha}$ denotes the $100(1-\alpha)$ percentile of the standard normal distribution. For example, if $\alpha = 0.05$, then

$$z_{1-\alpha} = z_{1-.05} = z_{.95} = 1.645.$$

Alternatively, for a given level of significance α, we reject H_0 if the

$$\text{p} - \text{value} = P(z > z_0) < \alpha.$$

The above probability can be calculated using the function normcdf$(z_0, 10^7, 0, 1)$ in TI-84. We see from Figure 6.1 that whenever $z_0 > 1.645$, the p-value $P(z > z_0)$ should be less than 0.05.

Left-tailed Test

If the hypotheses are

$$H_0 : p \geq p_0 \quad \text{vs.} \quad H_a : p < p_0,$$

then we reject the null hypothesis for a small value of z_0. For a given level α, we reject H_0 if

$$z_0 < z_{.05} \quad \text{or} \quad \text{p} - \text{value} = P(z \leq z_0) < \alpha.$$

The p-value can be computed using normcdf$(-10^7, z_0, 0, 1)$. We see in Figure 6.2 that the p-value is less than 0.05 whenever z_0 falls in the rejection region.

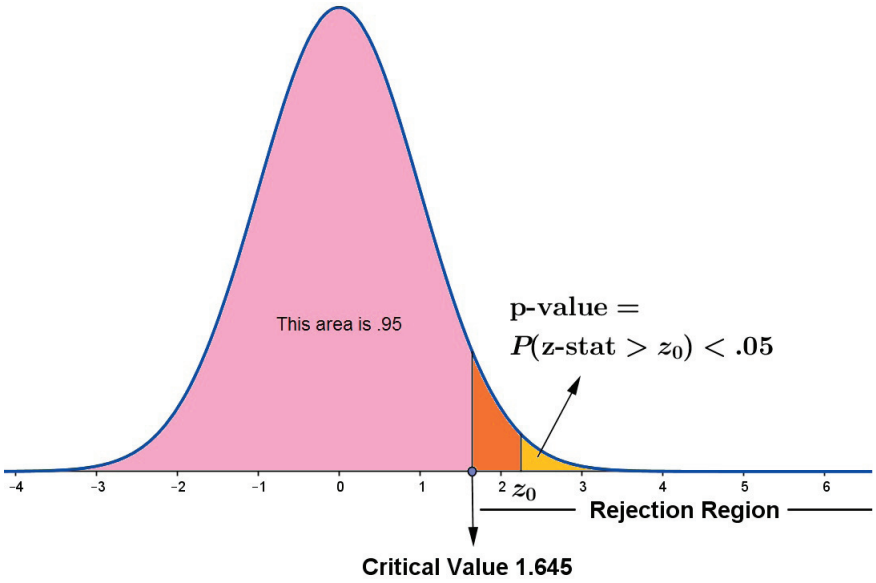


Figure 6.1: Illustration of critical point $z_{.95} = 1.645$ and the p-value for testing $H_0 : p \leq p_0$ vs. $H_a : p > p_0$ at the level of 0.05

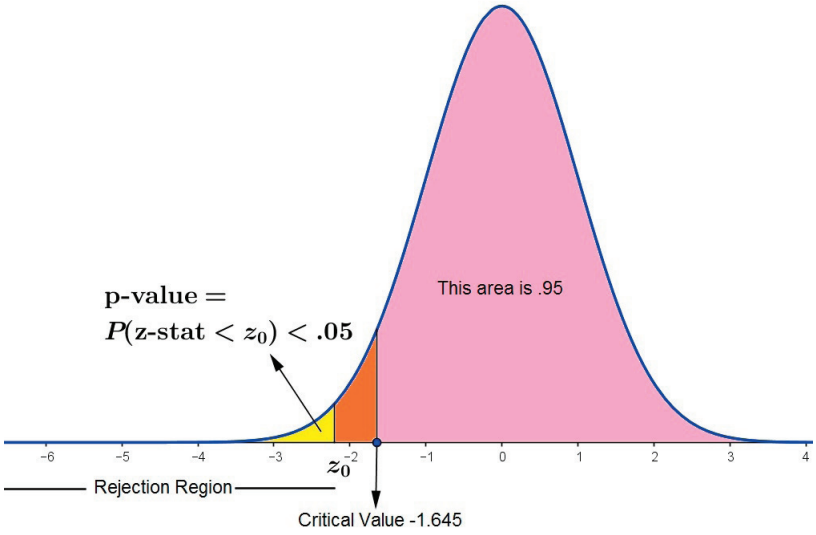


Figure 6.2: Illustration of critical point $z_{.05} = -1.645$ and the p-value for testing $H_0 : p \geq p_0$ vs. $H_a : p < p_0$ at the level of 0.05

Two-tailed Test

If the hypotheses are

$$H_0 : p = p_0 \quad \text{vs.} \quad H_a : p \neq p_0,$$

we reject the null hypothesis for a small or a large value of z_0. That is, reject H_0 if

$$z_0 > z_{1-\frac{\alpha}{2}} \text{ or } z_0 < -z_{1-\frac{\alpha}{2}}, \text{ or equivalently } |z_0| > z_{1-\frac{\alpha}{2}}.$$

For example, if we choose $\alpha = 0.05$, then $z_{1-\frac{\alpha}{2}} = z_{.975} = 1.96$.

Equivalently, we can reject H_0 if the p-value

$$P(z < -z_0) + P(z > z_0) < \alpha.$$

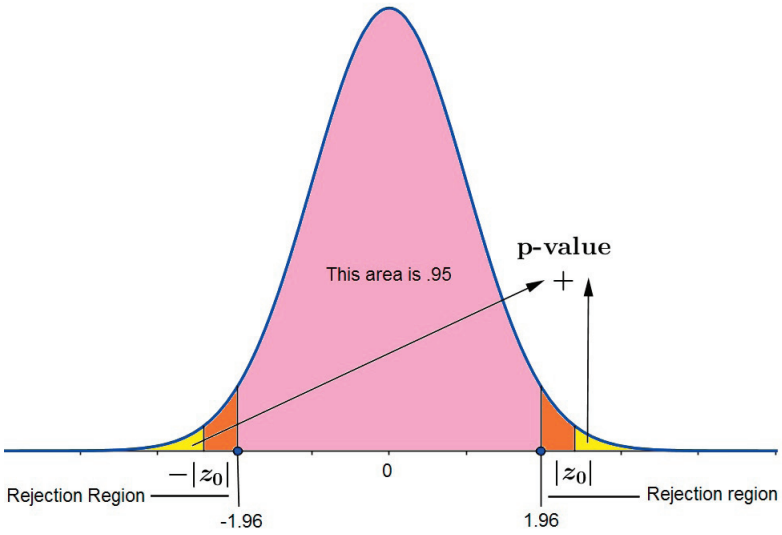


Figure 6.3: Illustration of critical points $z_{.025} = -1.96$, $z_{.975} = 1.96$ and the p-value for testing $H_0 : p \geq p_0$ vs. $H_a : p < p_0$ at the level of 0.05

■ **Example 6.2** A manufacturer of a machine part claims that only fewer than 3% of the parts could be defective. Inspection of a sample of 50 parts revealed that 4 parts were defective. Does this information provide sufficient evidence to conclude that the true percentage of defective parts is more than 3%? Test using the level of significance .05.

Solution: Let p denote the true proportion defective machine parts produced by the manufacturer.

$$H_0 : p \leq .03 \quad \text{vs.} \quad H_a : p > .03.$$

The sample proportion is $\widehat{p} = \frac{4}{50} = .08$

Test statistic:

$$z_0 = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.08 - .03}{\sqrt{\frac{.03 \times .97}{50}}} = 2.073 \text{ (observed value).}$$

The cut-off approach: As this is a right-tailed test, the cut-off value at the level .05 is $z_{95} = 1.645$. Since the observed value is grater than 1.645, we reject the null hypothesis.

The p-value approach: Alternatively, we can use the p-value approach as follows. The p-value, on the basis of the alternative hypothesis H_a, is the probability that the standard normal random variable Z greater than 2.073. That is,

$$P(z > 2.073) = \text{normcdf}(2.073, 10^7, 0, 1) = .019.$$

Since the p-value .019 is less than .05, we reject H_0. The data provide sufficient evidence to conclude that the true percentage of defective parts is more than 3 at the level .05. ∎

Cut-Off vs. P-value: If the decision of a test is based on comparing the value of the test statistic with the percentile (for example, $z_{.95}$), then the approach is referred to as the critical value approach; if it is based on the p-value, then the approach is referred to as the p-value approach.

Cut-Off

The cut-off value depends on the level of significance, and the alternative hypothesis of the test, and it should be computed for a given level of significance. Suppose we use the cut-off approach for Example 6.2. At the level of 0.05, the cut-off is

$$z_{.95} = 1.645.$$

Since the observed statistic is $2.073 > 1.645$, we reject the null hypothesis at the level 0.05. Can we reject the null hypothesis at the level, say, 0.025? The answer is yes, because the cut-off

$$z_{1-.025} = z_{.975} = 1.96,$$

and the observed statistic 2.073 is still greater than the cut-off. Notice that we need to calculate the cut-off for each level of significance and compare it to the observed statistic so as to reject/accept the null hypothesis.

P-value

The p-value does not depend on the given level of significance, and it depends only on the observed value and the alternative hypothesis.

For Example 6.2, the p-value is .019, and the null hypothesis is rejected at the level 0.05. We can also reject the null hypothesis at the level 0.025, because the p-value 0.019 is less than 0.025. As the calculation of p-value does not depend on the specified level of significance, the decision as to accept/reject the null hypothesis can be made for any level of significance. So we shall use only the p-value approach for all hypothesis testing problems in the sequel.

Why is the p-value computed at the boundary? In Example 6.2, why the p-value is computed at $p_0 = .03$, not at some values less than 0.03? For example, what is the p-value at .02? At $p_0 = .02$, the statistic

$$z_0 = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0 \times (1-p_0)}{n}}} = \frac{.08 - .02}{\sqrt{\frac{.02 \times .98}{50}}} = 3.0305$$

and the p-value is

$$P(z \geq 3.0305) = \text{normcdf}(3.0305, 10^7, 0, 1) = 0.0012,$$

which is less that the p-value of 0.019 at $p_0 = .03$. In fact, the p-value of the test at any value of p_0 less than .03 must be less than 0.019. In other words, the p-value of the test at the boundary 0.03 is the maximum. Thus, the null hypothesis $H_0 : p \leq .03$ is rejected at the level 0.05, if the p-value at the boundary $p_0 = .03$ is less than 0.05.

TI Calc

Calculating the test statistic, and the p-value using TI 84

To calculate the required quantities for Example 6.2, follow steps below.

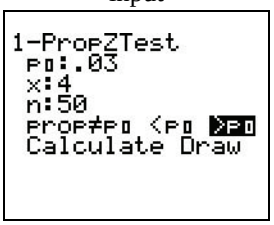
1. Select [STAT], [TESTS], [1-PropZTest]
2. enter $p_0 : .03$; $X : 4$; $n : 50$
3. Select $> p_0$ (this is based on the alternative hypothesis $H_a : p > p_0$)
4. Select [Calculate] and press [enter]

You get the following result:

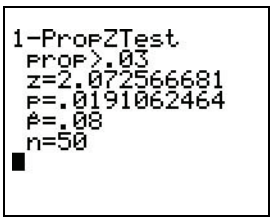
prop $> .03$ (your H_a); z = 2.073 (the value of the test statistic); p = .0191 (p-value);

$\hat{p} = .08$; n = 50

input



results



■ **Example 6.3** A pharmaceutical company claims that 75% of doctors prescribe one of its drugs to treat a particular disease. In a random sample of 40 doctors, 23 prescribed the drug to their patients. Does this information provide sufficient evidence to indicate that the actual percentage of doctors who prescribe the drug is less than 0.75?

a. Identify the parameter of interest, and the test you want to use.
b. Write the null and alternative hypotheses.
c. Write the value of the test statistic and p-value.
d. Write the conclusion.

Solution:

a. The parameter of interest is

$$p = \text{the true proportion of doctors who perscribe the drug.}$$

The one proportion z-test.

b. $H_0 : p \geq .75$ vs. $H_a : p < .75$.

c. Note p0: .75, X: 23, n: 40 and select $<$. Using these numbers in TI-84 ([STAT], [TESTS], [1-PropZTest]), we get the test statistic -2.556 and the p-value = .0053.

d. Since the p-value is less than .05, we can conclude, on the contrary to the manufacturer's claim, that less than 75% of doctors prescribe the drug.

■

■ **Example 6.4** According to the CDC Distracted Driving Study, in year 2011 about 17% of crashes in which someone was injured involved distracted driving (texting or attending phone calls while driving). An auto insurance agent believes that this percentage may be decreased now because of campaigns and TV ads by insurance companies. The agent investigated a sample of 120 crashes in which someone was injured, and found that 16 involved distracted driving. Does this information provide sufficient evidence to indicate that the percentage of crashes in which someone was injured involved distracted driving has decreased since 2011?

a. Identify the parameter of interest, and the test you want to use.

b. Write the null and alternative hypotheses.
c. Write the value of the test statistic and p-value.
d. Write the conclusion.

Solution:

a. The parameter of interest is

p = the true proportion of crahses in which someone was injured involved distracted driving at present.

The appropriate test is one proportion z-test.

b. $H_0 : p \geq .17$ vs. $H_a : p < .17$.

c. Note p0: .17, X: 16, n: 120 and select "<" Using these numbers in TI-84 ([STAT], [TESTS], [1-PropZTest]), we get the test statistic -1.0693 and the p-value = .1425.

d. Since the p-value is not less than .05, there is not sufficient evidence to indicate that the percentage has decreased since 2011.

■

6.4 Confidence Intervals for a Proportion

Confidence interval (CI) for a parameter is an interval estimate based on a sample from the population. The CI is constructed so that it would include the parameter with confidence $1-\alpha$, where $0 < \alpha < .5$. In most practical situations, $\alpha = .05$, and the confidence coefficient is $1-\alpha = .95$. Confidence interval with $1-\alpha = .95$ is referred to as the 95% confidence interval.

For large samples,

$$\frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \text{ is approximately normal with mean 0 and std deviation 1.}$$

As a result, we have

$$-z_{.975} \leq \frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} \leq z_{.975} \text{ with probability .95.}$$

Solving the above inequality for p, we find 95% CI for p as

$$\widehat{p} - z_{.975}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \;\leq\; p \;\leq\; \widehat{p} + z_{.975}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Note that $z_{.975}$ = invNorm(.975) = 1.96. The term

$$z_{.975}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \text{ is referred to as the Margin of Error (ME).}$$

Thus, the 95% CI for the proportion p is expressed as

$$\widehat{p} \pm z_{.975}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \widehat{p} \pm ME.$$

■ **Example 6.5** Nielsen company surveyed 225 owners of Android phones, and found that 160 of them planned to get another Android as their next phone.

a. Find a point estimate for the proportion of Android users who plan to get another Android.
b. Construct a 95% CI for the true proportion of Android users who plan to get another Android.

c. Interpret the meanings of the CI in part **b.**

d. Suppose an advertisement claimed that 70% of Android users plan to get another Android. Does the CI contradicts the claim?

Solution:

a. The point estimate is

$$\widehat{p} = \frac{160}{225} = .7111.$$

b. The margin of error is

$$z_{.975} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 1.96 \times \sqrt{\frac{.7111(1-.7111)}{225}} = 1.96 \times .0302 = .0592$$

So the 95% CI is $.7111 \pm .0592 = (.652, .770)$.

c. The true percentage of Android users who plan to get another Android is somewhere between 65 and 77 with confidence 95%.

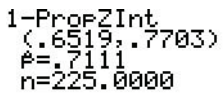
d. Since the above CI includes 0.70, the CI does not contradict the claim.

■

TI Calc

Construction of CI for a proportion using TI-84

Press [STAT], select TESTS and then [1-Prop ZInt...]
enter 160 for X, 225 for n, .95 for [C-Level], and select [Calculate]



The margin of error can be found as

$$\frac{.7703 - .6519}{2} = .0592.$$

■ **Example 6.6** A random sample of 80 automobile drivers from a city showed that only 60 of them fastening the seat belts while driving. Based on this information, it is desired to estimate the true percentage of drivers who are fastening seat belts while driving.

a. What interval estimating method do you use for this problem?

b. Find a 95% CI for the true proportion of drivers who are fastening seat belts while driving, and interpret its meanings.

c. Find a 90% CI for the true proportion of drivers who are fastening seat belts while driving.

d. Find a 95% CI for the true proportion of drivers who are not fastening seat belts while driving.

■

Solution:

a. One-Proportion Z interval.

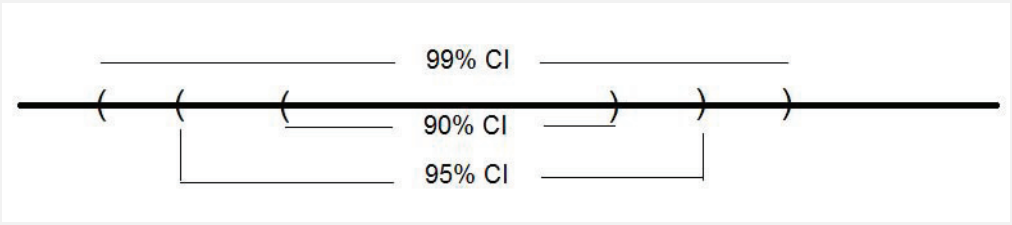
b. X = 60, n = 80, C-Level = .95; the 95% CI is (.655, .845). The true percentage of drivers who are fastening seat belts is between 65.5 and 84.5 with confidence 95%.

c. C-Level = .90; the 90% CI is (.670, .830).

d. This CI can be found (by entering X = 20, n = 80 and C-Level = .95) as (.155, .345). This CI can also be obtained from part **b** as

$$(1-.845, 1-.655) = (.155, .345).$$

Width of CI and Confidence Level: Note that in the preceding example, the 90% CI (.670, .830) is narrower than the 95% CI (.655, .845). In general, the width of the CI increases as the confidence level increases. Recall that the margin of error (ME) is one half of the width of a CI. So the margin of error increases with increasing confidence level.



As an example, suppose a 95% CI for a parameter based on a sample of size 25 has margin of error 0.07. Then the 90% CI based on the same sample has a margin of error less than 0.07 while the 99% CI based on the same sample has a margin of error larger than 0.07.

■ **Example 6.7** A new drug for treating a particular disease is to be tested for its effectiveness. A sample of 120 adults were asked to use the drug over a period of five days, and report the drug effects. Of the 120 patients, 92 patients told they were completely relieved from the disease. On the basis of this study, it is desired to find an interval estimate for the true success rate of the drug.

a. What interval estimation method is appropriate for this problem?
One-proportion Z interval

b. What is the point estimate of the true success rate?

$$\frac{92}{120} = .7667$$

c. Find a 95% CI for the success rate, and interpret its meanings.
Note that $X = 92$, $n = 120$ and the confidence level .95. Entering these numbers in [1-PropZint..], we find

$$(.691, .842)$$

That is, the true percentage of patients who were completely relieved from the disease is between 69.1% and 84.2% with confidence 95%. We can also say "the success rate of the drug is between 69.1% and 84.2% with confidence 95%"

d. What is the margin of error of the CI in part **b**?
The ME is $\frac{.842-.691}{2} = .0755$.

e. On the basis of the CI in part **b**, can we conclude that the true success rate is at least 69%? Explain.
Yes, because the left endpoint is greater than .69.

■

■ **Example 6.8** The purchasing department in a manufacturing firm has to decide as to accept/reject a large shipment of items. The manager of the purchasing department has decided to use the following strategy: He will inspect a random sample of 300 items, and if he finds evidence to indicate

the proportion of defective items in the shipment is no more than 4%, then he will buy the shipment. Suppose he finds 6 defective items in the sample.

a. Find a 95% CI for the true proportion of defective items in the shipment.
Here, $n = 300$ and $X = 6$. Using [1-PropZint...], we get

$$(.004, .036)$$

That is, the percentage of defective in the shipment is between .4% and 3.6%, with confidence 95%.

b. On the basis of the above CI, what could be the manager's decision? Explain.
Yes, because the maximum percentage is 3.6%, which is less than 4%.

c. What is the error rate of the statistical method that is being used to make the decision?
Since the decision is based on the 95% CI, there are 5% chances that the CI will not include the true percentage of defective. That is error rate is

$$100 - 95 = 5\%$$

■

7. Comparison of Two Proportions

We shall now see some methods of comparing two proportions. Let p_1 denote the proportion of individual with an attribute of interest in Population 1, and p_2 denote the same in Population 2. For example,

1. Let p_1 = the proportion of patients who got cured by taking treatment A, and p_2 = the proportion of patients with the same disease who got cured by taking treatment B. It is of interest to compare p_1 and p_2 to find the better treatment.
2. A researcher in occupational/environmet medicine believes that one of the causes of a particular disease is long-term exposure to a chemical. To test his belief, he examined a sample of adults and obtained the results in the from of following 2×2 table:

Group	Symptoms Present	Symptoms Absent	Totals
Exposed	X_1	$n_1 - X_1$	n_1
Unexposed	X_2	$n_2 - X_2$	n_2

Let p_e denote the proportion of people in the exposed group who got the symptoms, and p_u denote the same in the unexposed group. The problem is to compare p_e and p_u so as to find if there is an adverse effect of exposure.

7.1 Test for the Difference Between Two Proportions

Let X_1 denote the number of individuals with an attribute of interest in a sample n_1 individuals from the population 1, and let X_2 denote the number of individuals with an attribute of interest in a sample of n_2 individuals from the population 2. The data can be arranged as follows.

	Population 1	Population 2
sample size	n_1	n_2
no. of units in the sample with the attribute	X_1	X_2
sample proportion	$\widehat{p}_1 = \frac{X_1}{n_1}$	$\widehat{p}_2 = \frac{X_2}{n_2}$

On the basis of sample proportions, we need to test if there is a significant difference between the population proportions p_1 and p_2.

A Test Statistic for comparing two proportions is given by

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \text{ with } \widehat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$

For large samples, the above test statistic is distributed like z-scores or normally distributed with mean zero and variance one.

Right-Tailed Test

For given data X_1, n_1, X_2 and n_2, let z_0 denote the value of the above z-statistic. For testing

$$H_0 : p_1 \leq p_2 \quad \text{vs.} \quad H_a : p_1 > p_2,$$

the null hypothesis will be rejected if

$$z_0 > z_{1-\alpha} \quad \text{or} \quad \text{p-value} = P(z > z_0) < \alpha,$$

where α is the level of significance.

Left-Tailed Test

For testing

$$H_0 : p_1 \geq p_2 \quad \text{vs.} \quad H_a : p_1 < p_2,$$

the null hypothesis will be rejected the null hypothesis will be rejected if

$$z_0 < -z_{1-\alpha} \quad \text{or} \quad \text{p-value} = P(z < z_0) < \alpha,$$

where α is the level of significance.

Two-Tailed Test

For testing

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_a : p_1 \neq p_2,$$

the null hypothesis will be rejected if the p-value

$$z_0 > z_{1-\alpha/2} \quad \text{or} \quad z_0 < -z_{1-\alpha/2} \quad \text{or} \quad P(z < -|z_0|) + P(z > |z_0|) < \alpha,$$

where z_0 is an observed value of the statistic z, and α is the level of significance.

■ **Example 7.1** A physician believes that one of the causes of a particular disease is long-term exposure to a chemical. To test his belief, he examined a sample of 32 adults from the exposed group and found 13 adults with some symptoms whereas only 4 with the same symptoms in a sample of 25 unexposed adults. On the basis of these data, it is desired to test if there is a positive association between the prolonged exposure to the chemical and the disease. Test using $\alpha = 0.05$.

a. What test can be used?
b. State the null and alternative hypotheses.
c. Compute the value of the test statistic and p-value.
d. Write the conclusion of the test.

■

Solution:

a. Two-Proportion Z test or two-sample test for proportions.

b. Let p_e denote the proportion of adults with the symptoms in the exposed group, and p_u denote the same for the unexposed group.

$$H_0 : p_e \le p_u \quad \text{vs.} \quad H_a : p_e > p_u.$$

c. Here $X_1 = 13$, $n_1 = 32$, $X_2 = 4$ and $n_2 = 25$. The estimates are

$$\widehat{p}_e = \frac{13}{32} = .4063, \quad \widehat{p}_u = \frac{4}{25} = .1600 \quad \text{and} \quad \widehat{p} = \frac{13+4}{32+25} = .2982.$$

The test statistic is

$$\begin{aligned} z_0 &= \frac{\widehat{p}_e - \widehat{p}_u}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \\ &= \frac{.4063 - .1600}{\sqrt{.2982(1-.2982)\left(\frac{1}{32}+\frac{1}{25}\right)}} \\ &= 2.0170. \end{aligned}$$

We compute the p-value as $P(Z > z_0) = P(Z > 2.0170) = \text{normcdf}(2.0170, 10^7, 0, 1) = .0218$.

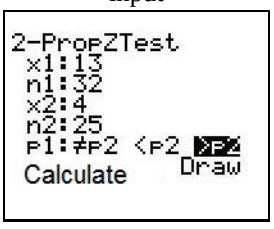
d. Since the p-value .0218 is less than .05, we reject H_0, and conclude that the prolonged exposure to chemicals has some positive effects on the disease.

TI Calc

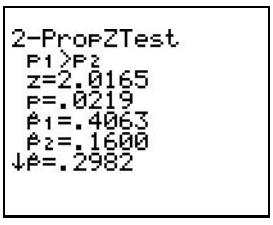
Two-Proportion Z test using TI-84:

Select [STAT], [TESTS] and [6: 2-Prop Z Test...]
enter the numbers as shown below.

input



result



Notice that the statistic we calculated above is 2.0170, and the one using TI-84 is 2.0165. This discrepancy leads to slightly different p-values, though practically the same.

■ **Example 7.2** Can a diagnostic test be recommended for use? To assess the effectiveness of a diagnostic test for detecting a certain disease, a medical researcher reports that 36 out of 40

diseased persons were correctly diagnosed by the test and 16 out of 80 non-diseased persons were incorrectly diagnosed. It is desired to test if the true positive diagnoses is higher than the false positive diagnoses.

a. What test can be used?
b. State the null and alternative hypotheses.
c. Compute the value of the test statistic and p-value.
d. Write the conclusion of the test.

Solution:

a. Two-Proportion Z test or two-sample test for proportions.
b. Let p_t and p_f denote respectively the true positive diagnoses and false positive diagnoses. Then, the hypotheses of interest are

$$H_0 : p_t \leq p_f \quad \text{vs.} \quad H_a : p_t > p_f .$$

c. $X_1 = 36,\ n_1 = 40,\ X_2 = 16$ and $n_2 = 80$. Using TI-84, we get the test statistic as

$$z_0 = 7.2947 \ \text{ and } \ \text{p} - \text{value} = 1.5079 \times 10^{-13} = 0$$

.
d. Since the p-value is zero, the test clearly indicates that the true positive diagnoses is greater than the false positive diagnoses.

■

■ **Example 7.3** Can meditation lower the risk of heart related problems such as heart attack and stroke? The most encouraging result was from one of the longest studies, conducted at Medical College of Wisconsin, in Milwaukee[1]. A group of 201 high-risk heart patients, male and female, were divided into two groups. One group meditated twice a day for 20 minutes; the control group relaxed but did not meditate. The group was followed for five years and the following results were obtained. In the above table, an event means the patient had suffered a heart attack, stroke or

Table 7.1: Meditation and heart related problems

	Meditation Group	Relaxed Group
Sample Size	$n_m = 100$	$n_r = 101$
No. of Events	$X_m = 20$	$X_r = 32$

died. In addition, the meditators tended to remain disease-free longer and to reduce their systolic blood pressure by an average of 5 mm (millimeters of mercury). On the basis of this result, can we conclude that mediation really lower blood pressure thereby reducing the risk of heart related problems?

a. Identify the parameters of interest.
b. What test do you use for the above problem?
c. State the null and alternative hypotheses.
d. Compute the value of the test statistic and p-value.
e. Write the conclusion of the test.

[1] http://www.secretstomeditation.com/highbloodpressure

Solution:

a. The parameters are

$$p_m = \text{true proportion of events among all heart patients who meditate regularly}$$

and

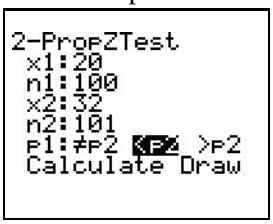
$$p_r = \text{true proportion of events among all heart patients who do not meditate}$$

b. Two-sample test for proportions.

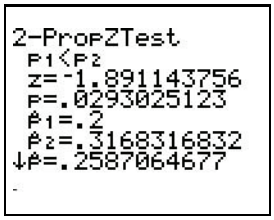
c. $H_0 : p_m \geq p_r$ vs. $H_a : p_m < p_r$.

d. Select [STAT], [TESTS] and [6: 2-PropZTest] in TI-84, and enter the data as shown below.

input



results



The z-statistic is 1.8911 and the p-value is 0.0293.

e. Since the p-value is less than 0.05, the data provide sufficient evidence to indicate that the meditation helps reducing the risk associated with high blood pressure.

■

■ **Example 7.4** Does overweight depend on the gender? The National Health and Nutrition Examination Survey reported the following statistics: In a sample of 546 boys in the age group 6-11, 87 of them are overweight, and in a sample of 508 girls in the same age group 74 of them overweight. Test if the proportion of boys who are overweight differs from that of girls.

a. Identify an appropriate test.

b. State the null and alternative hypotheses.

c. Compute the value of the test statistic and p-value.

d. Write the conclusion of the test.

Solution:

a. Two-Proportion Z test.

b.

$$H_0 : p_B = p_G \quad \text{vs.} \quad H_a : p_B \neq p_G,$$

where p_B denotes the proportion of overweight boys aged 6-11 in the entire population of all boys in that age group, and p_G denotes the same in girls aged 6-11.

c. Here, $X_1 = 87$, $n_1 = 546$, $X_2 = 74$ and $n_2 = 508$. Using TI-84 [STAT $\rightarrow$ TESTS $\rightarrow$ 2-PROP Z Test], we compute

$$z_0 = .6165 \text{ with p-value } .5376.$$

d. Since the p-value is larger than .05, we can't reject H_0. The test implies that percentage of overweight boys is not significantly different from that of girls.

■

Statistical Significance vs. Practical Significance: If the p-value of a test is very small, say, less than .05, we can say the effect (such as the difference between the hypothesized value and the true value) is statistically significant, but the effect may not be practically significant as will be seen in the following example.

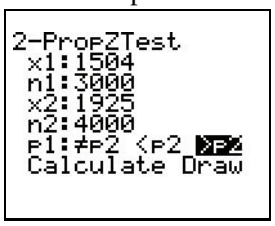
■ **Example 7.5** A drug was developed to cure a specific disease. To judge the effect of the drug, a sample of 3000 adults (treatment group) with the disease was treated with the drug for a period of 12 days, and a sample of 4000 adults (control group) with the same disease was given placebo (inert tablet, sugar pills) for the same period of 12 days. After the treatment, it was found that 1504 patients in the treatment group got cured, and 1925 patients in the control group got relieved from the disease. Do the data provide enough evidence to indicate that the drug is more than placebo in treating the disease? Test at the level 0.05.

Solution: Let us summarize the information as follows. The statistics and the p-value are

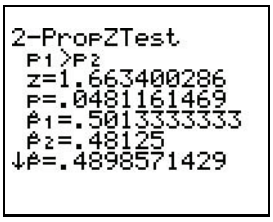
	treatment	control
Sample Sizes	$n_1 = 3000$	$n_2 = 4000$
No. of "Successes"	$X_1 = 1504$	$X_2 = 1925$

calculated using TI-84 as shown below.

input



result



The sample success rates are

0.5013 for the treatment group, and 0.4813 for the control group.

The difference

$$0.5013 - 0.4813 = .02$$

is statistically significant because the p-value of .0481 is smaller than 0.05. Notice that in the control group (those who took placebo treatment) the success rate was about 48% while in the treatment group it was 50%, only 2% higher than that of the control group. The increase by 2% in success rate may not worth, considering the treatment cost and time. So the improvement is statistically significant, but may not be practically significant. ■

■ **Example 7.6** A coronary prevention study was conducted to determine whether the administration of pravastatin to middle-aged men with high cholesterol levels over a period of five years reduces the risk of coronary events. In this context a coronary event is defined as a nonfatal heart attack or death from coronary heart disease. A group of 6595 men, aged 45 to 64 years, with high plasma cholesterol levels was randomly divided into two groups (a control group and a treatment group). The 3302 men in the treatment group received 40 mg of pravastatin daily while the 3293 men in the control group received a placebo. By the end of this five year trial, 174 of the 3302 men treated with pravastatin had experienced a cardiac event and 248 of the 3293 men treated with a placebo had experienced a cardiac event. Does this information provide sufficient evidence to indicate that pravastatin reduce the risk of coronary events? Test using $\alpha = .05$.

a. Identify the parameters of interest of the hypothesis test.

b. State the null and alternative hypotheses in terms of the parameters defined in part **a.**

c. Compute the appropriate test statistic and the p-value.

d. Write the conclusion of the test based on the p-value.

e. What is the maximum probability of rejecting the null hypothesis when it is actually true?

f. Can we conclude the same as in part **d** at the level of significance $\alpha = .01$? Explain.

■

7.2 Confidence Intervals for the Difference Between Two Proportions

Let $\widehat{p}_1$ denote the sample estimate of the population proportion p_1 based on a sample of size n_1, and let $\widehat{p}_2$ denote the sample estimate of the population proportion p_2 based on a sample of size n_2. For large samples, the quantity

$$\frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}} \text{ distributed like z-scores.}$$

As a result,

$$-z_{1-\frac{\alpha}{2}} \le \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}} \le z_{1-\frac{\alpha}{2}} \text{ with probability } 1-\alpha.$$

Solving the above inequality for $p_1 - p_2$, we obtain the following CI for $p_1 - p_2$:

Result 7.1 — A $100(1-\alpha)\%$ CI for the difference $p_1 - p_2$ is given by

$$\left(\widehat{p}_1 - \widehat{p}_2 - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}},\ \widehat{p}_1 - \widehat{p}_2 + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \right),$$

or equivalently,

$$\widehat{p}_1 - \widehat{p}_2 \pm ME, \quad \text{with } ME = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

Remark

Note that for a 95% CI, $1-\alpha = .95$, so $\alpha = .05$ and $\frac{\alpha}{2} = .025$. The critical value

$$z_{1-\frac{\alpha}{2}} = z_{1-.025} = z_{.975} = \text{invNorm}(.975) = 1.96.$$

For 90% CI, $1-\alpha = .9$ and so $\alpha = .1$ and $\frac{\alpha}{2} = .05$. The critical value

$$z_{1-\frac{\alpha}{2}} = z_{.95} = \text{invNorm}(.95, 0, 1) = 1.645.$$

■ **Example 7.7** In Example 7.2, we have $X_1 = 36,\ n_1 = 40$ for group of people with disease, and $X_2 = 16,\ n_2 = 80$ for the group of people without disease. We shall find a 95% CI for $p_d - p_f$, where

p_t = proportion of true positive diagnoses, and p_f = proportion of false positive diagnoses. Note that

$$\widehat{p}_1 = \frac{X_1}{n_1} = \frac{36}{40} = 0.90, \quad \widehat{p}_2 = \frac{X_2}{n_2} = \frac{16}{80} = 0.2, \quad n_1 = 40 \quad \text{and} \quad n_2 = 80.$$

To find a 95% CI, the z cut-off is $z_{.975} = 1.96$, and

$$ME = 1.96\sqrt{\frac{.9(1-.9)}{40} + \frac{.2(1-.2)}{80}} = 0.1278.$$

Therefore, the 95% CI for $p_t - p_f$ is given by

$$\widehat{p}_1 - \widehat{p}_2 \pm ME = .7 \pm .1278 = (.5722, .8278).$$

As the endpoints of the interval are positive, we can conclude that $p_t > p_f$. Furthermore, we can say that the true positive diagnoses is 57 to 83 percent higher than the false positive diagnosis. ■

Remark

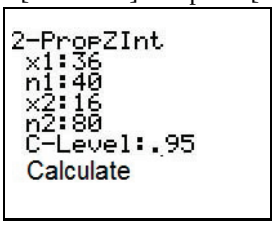
Comparison of p_1 and p_2 based on the CI for $p_1 - p_2$

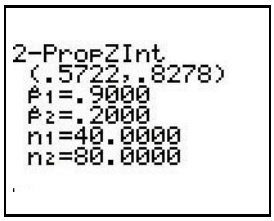
- The CI of the form $(+,+)$ indicates that $p_1 > p_2$.
- The CI of the form $(-,-)$ indicates $p_1 < p_2$.
- The CI of the form $(-,+)$ indicates that p_1 is not significantly different from p_2

TI Calc

Calculation of CI for $p_1 - p_2$ Using TI-84

Select [STAT], [TESTS] and [B: 2-Prop Zint..]
enter 36 for X_1, 40 for n_1, 16 for X_2, 80 for n_2 and .95 for [C-Level]
move the cursor on [Calculate] and press [enter]





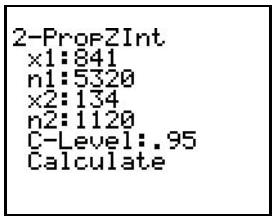
■ **Example 7.8** Angioplasty is a medical procedure to widen a blood vessel for smooth blood flow. A wire mesh tube, called stent, is placed in vessel to keep remain open. A study was conducted to compare the effectiveness of a *bare metal stent* with one that has been **coated with a drug** designed to prevent reblocking of the vessel. A total of 5,320 patients received bare metal stents, and of these, 841 needed treatment for reblocking within a year. A total of 1,120 received drug coated stents, and 134 of them required treatments for reblocking within a year. Find a 95% CI to check if the proportion of patients who received drug-coated stents and needed retreatment is less than that of those who received bare metal stent.

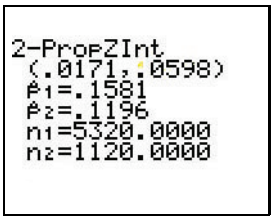
a. What interval estimation method do you use? [2 pts]
b. Find a 95% CI for the difference between the proportions. [3 pts]
c. What is the ME of the CI in part **b**? [2 pts]
d. Compare the two proportions on the basis of the CI in part **b.** [3 pts]

Solution:

a. 2-Proportion Z interval

b. Here, $X_1 = 841$, $n_1 = 5320$, $X_2 = 134$ and $n_2 = 1120$. Using TI-84, we find





Thus, the 95% CI for $p_1 - p_2$ is (.0171, .0598), where $p_1 =$ the proportion of patients who received bare metal stent, and needed treatments for reblocking, and p_2 is the same for the patients who received drug-coated stents.

c. Yes, because the endpoints are positive which implies that $p_1 > p_2$.

■

■ **Example 7.9** Using the data in Example 7.3, find a 90% CI for the difference $p_t - p_c$, where p_t is the true proportion of events in treatment group, and p_c is the same in the control group. Interpret the meanings of the CI.

Solution: For this example,

$$X_t = 20,\ n_t = 100,\ X_c = 32 \quad \text{and } n_c = 101.$$

TI Calc

- Select [STAT], [TESTS] and [B: 2-Prop Zint..]
- Enter 20 for X_1, 100 for n_1, 32 for X_2, 101 for n_2 and .90 for [C-Level]
- Move the cursor on [Calculate] and press [enter]. The results:

$$\widehat{p}_1 = 0.2, \quad \widehat{p}_2 = 0.3168, \text{and the CI is } (-.2175, -.0162)$$

Since both endpoints of the CI are negative, we can conclude that $p_t < p_c$ with 90% confidence. The adverse events in the meditating group is 1.6 to 22% lower than that of control group. ■

■ **Example 7.10** In a random sample of 485 females 335 are in favor of continuation of an environmental amendment in a state constitution. In the same state, 400 out of a sample of 500 males are in favor of continuation of the constitution.

a. Construct a 95% confidence interval for $p_f - p_m$, the difference in the proportions of females and males in the state who favor continuation of the amendment.

b. Does your CI in part **a** indicate that the proportion of females and males in the state who are in favor of continuation are different?

c. Will a 90% CI for $p_f - p_m$ indicate the proportions are different? Explain.

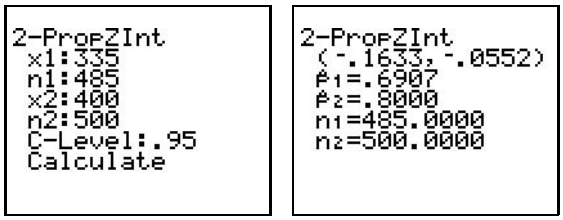
■

Solution:

a. Here $X_1 = 335,\ n_1 = 485,\ X_2 = 400$ and $n_2 = 500$. Select [STAT], [TESTS] and [B: 2-Prop Z Test]; enter these numbers and .95 for [C-Level] to get The 95% CI for $p_f - p_m$ is $(-.1633, -.0552)$.

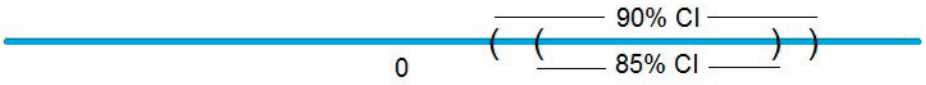
b. Yes, because the above CI does not include zero.

c. Since 90% CI should be included in $(-.1633, -.0552)$, it should not include zero. So 90% CI should also indicate that these two proportions are different.



■ **Example 7.11** Suppose that a 90% CI for the difference $p_1 - p_2$ indicates that $p_1 > p_2$. Does a 85% CI based on the same data indicate $p_1 > p_2$? What about a 95% CI?

Solution: Since the 90% CI shows that $p_1 > p_2$, the endpoints of CI should be positive. We also know that 85% CI should be included in the 90% CI, so the endpoints of this CI should also be positive, indicating $p_1 > p_2$. The 99% CI, however, is wider than the 90% CI, and so both of its endpoints are not necessarily positive. As a result, a 99% CI based on the same data may not indicate that $p_1 > p_2$.



■

■ **Example 7.12** The question of therapeutic superiority of streptokinase (SK) over tissue-type plasminogen activator (t-PA) for myocardial infarction (heart attack) is of considerable public health importance, since t-PA is approximately 10 times more expensive than SK. The Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) randomized clinical trial was carried out to answer this question, and part of the results are summarized below. It is desired to estimate the difference between the death rate of patients

Agent	Number of Patients	Number of Deaths
SK	20163	1453
t-PA	10363	654

who are taking SK and the death rate of patients who are taking t-PA.

a. Identify the parameters of interest.
p_S = proportion of deaths among all patients under SK treatement
p_t = proportion of deaths among all patients under t-PA treatment

b. Find a 95% CI for the difference between the death rates.
SK group: $X_1 = 1453$ and $n_1 = 20163$
t-PA group: $X_2 = 654$ and $n_2 = 10363$
Using TI-84: The 95% CI for $p_s - p_t$ is $(.003, .015)$

c. On the basis of the above CI, can we conclude that SK is superior to t-PA? Explain.
Since the CI is of the form $(+,+)$, we can say that the death rates among the patients who are treated with SK is .3 to 1.5 % more than that in t-PA treatment group.

d. Can we conclude the same as in part **c** on the basis of a 90% CI for the difference of death rates? Explain.
Yes, because the 90% CI should be included in the 95% CI, and it should also be of the form $(+,+)$

■

■ **Example 7.13** The age at which a woman gives birth to her first child may be an important factor in the risk of later developing breast cancer. An international study conducted by the World Health Organization (WHO) selected women with at least one birth and recorded if they had breast cancer or not and whether they had their first child before their 30th birthday or after. The results are summarized as follows. It is desired to estimate the difference between the cancer rates of

Age	Number of women	Cancer
at first birth > 30	3220	683
at first birth ≤ 30	10245	1498

these two groups of women.

a. Identify the parameters of interest.
$p_1 =$ proportion of cancer among women who gave birth later than 30 yrs of age
$p_2 =$ proportion of cancer among women who gave birth at the age of 30 or earlier

b. Find a 95% CI for the difference between the proportions of cancer in these two groups.
Here $n_1 = 3220$, $X_1 = 683$, $n_2 = 10245$ and $X_2 = 1498$
The 95% CI for $p_1 - p_2$ is $(.050, .082)$

c. What is the margin of error of your estimate in part **b**?
The ME is $\frac{.082-.050}{2} = .016$

d. Interpret the meanings of the CI in part **b**.
The cancer rate among women who gave birth later than 30 is 5 to 8.2% higher than that those who gave birth at the age of 30 or earlier.

■

8. Test and Confidence Interval for a Mean

Let $X_1, \ldots, X_n$ be a sample from a normal population with mean μ and standard deviation of σ. The sample mean $\bar{X}$ and the sample standard deviation S are calculated as

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

The population mean μ and the standard deviation σ are usually unknown, and it is desired to test if the population mean is a specified value. The test statistic is referred to as the t-statistic, and is given by

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, \tag{8.1}$$

where μ_0 is the specified value under the null hypothesis.

8.1 Hypothesis Test for a Population Mean

Right-Tailed Test

Suppose our hypotheses of interest are

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0,$$

where μ_0 is a specified value of the population mean. The statistic t defined in (8.1) is known to follow a distribution called

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

The numerical value of the t statistic based on an observed sample is denoted by t_0. This t_0 is also referred to as the observed value of the t statistic. The p-value for testing above hypothesis is given

by

$$P(t_{n-1} > t_0).$$

The null hypothesis will be rejected if the p-value is less than the level of significance α.

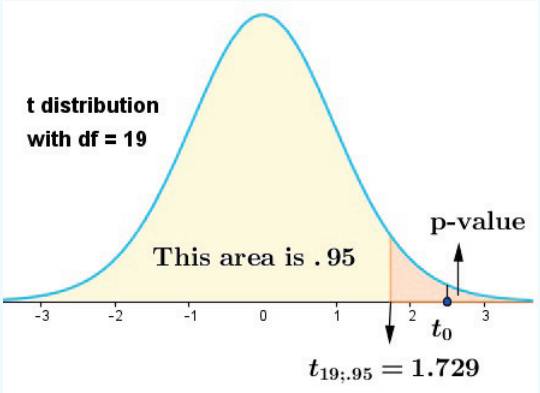


Figure 8.1: Illustration of critical value $t_{19;.95} = 1.729$ and the p-value

As an example, when $n = 20$, the 95th percentile of the t distribution with $n-1 = 19$ degrees of freedom is illustrated in Figure 8.1. If the hypotheses are

$$H_0 : \mu \le \mu_0 \quad \text{vs.} \quad H_a : \mu > \mu_0,$$

then the null hypothesis is rejected at the level of 0.05 if the test statistic

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > 1.729.$$

or the p-value $P(t_{19} > t_0) < 0.05$, where t_0 is an observed value of the test statistic.

Left-Tailed Test

For testing hypotheses

$$H_0 : \mu \ge \mu_0 \quad \text{vs.} \quad H_a : \mu < \mu_0,$$

the null hypothesis will be rejected if the p-value

$$P(t_{n-1} < t_0) < \alpha,$$

where t_0 is an observed value of the test statistic. The p-value can be calculated using TI-84 as

$$\begin{aligned} P(t_{n-1} < t_0) &= \text{tcdf}(-\infty, t_0, \text{degrees of freedom}) \\ &= \text{tcdf}(-10^7, t_0, n-1). \end{aligned}$$

Equivalently, the null hypothesis is rejected if

$$t_0 < t_{n-1;\alpha}.$$

Two-Tailed Test

If the hypotheses are

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \ne \mu_0,$$

the null hypothesis will be rejected if the

$$\text{p} - \text{value} = P(t_{n-1} > |t_0|) + P(t_{n-1} < -|t_0|) < \alpha.$$

We shall now illustrate applications of the t test in real life problems where it is of interest to assess the population characteristic based on a sample from the population.

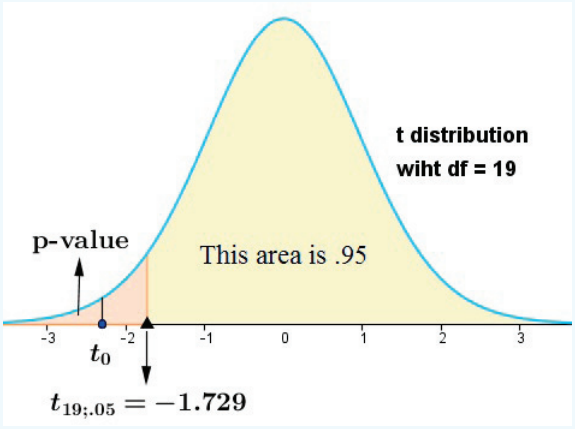


Figure 8.2: Illustration of critical value $t_{19;.05} = -1.729$ and the p-value

As an example, when $n = 20$, the 5th percentile of the t distribution with $n - 1 = 19$ degrees of freedom is illustrated in Figure 8.1. If the hypotheses are

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_a : \mu < \mu_0,$$

then the null hypothesis is rejected at the level of 0.05 if the test statistic

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -1.729.$$

or the p-value $P(t_{19} < t_0) < 0.05$, where t_0 is an observed value of the test statistic.

Table 8.1: Emission rates for a sample of 13 compact cars

car	1	2	3	4	5	6	7	8	9	10	11	12	13
emission rate per mile	.30	.35	.33	.27	.27	.32	.30	.24	.37	.22	.26	.33	.36

■ **Example 8.1** A sample of 13 compact cars was selected to assess the mean nitrogen-oxide emission of all compact cars. Each car was tested for nitrogen-oxide emissions, and the measurements (in grams per mile) as shown in Table 8.1 were collected: Suppose the emission standard limit for NOx is .4 gram per mile. Do the data provide sufficient evidence to conclude that, on average, the emission rate of all compact cars is less than .4 gram per mile? Test using $\alpha = .05$.

a. State the null and alternative hypotheses.
b. Calculate the t statistic and the p-value.
c. What can be concluded on the basis of the p-value?
d. For the testing method that you used, what is the probability of wrongly rejecting the null hypothesis when it is actually true?

Solution: The parameter of interest for this problem is $\mu =$ the mean amount emissions per mile for all compact cars.

a. As the objective of the test is to check if the mean μ is less than the emission standard limit .4 gram per mile, the hypotheses of interest are

$$H_0 : \mu \geq .4 \quad \text{vs.} \quad H_a : \mu < .4$$

b. The mean and SD of the sample are

$$\bar{x} = .3015 \quad \text{and} \quad s = .0471.$$

The test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{.3015 - .4}{.0471/\sqrt{13}} = -7.54.$$

Since this is a left-tailed test, the p-value is computed as

$$P(t_{n-1} \leq t_0) = P(t_{12} \leq -7.54) = 3.42864E - 06 \approx 0.$$

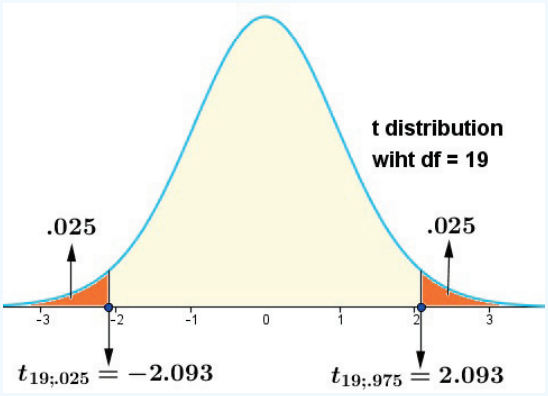


Figure 8.3: Illustration of critical points $t_{19;.025} = -2.093$ and $t_{19,.975} = 2.093$

As an example, suppose a t test based on a sample of size 20 is used to test the mean of a population with hypotheses $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$. Assuming that the level of significance is .05, the critical values are identified in Figure 8.3. Notice that the critical values are determined so that the left tail area and the right tail area are equal to .05/2 = .025. Since the t distribution is symmetric about zero, the left tail critical value is negative of the right tail critical value. Let t_0 be an observed value of the test statistic. The null hypothesis is rejected if

$$t_0 < -2.093 \text{ or } t_0 > 2.093.$$

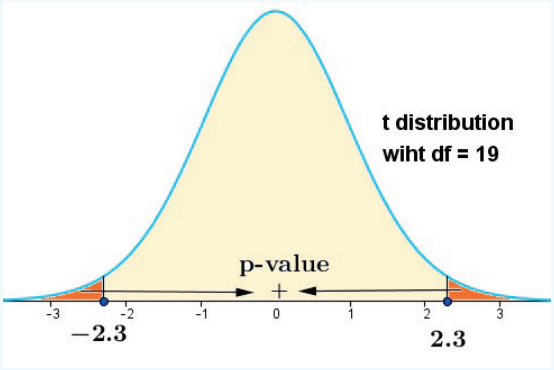


Figure 8.4: Illustration of p-value for an observed value $t_0 = 2.3$

Alternatively, the null hypothesis is rejected if the p-value

$$P(t_{19} < -|t_0|) + P(t_{19} > |t_0|) < \alpha.$$

This p-value is the sum of the orange colored areas in Figure 8.4 when $t_0 = 2.3$. Notice that the

The above p-value can be computed using TI-84 as follows:
Press [2nd] and [Distr], and select tcdf. Enter the values as shown.

$$\text{tcdf}(-10^7, -7.54, 12).$$

Press [enter] to get 3.42864E-06.

c. Since the above p-value is less than the level .05, we reject H_0, and conclude that the mean amount of emission per mile for all compact cars is less than .4.

d. Recall that the level of significance is the maximum probability of making type I error, and so the probability of rejecting the null hypothesis when it is actually true is no more than .05.

■

■ **Example 8.2** A automobile manufacturer claims that the average gas mileage of a compact car model 2013 is 28 mpg. A sample of 15 cars were test driven, and the following mileage were

recorded:

26 29 26 27 25 28 30 26 26 27 29 28 28 28 26

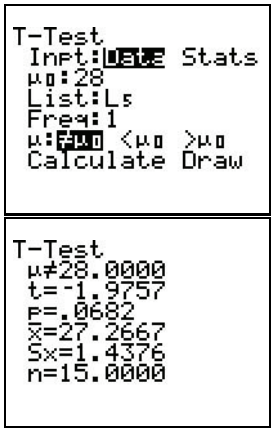
Do the data provide evidence to indicate that the average gas mileage is different from 28?

a. What is the parameter to be tested?
b. State the null and alternative hypotheses.
c. Calculate the t statistic and the p-value.
d. Write the conclusion.

■

Solution:

a. The parameter μ is the the true average mileage of all compact cars produced by the manufacturer in 2013.
b. $H_0 : \mu = 28$ vs. $H_a : \mu \neq 28$.
c. Enter the data in a list, say, L_5. Select [STAT], [TESTS] and [2: T-Test].
Select [Data], enter 28 for μ_0, enter [L5] for List.
Press [Calculate] to get the results as shown in the second figure.
The t-statistic is -1.9757 and p-value is .0682.
d. Since the p-value is not less than .05, the data do not provide enough evidence against the claim.



■ **Example 8.3** A machine is set to fill one liter of milk in plastic containers. At the end of a day's operation, a sample of 20 containers was selected and the actual amounts of milk in containers were measured (in ml) using an accurate method as follows: Does this sample provide enough

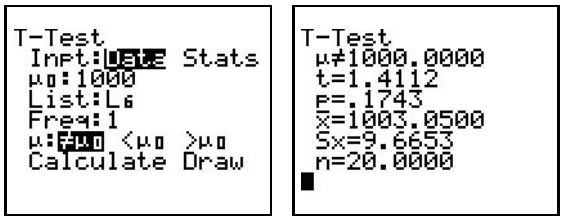
1010 1005 995 995 1004 999 990 1012 986 993
1005 1010 998 1021 995 1005 1008 1018 1016 996

evidence to indicate that the average amount of milk filled by the machine is different from a liter? Test using $\alpha = .05$.

a. What test do you use?
b. Identify the parameter of interest.
c. State the null and alternative hypotheses.
d. Calculate the t statistic and the p-value.
e. Write the conclusion.

Solution:

a. One-sample t test for mean
b. The true mean μ, the amount of milk filled by the machine in that day.
c. $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$.
d. Enter the data in a list, say L6. Select [STAT], [TESTS], [2: 1- T Test], and enter the data as shown below, slect [Calculate] and press [Enter]



e. Since the p-value is not less than .05, we can't reject H_0. The data do not provide enough evidence to indicate that the mean amount milk is different from one liter.

■

8.2 The t Confidence Interval for the Mean

Consider testing the mean μ of a population with the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0.$$

Recall the null hypothesis is rejected if $\left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| > t_{n-1;1-\alpha/2}$. The specified value μ_0 is accepted if

$$\left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| \leq t_{n-1;1-\alpha/2},$$

or equivalently

$$-t_{n-1;1-\alpha/2} \leq \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \leq t_{n-1;1-\alpha/2}.$$

Solving the inequality for μ_0, we see that $\mu = \mu_0$ is accepted if

$$\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}.$$

Thus, the $100(1-\alpha)\%$ CI for μ is given by

$$\left(\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right).$$

The above CI is referred to as the one-sample t interval or simply t interval for the mean. Also, note that the CI can written as

$$\bar{x} \pm t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm ME,$$

where ME is the

$$\text{margin of error} = t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}.$$

■ **Example 8.4** The following is a sample of size 12 from a population with mean μ.

```
13.8  9.1  9.7 11.6 13.4 11.1 14.2 12.6 12.7 10.9 12.7 10.7
```

Assuming that the population is approximately normal, find a 95% confidence interval for μ based on the above sample.

Solution: The mean and the standard deviation are computed as

$$\bar{x} = 11.875 \quad \text{and} \quad s = 1.617.$$

To find the 95% confidence interval for the mean, the required t percentile is

$$t_{n-1;1-\frac{\alpha}{2}} = t_{11;.975} = 2.201.$$

$$\begin{aligned}\left(\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right) &= \left(11.875 - 2.201\frac{1.617}{\sqrt{12}}, 11.875 + 2.201\frac{1.617}{\sqrt{12}}\right) \\ &= (11.875 - 1.027,\ 11.875 + 1.027) \\ &= (10.848, 12.902).\end{aligned}$$

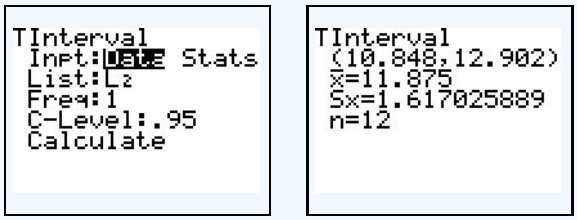
Thus, the population mean is between 10.848 and 12.902 with confidence 95%. The margin of error (ME) is 1.027. ■

TI Calc

Calculation of t Confidence Interval using TI-84

1. Select [STAT], [TESTS], [8: T Interval...]
2. Under [T Interval], select [Data] if data are available or [Stat] if only the mean $\bar{x}$ and the standard deviation s are available
3. Select the list where the data are stored
4. Enter the confidence level for [C-level]
5. Select [Calculate], and press [Enter]

The following figure illustrates the calculation of 95% confidence interval for the mean in Example 8.4, assuming that the data are stored in list L2:



Interpretation of a 95% Confidence Interval: In the above example, a single sample of size 12 yielded the 95% CI as (10.848, 12.902). Also, we can say that the population mean is somewhere between 10.848 and 12.902 with confidence 95%. This statement "95% confidence" can be explained as follows:

Consider 95% confidence intervals (using the t interval method) based on all possible samples of size 12 from the population. That is, for each sample we construct the 95% CI and so we have a collection of CIs. Among all these CIs, about 95% of them would include the population mean.

Does this particular 95% CI (10.848, 12.902) include the population mean? The answer is we are not sure. Since we know that 95% of all possible samples yield CIs that would include the population mean, we just believe that this interval $(10.848, 12.902)$ is one of them with 95% confidence.

■ **Example 8.5** The average weight of 18 weeks old pigs with the traditional feed is 80 kg. An animal feed manufacturer claims that a new supplemental diet for pigs increase the mean weight by at least 5 kg over a period of 17 weeks of feeding. A sample of 15 one week old baby pigs was given supplemental diet over a period of 17 weeks, and their weights (in kg) are recorded as given below.

```
84.7  88.0  90.5  87.4  85.1  85.6  82.9  84.3  87.0
92.2  80.5  88.2  91.9  81.9  85.0
```

Find a 95% confidence interval for the mean weight of all such one week old pigs after 17 weeks of supplemental diet.

a. What interval estimation method is appropriate for this problem?

b. What is the parameter of interest?

c. Calculate the 95% CI for the mean weight?

d. Write the CI in part **c** in the form "point estimate ± ME."

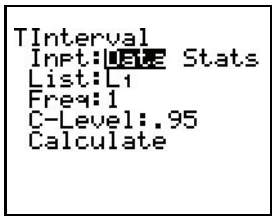
e. Does this 95% confidence interval support the manufacturer's claim?

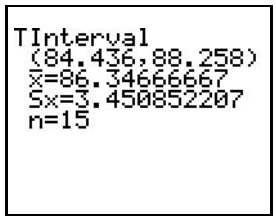
Solution:

a. One-Sample *t* interval.

b. The mean weight of all 18 weeks old pigs who were fed supplemental diet.

c. Assuming that the data are stored in list L1, we find the 95% CI as follows:





d. The CI is $(84.436, 88.258)$. The ME is

$$\frac{88.258 - 84.436}{2} = 1.911.$$

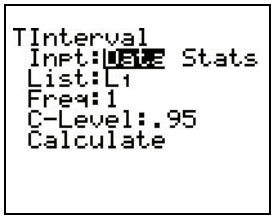
So, The CI can be expressed as 86.347 ± 1.911.

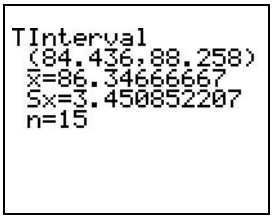
e. The CI indicates that the mean weight is between 84.4 and 88.3 kg with confidence 95%. In particular, we see that the true mean could be less than 85 kg. According to the manufacturer's claim, the mean weight should be at least $80 + 5 = 85$ kg. Since the confidence interval includes 85, on the basis of this confidence interval, we can not conclude that the mean weight increases by at least 5 kg after 17 weeks of supplemental diet.

■

TI Calc

Calculation of t-interval using TI-84: Assuming that the data are stored in list L1, we find the 95% CI as follows:





■ **Example 8.6** For the mileage data in Example 8.2, let us find a 95% CI for the mean mileage. From Example 124, we have

$$n = 15, \quad \bar{x} = 27.2667 \quad \text{and} \quad s = 1.4376.$$

The percentile required for the 95% CI for the mean is (using TI-84)

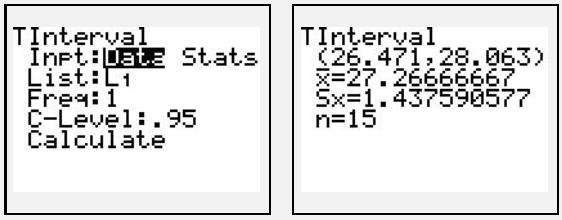
$$t_{n-1;.975} = t_{14;.975} = \text{invT}(.975, 14) = 2.145.$$

TI-84 function "invT" can be accessed by selecting [2nd] and [Distr]. So the 95% CI is

$$\left(27.2667 - 2.145\frac{1.4376}{\sqrt{15}},\ 27.2667 + 2.145\frac{1.4376}{\sqrt{15}}\right) = (26.47,\ 28.06).$$

The mean mileage of the compact cars is between 26.5 and 28 miles with confidence 95%.

Using [STAT], [TESTS], [TInterval]:



■

■ **Example 8.7** An independent agency decided to test the claim by Apple that their iPad battery life is 10 hours. A sample of 15 iPads was tested, and the mean and standard deviation of battery times are

$$\bar{x} = 9.8 \quad \text{and} \quad s = .64.$$

Suppose the agency decides to test the claim on the basis of a 95% CI for the mean battery time.

a. What interval estimation method can be used? [One-sample t interval]

b. Find the 95% CI for the mean battery time.
Select [STAT], [TESTS], [8: TInterval...], and select [Stats]. Enter 9.8 for $\bar{x}$, .64 for S_X, 15 for n, and .95 for [C-Level]; Select [Calculate] and press enter to get

$$(9.45, 10.15)$$

c. Write the CI in part **b** in the form "point estimate $\pm$ ME."
$ME = \frac{10.15-9.45}{2} = .35$. The CI is $\bar{x} \pm ME = 9.8 \pm .35$

d. What do you conclude on the basis of the CI in part **b.**
The average battery life is between 9.45 and 10.15 hours with confidence 95%.

e. What assumption is necessary for the CI in part **b** to be valid?
The life hours follow a normal distribution.

■

8.3 Exercises

1. According to recent National Health and Nutrition Examination Surveys (NHANES), the mean height of adult men in the USA is 69.3 inches with standard deviation 3 inches. A sociologist believes that taller people may be more likely to be promoted to positions of leadership, and hypothesizes that the mean height μ of all male executives may be greater than the national average. A simple random sample of 100 male executives has a mean height of 69.9 inch with standard deviation $S = 3$ inches. On the basis of this sample, can we conclude that, on the average, the male business executives are taller than the general male population at the level $\alpha = .05$?
   a. State the null and alternative hypotheses.
   b. Calculate the t statistic.
   c. Calculate the p-value.
   d. Write the conclusion.
   e. For the testing method that you used, what is the probability of wrongly rejecting the null hypothesis when it is actually true?

2. A automobile manufacturer claims that the average gas mileage of a compact car model 2013 is 28 mpg. A sample 15 cars were test driven, and the following mileage were recorded:

   26 29 26 27 25 28 30 26 26 27 29 28 28 28 26

   It is desired to estimate the true mean mileage of this particular model.
   - **a.** What interval estimation method is appropriate?
   - **b.** Find a 95% CI for the true mean mileage.
   - **c.** What is the margin of error of your CI?
   - **d.** On the basis of the CI in part **b**, can we conclude that the manufacturer's claim is not true? Explain.

9. Comparison of Two Means

9.1 Hypothesis Tests for Comparing Two Normal Means

We shall now consider hypothesis tests for comparing two normal population means. The setup is as follows. There are two testing methods for comparing two normal means are available:

	Populations	
	1	2
Population (mean, var)	(μ_1, σ_1^2)	(μ_2, σ_2^2)
Sample Sizes	n_1	n_2
Samples	$x_{11}, x_{12}, \ldots, x_{1n_1}$	$x_{21}, x_{22}, \ldots, x_{2n_2}$
Sample Means	$\bar{x_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$	$\bar{x_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$
Sample Variances	$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x_1})^2$	$s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2i}-\bar{x_2})^2$

1. The first one is referred to as the **two-sample t-test**, and is valid under the assumption that the samples are independent and the population variances are equal.
2. The second one is referred to as the **Welch test or two-sample t-test without assuming equality of variances** and is valid as long as the samples are independent and no assumption on variances is needed.

We shall first describe he two-sample t test assuming that the population variances are equal.

9.2 The Two-Sample t-test

The test statistic for testing equality of two normal means, under the assumptions that,

1. the samples are from normal populations,
2. the samples are independent, and
3. the population variances are equal,

is defined by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \text{ where } s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}. \tag{9.1}$$

Notice that the variance s_p^2 is obtained by pooling the independent sample variances s_1^2 and s_2^2, and so it is referred to as the pooled variance estimate. This t-statistic defined above has the t-distribution with degrees of freedom n_1+n_2-2.

Two-Tailed Test

For testing hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 \neq \mu_2,$$

the p-value is given by

$$\text{p-value} = P(t_{n_1+n_2-2} < -|t_0|) + P(t_{n_1+n_2-2} > |t_0|),$$

where t_0 is an observed value of t in (9.1). The cut-off approach is to reject H_0 if

$$|t_0| > t_{n_1+n_2-2;1-\frac{\alpha}{2}},$$

where α is the level of significance.

Right-Tailed Test

For testing hypotheses

$$H_0 : \mu_1 \leq \mu_2 \quad \text{vs.} \quad H_a : \mu_1 > \mu_2,$$

the p-value is given by

$$\text{p-value} = P(t_{n_1+n_2-2} > t_0),$$

where t_0 is an observed value of t in (9.1). The cut-off approach is to reject H_0 if

$$t_0 > t_{n_1+n_2-2;1-\alpha}.$$

Left-Tailed Test

For testing hypotheses

$$H_0 : \mu_1 \geq \mu_2 \quad \text{vs.} \quad H_a : \mu_1 < \mu_2,$$

the p-value is given by

$$\text{p-value} = P(t_{n_1+n_2-2} < t_0),$$

where t_0 is an observed value of t in (9.1). The cut-off approach is to reject H_0 if

$$t_0 < t_{n_1+n_2-2;1-\alpha}.$$

■ **Example 9.1** Sam Sleep researcher[1] hypothesizes that people who are allowed to sleep for eight hours will score significantly higher than people who are allowed to sleep for only four hours on a cognitive skills test. He brings sixteen participants into his sleep lab and randomly assigns them to one of two groups. In one group he has participants sleep for eight hours and in the other group he has them sleep for four hours. The next morning he administers the SCAT (Sam's Cognitive Ability Test) to all participants. Scores on the SCAT range from 1-9 with high scores representing better performance.

[1] source: http://web.mst.edu/∼psyworld/texample.htm

SCAT scores

8 hours sleep group (X1)	5	7	5	3	5	3	3	9
4 hours sleep group (X2)	8	1	4	6	6	4	1	2

Based on the above data, it is desired to test if the mean score for the 8-hour sleep group is higher than that for the 4-hour sleep group at the level $\alpha = .05$.

a. Identify the population means that are to be compared.
b. State the null and alternative hypotheses.
c. Assuming that the population variances are equal, calculate the t-statistic.
d. Find the p-value.
e. Write the conclusion.
f. State the assumptions under which the above test is valid.

Solution:

a. μ_1 = the mean score of all people who sleep 8 hours; μ_2 = the mean score of all people who sleep 4 hours.

b. $H_0 : \mu_1 \leq \mu_2$ vs. $H_a : \mu_1 > \mu_2$.

c. For these data,

$$\bar{x}_1 = 5, \;\; s_1^2 = 4.5714 \;\; \bar{x}_2 = 4, \;\; s_2^2 = 6.5714$$

and

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(8-1)4.5714 + (8-1)6.5714}{8+8-2} = 5.5714.$$

The t-statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{5-4}{\sqrt{5.5714\left(\frac{1}{8}+\frac{1}{8}\right)}} = 0.8473.$$

d. p-value $= P(t_{14} > .8473) = \text{tcdf}(0.8473, 10^7, 14) = 0.2055$

e. Since the p-value is not less than .05, we can not conclude that the mean score for 8 hour sleep group is higher than that for 4 hour sleep group.

f. The samples are from normal populations and they are independent.

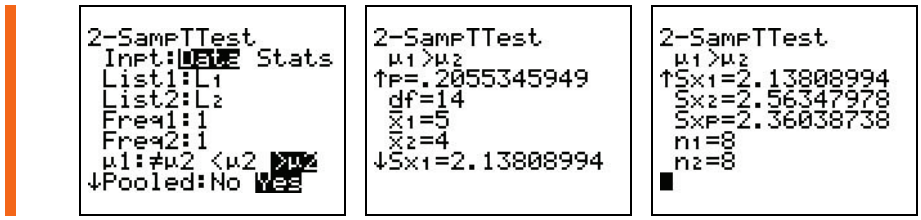
■

TI Calc

Two-Sample t Test for Comparing Two Means using TI-84

Calculation of statistics and p-value for Example 9.1 using TI-84. We also assume the data were stored in lists L1 and L2.

1. Select [STAT], [TESTS], [4: 2 SampTTest]
2. Enter the values as shown in the following figure.
3. Note that you should choose [Pooled: Yes], as we assume that population variances are equal.
4. Select [Calculate], and press [ENTER].



9.3 The Welch Test for Comparing Two Means

As noted earlier, this test is valid under the assumptions that

1. the samples are from normal populations, and
2. the samples are independent.

No assumption on the population variances is made. The test statistic is given by

$$w = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

This w statistic has an approximate t distribution with degrees of freedom

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}\right)}.$$

Follow the steps to carry out the Welch test:

1. For given two data sets, compute the means and variances:
$$\bar{x_1},\ \bar{x_2},\ s_1^2 \text{ and } s_2^2.$$
2. Calculate the degrees of freedom f using the above formula.
3. Calculate the p-value using the t distribution with degrees of freedom f.

■ **Example 9.2** In Example 9.1, we assumed that the population variances are equal. We shall now use the Welch test when the population variances are unknown and arbitrary. Consider

$$H_0 : \mu_1 \leq \mu_2 \quad \text{vs.} \quad H_a : \mu_1 > \mu_2,$$

where μ_1 = the mean of all people who sleep 8 hours, and μ_2 = the mean of all people who sleep 4 hours. Recall that for the "sleep data" in Example 9.1, the means and variances are

$$\bar{x_1} = 5,\ \ s_1^2 = 4.5714\ \ \bar{x_2} = 4,\ \ s_2^2 = 6.5714.$$

The test statistic

$$w = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5-4}{\sqrt{\frac{4.5714}{8} + \frac{6.5714}{8}}} = 0.8473.$$

The approximate degrees of freedom is

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}\right)} = \frac{\left(\frac{4.5714}{8} + \frac{6.5714}{8}\right)^2}{\left(\frac{4.5714^2}{8^2(8-1)} + \frac{6.5714^2}{8^2(8-1)}\right)} = 13.5631$$

The p-value is

$$P(t_{13.5631} > 0.8473) = \text{tcdf}(.8473, 10^7, 13.5631) = 0.2058.$$

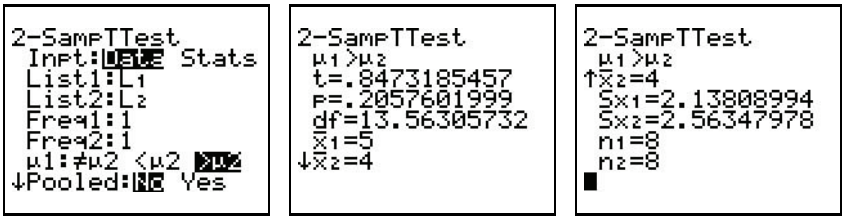
Since the p-value is not less than 0.05, we can not conclude that $\mu_1 > \mu_2$. In other words, the data do not provide enough evidence to support $H_a : \mu_1 > \mu_2$. ■

TI Calc

The Welch Test for Comparing Two Means using TI-84

Calculation of the Welch statistics and p-value for Example 9.1 using TI-84. We also assume the data were stored in lists L1 and L2.

1. Select [STAT], [TESTS], [4: 2 SampTTest]
2. Enter the values as shown in the following figure.
3. Note that you should choose [Pooled: No], as we do not assume that the population variances are equal.
4. Select [Calculate], and press [ENTER].



■ **Example 9.3** Two different drugs A and B for relieving pain are to be compared. A sample of 15 adults in the same age group was asked to use drug A when they suffer from headache, and to record the time (in minutes) until they are relieved completely. Another independent sample of 13 adults in similar age group was asked to use drug B for headache, and to record the time it takes to relieve pain. The results are as follows. Do the statistics in the following table provide sufficient

	drug A	drug B
Sample Size	$n_1 = 15$	$n_2 = 13$
Sample Means	$\bar{x}_1 = 25$	$\bar{x}_2 = 23$
Sample SD	$s_1 = 2.3$	$s_2 = 1.9$

evidence to conclude that the mean time to relieve pain for drug A is different from that for drug B?

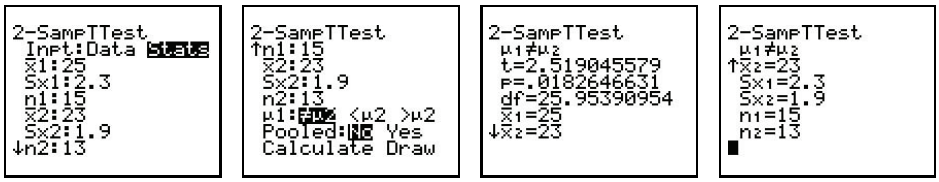
a. What test do we use for the above problem?
b. What are the parameters of interest?
c. State the null and alternative hypotheses.
d. What are the values of the test statistic and p-value?
e. What decision can be made on the basis of the p-value in part **d**?

Solution:

a. Since the problem is comparison of two means, and no assumption is made on the variances, the Welch test is appropriate.
b. μ_A = the mean time to relieve pain for drug A, and μ_B = the mean time to relieve pain for drug B.
c. As we are interested in finding if the mean times for these two drugs are different, the appropriate hypotheses are

$$H_0 : \mu_A = \mu_B \quad \text{vs.} \quad H_a : \mu_A \neq \mu_B.$$

d. Select [STAT], [TESTS], [4: 2 SampTTtest] from TI-84; enter the data as shown below:



The test statistic is 2.5190 and the p-value is 0.0183.

e. Since the p-value is less than 0.05, the null hypothesis is rejected. The data provide enough evidence to support the alternative hypothesis that the mean times to relieve pain for these two drugs are different at the level 0.05.

■

■ **Example 9.4** In order to compare the effects of two different fertilizers *A* and *B* on production of oranges, 20 randomly selected plots were treated with *A* and 16 randomly selected plots were treated with *B*. The results (in pounds) are as follows.

	Fertilizers	
	A	B
sample size	$n_A = 20$	$n_B = 16$
sample means	$\bar{x}_A = 560$	$n_B = 527$
sample SD	$s_A = 12.34$	$s_B = 10.45$

■

The society for farmers producing oranges may recommend the fertilizer *A* to *B*, if the average increase in the yield is 20 lbs or more. Carry out a test to see if the data provide enough evidence to support that the average increase is 20 lbs or more.

a. What are the parameters of interest.
b. State the null and alternative hypotheses.
c. Which test is appropriate for above hypotheses.
d. Calculate the test statistic and the p-value.
e. What is your conclusion?
f. What are the necessary assumptions for the conclusion to be valid?

Solution:

a. The parameters of interest are

$$\mu_A = \text{the mean yield per plot with fertilizer A}$$

and

$$\mu_B = \text{the mean yield per plot with fertilizer B.}$$

b.

$$H_0 : \mu_A - \mu_B \leq 20 \quad \text{vs.} \quad H_a : \mu_A - \mu_B > 20$$

or

$$H_0 : (\mu_A - 20) - \mu_B \leq 0 \quad \text{vs.} \quad H_a : (\mu_A - 20) - \mu_B > 0$$

c. As the samples are independent and no assumption is made on population variances, the Welch test is appropriate.

d. Using TI-84:

Select [STAT], [TESTS], [4: 2-SampTTest] and [Stats]; enter $560 - 20$ for $\bar{x}_1$, 12.34 for S_{x1}, 20 for n_1, 527 for $\bar{x}_2$, 10.45 for s_{x2}, 16 for n_2, select $> \mu_2$, and [No] for [Pooled]; press [Calculate] to get $t = 3.4212$ and $\text{p} - \text{value} = 0.0008$.	2-SampTTest μ1>μ2 t=3.421178648 p=8.222241E-4 df=33.86358268 x̄1=540 ↓x̄2=527

e. Since the p-value is close to 0, it is less than any practical level of significance, we reject the null hypothesis and conclude that on average the yield with fertilizer A is 20 lbs more than that of B.

f. The necessary assumptions are that the samples are independent, and the sample yields follow normal distributions.

■ **Example 9.5** An insurance company wants to find if the average speed at which men derive cars is greater than that of women drivers. The company took a sample of 27 cars driven by men on a highway,and found that the mean speed to be 73 miles pere hour with the SD of 2.3 miles per hour. Another sample of 18 cars driven by women in the same highway yielded the mean speed of 69 miles with the SD of 2.5 miles. It is desired to test if the mean speed by men exceeds that of women at the level .05.

a. What testing method is appropriate for this problem? Explain.

b. Compute the test statistic and p-value.

c. What is your conclusion on the basis of the p-value?

d. State the assumptions under which the test is valid.

■

9.4 Confidence Intervals for the Difference Between Two Means

We shall see a method for computing confidence intervals for the difference between two population means assuming that

1 the samples are independent, and
2 the samples are from normal populations.

The samples and statistics are as described earlier for the testing problems. Consider testing

	Populations	
	1	2
Population (mean, var)	(μ_1, σ_1^2)	(μ_2, σ_2^2)
Sample Sizes	n_1	n_2
Samples	$x_{11}, x_{12}, \ldots, x_{1n_1}$	$x_{21}, x_{22}, \ldots, x_{2n_2}$
Sample Means	$\bar{x_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$	$\bar{x_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$
Sample Variances	$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x_1})^2$	$s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x_2})^2$

$$H_0 : \mu_1 - \mu_2 = d \quad \text{vs.} \quad H_a : \mu_1 - \mu_2 \neq d.$$

Then the Welch test statistic is given by

$$w = \frac{\bar{x}_1 - \bar{x}_2 - d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The set of values of d for which the null hypothesis is not rejected is given by

$$-t_{f;1-\frac{\alpha}{2}} \leq \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{f;1-\frac{\alpha}{2}},$$

where

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}\right)}.$$

Solving the above inequality for d, we write the CI for $\mu_1 - \mu_2$ as

$$\bar{x}_1 - \bar{x}_2 - t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq d \leq \bar{x}_1 - \bar{x}_2 + t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Thus, the Welch CI for $\mu_1 - \mu_2$ is given by

$$\left(\bar{x}_1 - \bar{x}_2 - t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \bar{x}_1 - \bar{x}_2 - t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) = \bar{x}_1 - \bar{x}_2 \pm t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Follow the steps to find the Welch CI for the difference between two means:

1. For given two data sets, compute the means and variances:
$$\bar{x}_1,\ \bar{x}_2,\ s_1^2 \text{ and } s_2^2.$$
2. Calculate the degrees of freedom f using the above formula.
3. Find the t critical value $t_{f;1-\frac{\alpha}{2}} = \text{invT}\left(1 - \frac{\alpha}{2}, f\right)$.
4. Calculate
$$\bar{x}_1 - \bar{x}_2 \pm t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

■ **Example 9.6** Let us calculate a 95% confidence interval for $\mu_1 - \mu_2$ for Example 9.1, where $\mu_1 =$ the mean SCAT score all people who sleep 8 hours and $\mu_2 =$ the mean SCAT score all people who sleep 4 hours. For this example,

$$\bar{x}_1 = 5,\ \ s_1^2 = 4.5714\ \ \bar{x}_2 = 4,\ \ s_2^2 = 6.5714.$$

and

$$f = 13.5631 \quad \text{and} \quad t_{3.5631;.975} = \text{invT}(.975, 13.5631) = 2.1513.$$

Thus, the 95% CI for $\mu_1 - \mu_2$ is

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 \ &\pm\ t_{f;1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 5 - 4 \pm 2.1513\sqrt{\frac{4.5714}{8} + \frac{6.5714}{8}} \\ &= 1 \pm 2.5389 \\ &= (-1.5389, 3.5389). \end{aligned}$$

Notice that the margin of error of the above CI is 2.5389. Since the CI includes zero, we can not conclude that the means μ_1 and μ_2 are significantly different. ■

■ **Example 9.7** Do right-handed people live on average longer than left-handed people? One reason, some psychologists suggested[2], is that left-handers live in a world designed for right-handers. The stress of being left-handed in a right-handed world leads to earlier deaths among left-handers. Several studies have compared the life expectancies of left-handers and right-handers. The following results were obtained from one such study.

	Right-handed	Left-handed
Sample Size	$n_1 = 888$	$n_2 = 99$
Mean age at death	$\bar{x}_1 = 75.2$	$\bar{x}_2 = 66.8$
SD	$s_1 = 15.1$	$s_2 = 25.3$

Find a 95% confidence interval for the difference between average life expectancies of right-handed and left-handed people.

a. What interval estimate method is appropriate for this problem?
b. What is the parameter of interest?
c. Find a 95% confidence interval for the parameter.
d. What can we conclude from the confidence interval?

Solution:

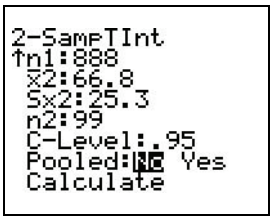
a. As the samples are independent, the Welch confidence interval (or two-sample t interval without assuming equality of variances) is appropriate for this problem.

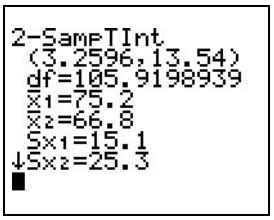
b. The parameter of interest is

$$D = \mu_1 - \mu_2,$$

where μ_1 is the average life expectancy of all right-handed people, and μ_2 is the average life expectancy of left-handed people.

c. Using TI-84: Select [STAT], [TESTS], [0: 2-SampTint...] from TI-84; enter the data as shown below:





The 95% CI for D is (3.26, 13.54).

d. On average, the life expectancy of right-handed people is 3.26 to 13.54 years greater than that of left-handed people with 95% confidence.

■

■ **Example 9.8** Do women pay less on average for car insurance than men? In almost all states in the USA, the average auto insurance premium for women is less than that of men. To judge the difference between the average annual premiums, a sample of 32 women and a sample of 29 men with similar policies were selected from Louisiana, and the following results were obtained.

	Men	Women
Sample Size	$n_1 = 29$	$n_2 = 32$
Mean Premium	$\bar{x}_1 = \$1{,}310$	$\bar{x}_2 = \$1{,}095$
SD	$s_1 = 57.5$	$s_2 = 68.7$

It is desired to estimate the true difference between the average auto insurance premiums for men and women in Louisiana.

[2]The New York Times; April 4, 1991

a. What interval estimate method is appropriate for this problem?
b. What is the parameter of interest?
c. Find a 95% confidence interval for the parameter.
d. What can we conclude from the confidence interval?

Solution:

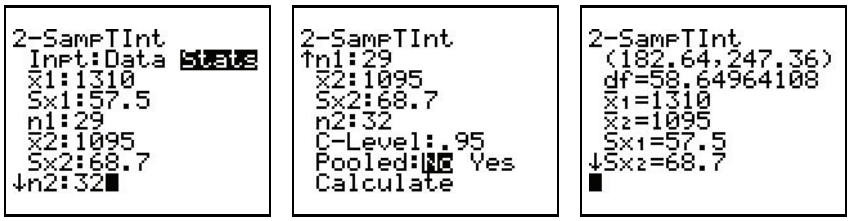
a. As the samples are independent, the Welch confidence interval (or two-sample t interval without assuming equality of variances) is appropriate for this problem.

b. The parameter of interest is

$$D = \mu_1 - \mu_2,$$

where μ_1 is the mean premium for women and μ_2 is the same for men.

c. Using TI-84: Select [STAT], [TESTS], [0: 2-SampTint...] from TI-84; enter the data as shown below:



The 95% CI for D is (182.64, 247.36).

d. On average, the premium for men is 183 to 247 dollars greater than that of women with 95% confidence.

■

■ **Example 9.9** An insurance company wants to find if the average speed at which men derive cars is greater than that of women drivers. The company took a sample of 27 cars driven by men on a highway,and found that the mean speed to be 73 miles pere hour with the SD of 2.3 miles per hour. Another sample of 18 cars driven by women in the same highway yielded the mean speed of 69 miles with the SD of 2.5 miles. It is desired to estimate the difference between the means.

a. What interval estimation method is appropriate for this problem? Explain.

b. Compute a 95% CI for the mean difference.

c. What is margin of error?

d. Interpret the meanings of the CI in part **b**.

e. State the assumptions under which the CI is valid.

■

9.5 The Matched-Pair t-Test

There are situations where the two samples are not independent. If the sample data are collected from the same set individuals over different time periods are not independent because the measurements at later time point may depend on the measurements at the initial time point. Another case where samples are not independent is the problem of comparing average weights of women before and after participating a diet program. Samples from siblings or twins are also not independent. In such cases, we should use the matched-pair t-test as described below.

Let $x_1,...,x_n$ denote measurements from a sample of n individuals at time point 1, and let $y_1,...,y_n$ denote the same at time point 2. The data for a matched-pair design are arranged as in the following table: Let μ_x and μ_y denote the means of x and y measurements in the populations,

Table 9.1: Matched-pair measurements

subjects	1	2	...	n
measurements (x)	x_1	x_2	...	x_n
measurements (y)	y_1	y_2	...	y_n
difference $d = x - y$	d_1	d_2	...	d_n

respectively. A matched-pair test is used to test on

$$\mu_d = \mu_x - \mu_y.$$

The test statistic is given by

$$t_d = \frac{\bar{d}}{s_d/\sqrt{n}},$$

where

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad s_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2.$$

Suppose it is desired to test

$$H_0 : \mu_d = 0 \quad \text{vs.} \quad H_a : \mu_d \neq 0.$$

Under the assumption that the d_i's follow a normal distribution, the test statistic

$$t_d \text{ has a } t\text{-distribution with degrees of freedom } n-1$$

when $H_0 : \mu_d = 0$ is true. Let t_{d0} be an observed value of t_d. The matched-pair t test rejects the above null hypothesis whenever the

$$\text{p-value} = P(t_{n-1} > |t_{d0}|) + P(t_{n-1} = \text{tcdf}(|t_{d0}|, 10^7, n-1) + \text{tcdf}(-10^7, -|t_{d0}|, n-1 < \alpha.$$

Notice that we simply apply one-sample t test using the data or the differences.

■ **Example 9.10** The manufacturer of an air conditioning/heating units introduced a new energy efficient unit which is supposed consume less electricity thereby saving money for homeowners. A sample of 15 houses with the new unit is selected, and the electric bills for the month of January after installing the new unit and for the same month a year before. The electric charges (in nearest dollar amount) are shown in Table 9.2. Test, on average, the new unit really reduces the electric cost?

a. Identify an appropriate test, and the necessary assumptions for the test to be valid.
b. Identify the parameters of interest for the hypothesis test.
c. State the appropriate null and alternative hypotheses.

Table 9.2: Electric costs for the month of January a year before and after installing the new unit

House	Before	After	d	House	Before	After	d
1	155	146	9	9	156	147	9
2	169	154	15	10	158	149	9
3	156	145	11	11	171	160	12
4	162	150	12	12	166	155	11
5	149	146	3	13	166	152	14
6	167	154	13	14	166	152	14
7	173	154	19	15	155	148	7
8	155	147	8				

d. Compute the test statistic and the p-value.

e. Write the conclusion of the test.

Solution:

a. Since the sample measurements "before" and "after" are collected from the same houses, the samples are dependent, and under the assumption that the differences follow a normal model, appropriate test is the matched-pair t test.

b. The parameters are

$$\mu_B = \text{the mean electric cost before installing the new unit,}$$

and

$$\mu_A = \text{the mean electric cost after installing the new unit.}$$

c. Let $\mu_D = \mu_B - \mu_A$. Since the problem of interest here is to test if the average electric cost before installing the new unit is higher than the average cost after installing the unit, the alternative hypothesis should be $\mu_D > 0$. So

$$H_0 : \mu_D \leq 0 \quad \text{vs.} \quad H_a : \mu_D > 0.$$

d. For this problem, the mean and standard deviation of d_i's are

$$\bar{d} = 11.07 \quad \text{and} \quad s_d = 3.83.$$

To get $\bar{d}$ and s_d, enter the d_i's in a "List" of TI-84, select [STAT], [CALC] and find the [1: 1-Var Stats]. The t-statistic is

$$\frac{\bar{d}}{s_d/\sqrt{n}} = \frac{11.07}{3.83/\sqrt{15}} = 11.19.$$

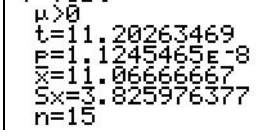
The p-value is $P(t_{14} > 11.19) = \text{tcdf}(11.19, 10^7, 14) = 0$.

e. Since the p-value is zero, the null hypothesis is rejected for any practical level of significance. There is strong evidence to support that $\mu_A < \mu_B$ or $\mu_B > \mu_A$.

■

TI Calc

Matched-Pair t Test Using TI-84: Let us calculate the test statistic and the p-value for the preceding example. Enter the values of the differences d in a list, say, L1. Select [STAT], [TESTS], [T-Test] and [Data], identify the list, select $> \mu_0$ and press [Calculate] to get test statistic 11.2026 and p-value 0



■ **Example 9.11** Are the front brake pads worn out faster than the ones in the rear? To find out, in a sample of 10 automobiles, the mileage (in 1,000 miles) at which the new front brake pads were worn to 4 mm from their original thickness was recorded. For the same sample of automobiles, the mileage at which the new rear brake pads were worn to 4 mm from their original thickness was recorded. The data are as shown below.

Automobile	1	2	3	4	5	6	7	8	9	10
Rear	47.1	49.0	50.7	55.9	46.4	47.4	48.8	50.2	52.0	50.6
Front	44.8	42.5	48.4	47.0	43.8	44.3	42.2	47.8	46.1	42.2
Differences, d	2.3	6.5	2.3	8.9	2.6	3.1	6.6	2.4	5.9	8.4

Test, on average, the rear brake pads last longer than the front ones.

a. Identify an appropriate test, and the necessary assumptions for the test to be valid.
b. Identify the parameters of interest for the hypothesis test.
c. State the appropriate null and alternative hypotheses.
d. Compute the test statistic and the p-value.
e. Write the conclusion of the test.

Solution:

a. As the sample measurements was collected from the same 10 automobiles, the samples are dependent, and so we should use the matched-pair t test.
b. The parameters of interest are

$$\mu_R = \text{the mean mileage of cars when the rear brake pads were worn to 4 mm}$$

and

$$\mu_F = \text{the mean mileage of cars when the front brake pads were worn to 4 mm}$$

c. $H_0 : \mu_R \leq \mu_F$ vs. $H_a : \mu_R > \mu_F$.
d. Enter the differences in a list of TI-84, and calculate (refer to page 142):

$$\text{test statistic} = 5.8566 \quad \text{and} \quad \text{p} - \text{value} = .0001.$$

e. Since the p-value is much smaller than any practical level of significance, we reject H_0 and conclude that, on the average the front brake pads worn out faster than the rear ones.

■

■ **Example 9.12** Does name-brand gas give better mileage than the cheap gas? There has been a common belief among consumers that the name-brand gas is likely to give better gas milage than cheap gas available at independent gas station. A group of 12 college students decided to check on this belief, and used cheap gas in their cars over a month and recorded the gas mileage. Then they all used name-brand gas over the next month and noted the gas mileage. The results are as follows.

Student	1	2	3	4	5	6	7	8	9	10	11	12
						mileage						
Name-Brand Gas	30	27	27	29	30	28	31	29	30	28	30	28
Cheap Gas	27	26	28	27	31	27	29	25	30	32	27	27

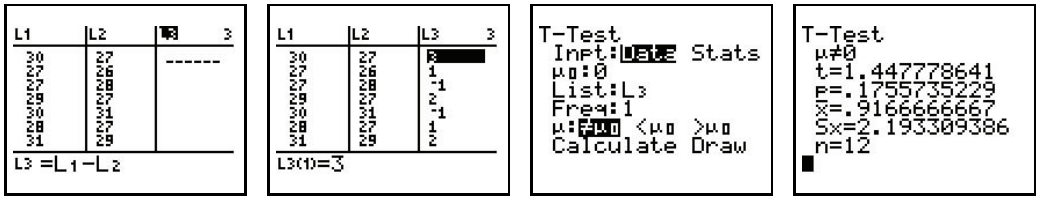
Do the data indicate that the average mileage with name-brand gas is different from that with cheap gas? ■

a. What test is appropriate for the problem?

b. Identify the parameter of interest.
c. State the appropriate hypotheses for the problem of interest.
d. Calculate the test statistic and the p-value.
e. Write the conclusion.

Solution:

a. Since the data were collected from the same cars, the samples are dependent and so the matched-pair t test is appropriate.
b. $\mu_d =$ the mean difference of mileage with the name-brand gas and the cheap gas.
c. $H_0 : \mu_d = 0$ vs. $H_a : \mu_d \neq 0$.
d. To find the differences of the mileage,
1 enter the mileage for the name-brand gas in, say, L1 and those for the cheap gas in L2.
2 Scroll to the top of the L3 column, press [2nd] and L1, press minus, press [2nd] and L2, then press [Enter].
3 All the differences appear in L3. Carry out the t test as shown below.



e. Since the p-value of 0.1756 is not less than .05, the data do not provide sufficient evidence to indicate that the mean gas mileage for name-brand gas is higher than that of cheap gas.

9.6 Matched-Pair Confidence Intervals

We shall now describe interval estimation method for matched-pair data. Consider the data in Table 9.1, and the hypotheses

$$H_0 : \mu_d = d_0 \quad \text{vs.} \quad H_a : \mu_d \neq d_0.$$

For a given level of significance α, the null hypothesis is rejected if

$$\frac{\bar{d} - d_0}{s_d/\sqrt{n}} < -t_{n-1;1-\frac{\alpha}{2}} \quad \text{or} \quad \frac{\bar{d} - d_0}{s_d/\sqrt{n}} > t_{n-1;1-\frac{\alpha}{2}}.$$

The set of values of d_0 for which the null hypothesis is not rejected is a confidence intervals, and is given by

$$\frac{\bar{d} - d_0}{s_d/\sqrt{n}} \geq -t_{n-1;1-\frac{\alpha}{2}} \quad \text{and} \quad \frac{\bar{d} - d_0}{s_d/\sqrt{n}} \leq t_{n-1;1-\frac{\alpha}{2}},$$

equivalently,

$$-t_{n-1;1-\frac{\alpha}{2}} \leq \frac{\bar{d} - d_0}{s_d/\sqrt{n}} \leq t_{n-1;1-\frac{\alpha}{2}}.$$

Solving the inequality for d_0, we obtain

$$\bar{d} - t_{n-1;1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \leq d_0 \leq \bar{d} + t_{n-1;1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}.$$

Thus, the CI for μ_d is given by

$$\left(\bar{d} - t_{n-1;1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}},\ \bar{d} + t_{n-1;1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} \right) = \bar{d} \pm t_{n-1;1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}},$$

which is the one-sample t interval based on the differences d_i's.

■ **Example 9.13** For Example 9.10, find a 95% CI for the difference between the mean electric cost before installing the new energy efficient unit and the mean electric cost after installing the unit. Recall that the differences are

```
d:  9  15  11  12  3  13  19  8  9  9  12  11  14  14  7
```

The mean and the standard deviation of the d_i's are

$$\bar{d} = 11.0667 \quad \text{and} \quad s_d = 3.8260.$$

Noting that the sample size is $n = 15$, the t critical value to find a 95% confidence interval is

$$t_{n-1;1-\frac{\alpha}{2}} = t_{14;.975} = \text{invT}(.975, 14) = 2.1448.$$

So the 95% CI is

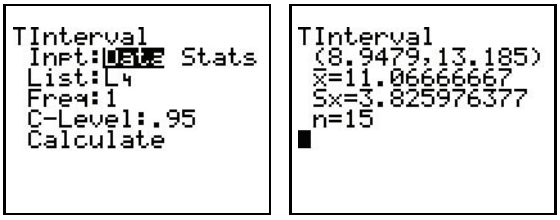
$$\bar{d} \pm t_{n-1;1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}} = 11.0667 \pm 2.1448 \frac{3.8260}{\sqrt{15}} = (8.95, 13.19).$$

Thus, on average, by using the new unit one could save 9 to 13 dollars in monthly electric cost. ■

TI Calc

Calculation of Matched-Pair CI for Example 9.13 Using TI-84:

a. Enter the values of d in a list, say, L4.
b. Select [STAT], [TESTS], [TInterval...]
c. Enter the values as shown below, and [Calculate]



■ **Example 9.14** In a clinical cohort study, a sample of 15 participants attended clinical examinations approximately every four years. Suppose we want to compare systolic blood pressures between examinations (that is, changes over 4 years). The data in Table 9.3 are systolic blood pressures measured at the first and second examinations for the sample of 15 participants. We

Table 9.3: Systolic blood pressures at two examinations

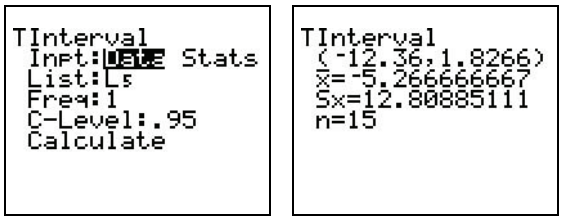
Subject	Examination 1	Examination2	Difference
1	168	141	-27
2	111	119	8
3	139	122	-17
4	127	127	0
5	155	125	-30
6	115	123	8
7	125	113	-12
8	123	106	-17
9	130	131	1
10	137	142	5
11	130	131	1
12	129	135	6
13	112	119	7
14	141	130	-11
15	122	121	-1

now estimate the mean difference in blood pressures over 4 years. This is similar to a one-sample problem with a continuous outcome except that we are now using the data on differences.

a. What interval estimation method is appropriate for the above problem?
b. Find a 95% CI for the mean difference in blood pressure.
c. Interpret the meanings of the CI.

Solution:

a. Since the blood pressures are measured from the same sample of individuals, the samples are dependent, and so we should use matched-pair CI.
b. Enter the values of the difference in a list, say L5, and find one-sample t interval. In TI-84, it is [8: TInterval ...]. The 95% CI is (-12.36, 1.83).



c. We are 95% confident that the mean difference in systolic blood pressures between examinations 1 and 2 (approximately 4 years apart) is between -12.4 and 1.8. Since the CI includes zero, the mean difference in blood pressures over time is not statistically significant.

■

Crossover Trials: Crossover trials are clinical trials in which each subject receives both of the two treatments (e.g., an experimental treatment and a control treatment involving placebo). Participants are randomly assigned to receive one of the treatments and, after a wash-out-period, the other treatment. Outcomes are measured after each treatment from each participant. An advantage of the crossover trial is that each participant acts as his/her own control, and so fewer participants are generally required to demonstrate an effect. If the outcomes can be measured on continuous scale, then the treatment effect can be assessed using matched-pair method.

■ **Example 9.15** A crossover trial is conducted to evaluate the effectiveness of a new drug designed to reduce symptoms of depression in adults over 60 years of age who suffered a stroke. Symptoms of depression are measured on a scale of $0-100$ with higher scores indicative of more frequent and severe symptoms of depression. The trial was run as a crossover trial in which each patient was blind to the treatment assignment and the order of treatments. That is, some participants receive first placebo and then new drug while others receive firs new drug and then placebo. After each treatment, depressive symptoms were measured in each patient. The difference in depressive symptoms was measured in each participant by subtracting the depressive symptom score after taking the new treatment from the depressive symptom score after taking the placebo. A total of 72 participants completed the trial and the data are as follows. It is desired to find a confidence

Sample Size n	Mean Difference of Symptoms Scores	SD of Differences
72	14.6	5.4

interval for the mean difference of symptoms scores.

a. What interval estimation method is appropriate for this problem?

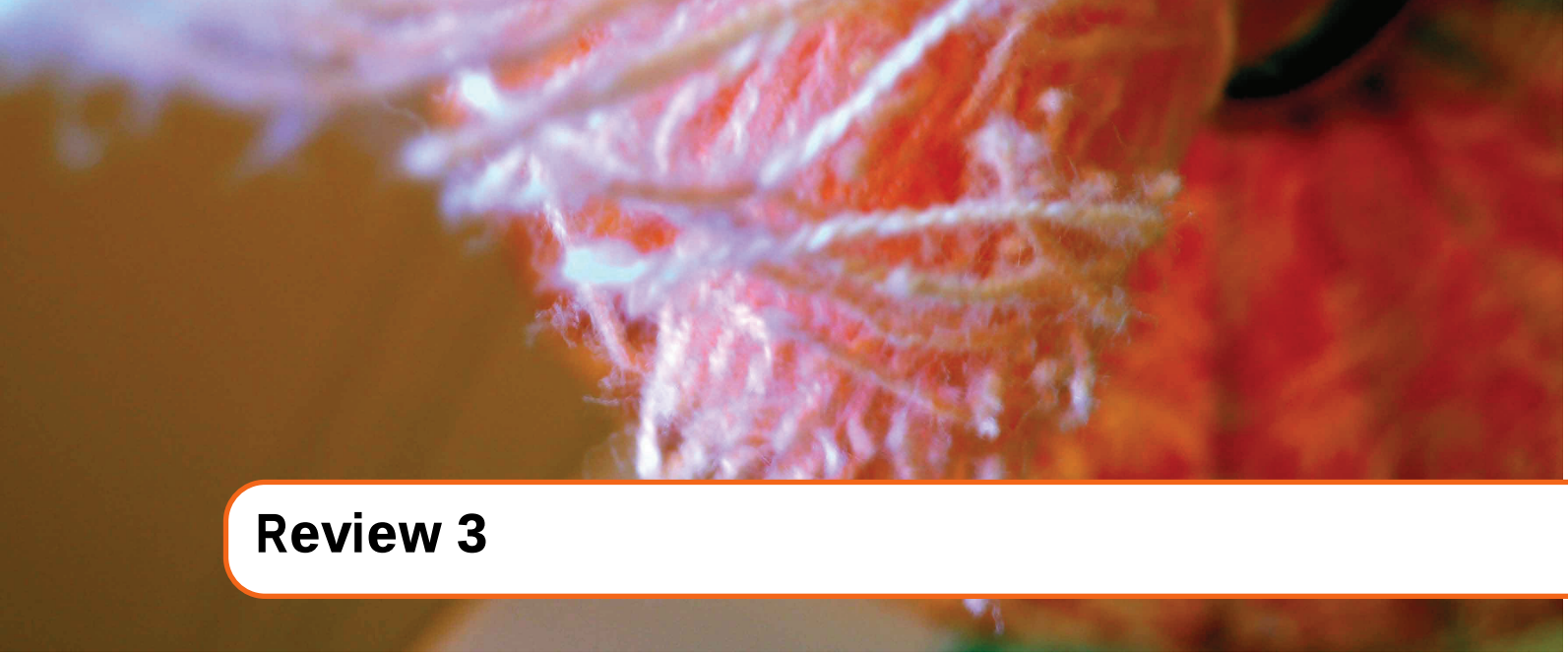
b. Find a 95% CI for the mean difference of symptoms scores.

c. What is the margin of error of the CI in part **b**?

d. On the basis of the confidence interval, can we conclude that the new drug is effective in reducing the symptoms of depression? Explain.

e. Can we conclude the same as in part **d** on the basis of a 90% CI? Explain.

■



Review 3

Topics Covered after the First Test

The following are just outlines of the materials covered after the second test. You should read the notes for more details. Go through all the worked out examples in the notes.

One Proportion	Hypothesis Test: [STAT], [TESTS] [1-PropZTest] Conf Interval: [STAT], [TESTS], [1-PropZint...]
Two Proportions	Hypothesis Test: [STAT], [TESTS] [2-PropZTest] CI for $p_1 - p_2$: [STAT], [TESTS] [2-PropZInt...]
Mean	Hypothesis Test: [STAT], [TESTS] [T-Test] CI for Mean: [STAT], [TESTS] [TInterval]
Comparing Two Means (independent samples)	Hypothesis Test: [STAT], [TESTS] [2-SampTTest] Conf Interval: [STAT], [TESTS], [2-SampTInt...]
Matched Pair (dependent samples)	Apply one-sample t method for the differences

1. ## Some Basic Terminologies in Hypothesis Test

   - type I error (false positive)
   - type II error (false negative)
   - Level of significance
   - Power
   - Margin of Error

2. **Some Important Results**

- If a test rejects H_0 at the level of significance .05, then the test should also reject H_0 at any level of significance $> .05$.
- If H_a is right-sided (i.e., $>$) and the test statistic is large, then the p-value should be small.
- If H_a is left-sided (i.e., $<$) and the test statistic is large negative, then the p-value should be small.
- The maximum probability of rejecting the null hypothesis when it is actually true is the level of significance α.
- The width of a CI increases with increasing confidence level.
- The margin of error of a CI is one half of the width of the CI.
- The margin of error increases with increasing confidence level.
- If a CI for the difference of proportions, say, $p_1 - p_2$ is of the form
  * $(+,+)$ then $p_1 > p_2$
  * $(-,+)$ then p_1 and p_2 are not significantly different
  * $(-,-)$ then $p_1 < p_2$

3. **Hypothesis Test and Confidence Interval for a Proportion P**

Let X denote the number of "successes" in a sequence of n trials. The sample proportion is $\widehat{p} = \frac{X}{n}$.

To compute the p-value and the test statistic using TI calculator:

select STAT $\rightarrow$ TESTS $\rightarrow$ 1-PropZTest;
enter the value of X, n and identify the alternative hypothesis;
select [Calculate], and press enter;

To compute the CI for p using TI calculator:

select STAT $\rightarrow$ TESTS $\rightarrow$ 1-PropZint;
enter the value of X, n and C-level; select Calculate, and press enter; For example, when $X = 21$, $n = 54$, to compute the 95% confidence CI for p, enter 21 for X, 54 for n, .95 for C-Level, highlight [Calculate] and press enter to get (.259, .519). The margin of error is

$$(.519 - .259)/2 = .13.$$

Furthermore, note that $\widehat{p} = 0.389$, and the CI is

$$\widehat{p} \pm ME = .389 \pm .13 = (.259, .519).$$

The true proportion is between .259 and .519 with confidence 95%.

4. **Comparison Between Two Proportions**

Here, we have X_1 successes out of n_1 trials, and X_2 successes out of n_2 trials. Also,

$$p_1 = \text{ true proportion in the first population,}$$

$$p_2 = \text{true proportion in the second population.}$$

Hypothesis test and CI for $p_1 - p_2$ can be calculated using the TI calculator as follows: Suppose $X_1 = 20$, $n_1 = 40$, $X_2 = 15$ and $n_2 = 50$, and we like to test

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_a : p_1 > p_2.$$

To calculate the p-value:

select STAT $\rightarrow$ TESTS $\rightarrow$ 2-PropZTest:

enter 20 for X_1, 40 for n_1, 15 for X_2, 50 for n_2 and select $> p_2$.
Select [Calculate] and press enter to get: test-statistic Z = 1.933
p-value p = .027. Since the p-value is less than 0.05, we reject H_0. That is H_a is true.
We conclude that p_1 is greater than p_2.

90% CI for $p_1 - p_2$
Select STAT $\rightarrow$ TESTS $\rightarrow$ 2-PropZInt...
enter 20 for X_1, 40 for n_1, 15 for X_2, 50 for n_2 and .90 for [C-Level];
select [Calculate] and press enter to get:
(0.032, .368).
Note that both endpoints of the CI for $p_1 - p_2$ are positive. This implies that $p_1 > p_2$. Furthermore, the difference $p_1 - p_2$ is between 0.032 and .368 with 90% confidence. We can also say that p_1 is 0.032 to .368 larger than p_2.

Some Problems from the Topics Covered after the Second Test

1. In the following choose the correct answer.
   - **i.** In the following identify the correct answer.
     - **i.** The power of a test is
       - **a.** the probability of rejecting null hypothesis when it is false;
       - **b.** the probability of accepting null hypothesis when it is true;
       - **c.** the probability of making type I error;
       - **d.** none of the above.
     - **ii.** The type II error is
       - **a.** accepting null hypothesis when it is false
       - **b.** wrongly rejecting null hypothesis when it is true
       - **c.** accepting the null hypothesis when it is true
       - **d.** also called false positive rate
     - **iii.** The p-value for a hypothesis test is 0.044. Then
       - **a.** the null hypothesis will be rejected for any level of significance of .05 or less.
       - **b.** the null hypothesis will be rejected for any level of significance of .05 or greater.
       - **c.** the null hypothesis will be rejected at the level of significance 0.01.
       - **d.** none of the above.
     - **iv.** A 95% CI for the mean of a population on the basis of a sample is 12 ± 2.
       - **a.** The margin of error of the above CI is 1.
       - **b.** Based on the above CI, we can say that the population mean is greater than or equal to 10.
       - **c.** A 90% CI based on the same sample must be wider than $(10, 14)$.
       - **d.** A 99% CI based on the same sample must be shorter than $(10, 14)$.
     - **v.** Three CIs are constructed for a population proportion p based on a sample with sample proportion $\widehat{p} = .3$. Identify the one that could be a reasonable CI.
       - **a.** 90% CI: (.25, .35)
       - **b.** 95% CI: (.40, .50)
       - **c.** 99% CI: (.10, .25)
2. Three CIs are constructed for a population proportion p based on a sample. They are
   - **a.** 90% CI: (.63, .75)
   - **b.** 95% CI: (.57, .81)
   - **c.** 99% CI: (.33, 1.05)

   Identify the one that is incorrect.

3. A survey was conducted to estimate the percentage of households in a city with annual income \$65,000 or more. In a sample of 300 households, 100 are with annual income of \$65,000 or more.
   a. What is the appropriate interval estimating method?

   b. What is the sample estimate of the percentage?

   c. Find a 95% CI for the true percentage of households with annual income \$65,000 or more.

   d. On the basis of the 95% CI, can we say that "with 95% confidence about 28 to 39 percentage of households in the sample make annual income \$65,000 or more? Explain.

4. A random sample of 100 electronic components was tested to estimate the average lifetime. The sample mean is 125 hours with the standard deviation of 5 hours. It is desired to estimate the mean life hours of all electronic components produced by the manufacturer.
   a. What is the appropriate interval estimating method?

   b. Find a 95% CI for the mean.

   c. What is the ME of your CI?

   d. On the basis of the 95% CI, can we conclude that the mean life hours is at least 115 hours? Explain.

5. According to the credit rating company Equifax, credit limits on newly issued credit cards have increased between Jan 2011 and May 2011. The following data were collected to estimate the mean difference.

Jan 2011	May 2011
$n_1 = 400$	$n_2 = 500$
$\bar{x}_1 = \$2635$	$\bar{x}_2 = \$2887$
$s_1 = \$365$	$s_2 = \$412$

   a. What interval estimation method is appropriate for this problem?

   b. Find a 95% CI for the difference between the mean credit limits in Jan 2011 and May 2011.

c. Does the CI indicate that the mean credit limit has been increased since Jan 2011? Explain.

6. The manufacturer of a gasoline additive claims that the use of this additive increase gasoline mileage. A random sample of 10 cars were driven one week without additive, and then for one week with the gasoline additive with the following data: It is desire to estimate the

mileage data

cars	1	2	3	4	5	6	7	8	9	10
no additive	24	24	23	25	25	25	28	28	24	26
additive	28	27	27	28	30	28	25	28	27	25

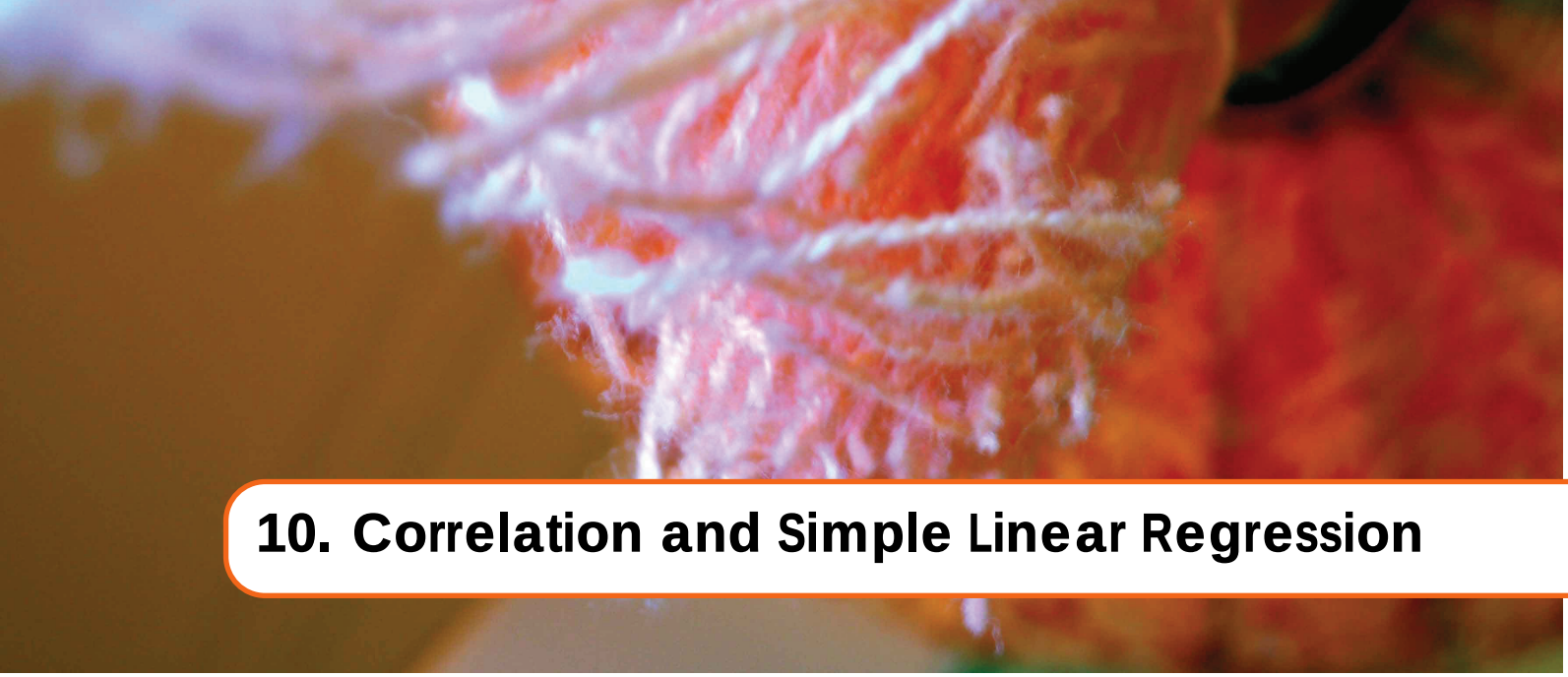
mean difference in mileage of gasoline without additive and with additive.

a. What interval estimating method is appropriate for this problem?

b. Find a 95% CI for the mean difference in mileage.

c. What is the average difference?

d. On the basis of the CI in part **b**, can we conclude that the gasoline additive increase the mean mileage? Explain.



10. Correlation and Simple Linear Regression

10.1 Correlation Coefficient

The correlation coefficient, denoted by r, is a measure of the strength of the linear relationship between two variables. It is a numerical value that measures the strength of the linear relationship and the direction of the relationship. The correlation coefficient ranges in value between -1.0 and +1.0. The correlation coefficient is interpreted based on its sign (positive or negative) and its absolute value.

- A perfect positive correlation has a coefficient of 1.0.
- A perfect negative correlation has a coefficient of -1.0.
- When there is no association between two variables, the correlation coefficient has a value very close to zero.
- Thus, the correlation coefficient r satisfies

$$-1 \leq r \leq 1.$$

Examples where the correlation analysis is useful are

- $X =$ midterm scores of students in a course
  $Y =$ scores in the final exam
- $X =$ heights of a sample of soldiers
  $Y =$ weights of the same sample of soldiers
- $X =$ monthly expenditure on advertisement of a product
  $Y =$ monthly sales revenue
- $X =$ IQ score
  $Y =$ math score

■ **Example 10.1** The following are the IQ scores and math scores of a sample of 20 college students:

students	1	2	3	4	5	6	7	8	9	10
IQ score, X	110	118	93	94	97	116	109	109	108	93
math score, Y	72	76	65	68	67	77	73	74	70	66
students	11	12	13	14	15	16	17	18	19	20
IQ score, X	96	96	91	95	108	100	89	100	98	100
math score, Y	67	66	67	66	75	69	66	72	70	69

The correlation coefficient is a valid measure of association only when the variables have linear relationship. A linear relation between two variables can be checked by examining the sca ter plot; the plot in which X values are plotted on x-axis and the corresponding Y values are plotted on y-axis. ∎

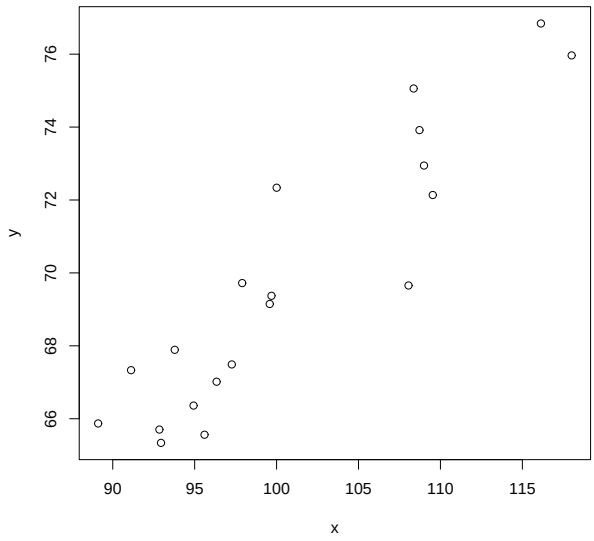


Figure 10.1: Scatter plot of IQ scores x and math scores y

The scatter plot clearly indicates that there is a positive linear association between the variables X and Y. To quantify the association by a numerical measure, we calculate the correlation coefficient r as follows.

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \\ &= \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{(n-1)S_XS_Y}, \end{aligned}$$

where S_X is the standard deviation of X values and S_Y denotes the standard deviation of Y values.

For the above data on math scores and IQ scores,

$$\sum_{i=1}^{20}X_iY_i = 141446,\ \bar{X} = 101,\ \bar{Y} = 69.75,\ S_X = 8.45,\ \text{and } S_y = 3.73.$$

Noting that the sample size $n = 20$, the correlation coefficient between the IQ scores and math

score is

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{(n-1)S_X S_Y} \\ &= \frac{141446 - 20(101)(69.75)}{19(8.45)(3.73)} \\ &= 0.9216. \end{aligned}$$

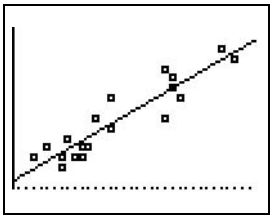
TI Calc

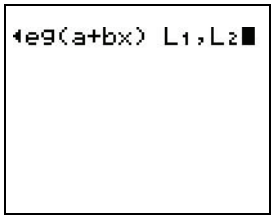
Scatter plot and correlation using TI-84:

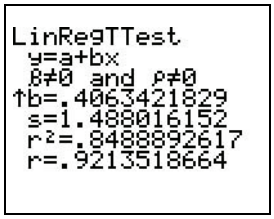
1. Enter the X values in L1, and Y values in L2
2. Press [2nd], [STATPLOT], and [Plot 1 On]
3. On Plot 1 screen, select the first icon for Type, identify L1 for [Xlist] and L2 for [Ylist]
4. Press [GRAPH] to get the first figure below. If it does not work, try pressing [ZOOM] and [ZOOM STAT]

To find correlation:

1. This step is required for the first time. Press [2nd] and [CATALOG]; scroll down to [DiagnosticOn]; press [ENTER] and press [ENTER] again to see [Done].
2. Press [STAT], [CALC], select [8:LinReg(a+bx)]
3. Identify the lists by pressing [2nd],L1,comma,[2nd] and L2
4. Press [ENTER] to get the results as shown in the second figure







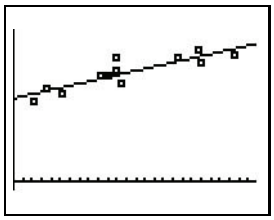
■ **Example 10.2** Do heavier people burn more calories? Metabolic rate is important in studies of weight gain, dieting and exercise. Data on the lean body mass and resting metabolic rate were collected from a sample of 12 women who are subjects in a study of dieting. Lean body mass is a person's weight leaving out all fat. Metabolic rates is measured in calories burned per 24 hours.

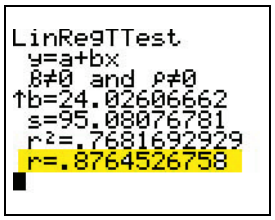
Lean body mass and metabolic rate												
Mass (kg)	36.1	54.6	48.5	42.0	50.6	42.0	40.3	33.1	42.4	34.5	51.1	41.2
Rate	995	1425	1396	1418	1502	1256	1189	913	1124	1052	1347	1204

a. Construct scatter plot and interpret it.

b. Compute the correlation coefficient.

Solution:





a. The scatter plot indicates a positive linear relation between lean body mass and metabolic rate.

b. The correlation coefficient

$$r = 0.88.$$

There is a strong positive linear relation between lean body mass and metabolic rate.

■

Some Features of Correlation

1. Correlation is used as a measure of association when both X and Y are random variables.
2. Values of correlation coefficient being close to 1 or -1 indicate that the variables are strongly related.
3. Correlation near zero indicates that the variables are not associated or uncorrelated.
4. The correlation coefficient is not affected by the units of measurements. For example,consider the following data represent heights and weights of a sample of 10 people.

```
height: 69.3 66.3 69.0 68.2 64.5 64.4 65.0 65.8 66.3 68.1 (in inches)
weight: 66.4 63.0 68.8 67.9 60.4 57.2 60.8 61.4 61.8 68.0 (in kilograms)
```

The correlation coefficient is 0.936. Suppose we transform the data by converting inches to centimeters and kilogram to pound. Noting that 1 inch = 2.54 cm and 1 kilogram = 2.2 lbs, the transformed data are

```
height: 176.1 168.3 175.1 173.2 163.9 163.5 165.0 167.1 168.4 172.9 (in cms)
weight: 146.0 138.5 151.3 149.3 132.9 125.8 133.8 135.0 135.9 149.6 (in lbs)
```

The correlation coefficient based on the above transformed data is also 0.936.

5. Correlation near 1 or -1 simply tells us that the two variables are strongly associated. However, a correlation between two variables does not necessarily imply that one causes the other. For example, there are numerous studies indicate that there is a high positive correlation between the salary of a person and his/her height. However, there is no logic to support that there is a connection between heights and salaries. The correlation could be due to coincident effects of some common cause.
6. [1]Empirically observed correlation is a necessary but not sufficient condition for causality. Correlation is not causation but it sure is a hint.

10.2 Linear Regression

Linear regression is an approach for modeling the relationship between two variables referred to as the response variable or dependent variable (Y) and the explanatory variable (also called predictor variable) x. Here, Y is a random variable whereas x is fixed. As an example,

- a company may want to understand how past advertising expenditures have related to sales in order to make future plans on advertising and increasing sales revenue. The response or dependent variable in this instance is sales (Y) and the explanatory variable is advertising expenditures (x).
- One may want to assess the relation between the price of a house Y and the living area x in a neighborhood.

The case of one explanatory variable is called *simple linear regression*. Linear regression model with more than one explanatory variables is called *multiple linear regression*. Examples where linear relationship between two variables are:

[1]Tufte, E. R. (2006). The Cognitive Style of Power Point: Pitching Out Corrupts Within (2nd ed.). Cheshire, Connecticut: Graphics Press.

1. Y = the price of a used car
   x = the age of the car
2. Y = salinity of water
   x = electric conductivity measurements
3. Y = the gas mileage of a car
   x = the weight of the car
4. Y = monthly sales of a name-brand TV
   x = monthly expenditures on advertisement.

The data on x and Y are collected as shown in the following table.

subjects	1	2	3	...	n
x	x_1	x_2	x_3	...	x_n
Y	y_1	y_2	y_3	...	y_n

10.3 Model Fitting

To fit a linear model that describes the relationship between the variables Y and X mathematically, follow the steps:

1. Construct the scatter plot by plotting the x-values on the x-axis and the corresponding Y-values on the y-axis. If the data exhibit a linear relation (see Figure 10.3), then proceed as follows.
2. We assume the linear regression model that

$$Y = a + bx + \varepsilon, \tag{10.1}$$

   where
   a = intercept parameter of the model
   b = the slope parameter
   ε = random error.
3. For given pairs of data $(x_1, Y_1), \ldots, (x_n, Y_n)$, the model parameters are estimated by the values a and b that minimize the error sum of squares

$$\sum_{i=1}^{n} (Y_i - a - bx_i)^2.$$

   Such estimates are called the least square estimates, and they are given by

$$\widehat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}}{(n-1)s_x^2},$$

   where s_x^2 is the variance of the x-data, and

$$\widehat{a} = \overline{Y} - \widehat{b}\,\overline{x}.$$

4. The fitted model is

$$\widehat{Y} = \widehat{a} + \widehat{b}\,x.$$

5. The squared correlation between Y and x is called the **the coefficient of determination,** and is denoted by r^2. The coefficient of determination is the proportion of variation in Y explained by model's prediction. Note that

$$-\leq r \leq 1 \qquad \Longleftrightarrow \qquad 0 \leq r^2 \leq 1.$$

The r^2 also indicates how well the data fit the linear regression model. If r^2 is near 1, then we can say that the model is well fitted, and the fitted model can be used for predicting Y for a given x value.

■ **Example 10.3** The data in Table 10.1 represent the salinity level in water and the corresponding electric conductivity measurement.

In order to find if the variables Y and x are linearly related, we first construct the **scatter plot** by plotting X value on x-axis and the corresponding Y value on y-axis. This scatter plot can be obtained using TI-84 as follows:

1. Enter the data on x in list L1, and those on Y in list L2.
2. Press [2nd] [STAT PLOT] and select [Plot 1 On]. Then press [ENTER] on each of the options as shown in the second box of Figure 10.2.
3. Press [GRAPH] (or [ZOOM] and [9: ZOOM STAT]) to get the scatter plot as shown in the third box of Figure 10.2.

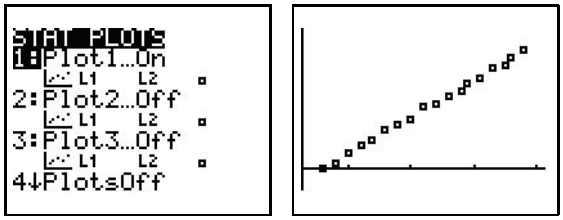


Figure 10.2: Illustration of scatter plot using TI-83/84

The plot in the second box clearly indicates that these two variables are linearly related. The parameters are estimated by the least square method as

$$\widehat{b} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{(n-1)s^2} = \frac{296.21 - 18 \times 3.256 \times 4.292}{17.383} = 2.569$$

and

$$\widehat{a} = 4.292 - 2.569 \times 3.256 = -4.073.$$

Thus the fitted model is

$$\widehat{Y} = -4.073 + 2.569\, x.$$

■

Checking Model Adequacy: As the coefficient of determinant $r^2 = .997$ (see the third box of TI calculator output below) is very close 1, we can say that the linear regression model is well fitted.

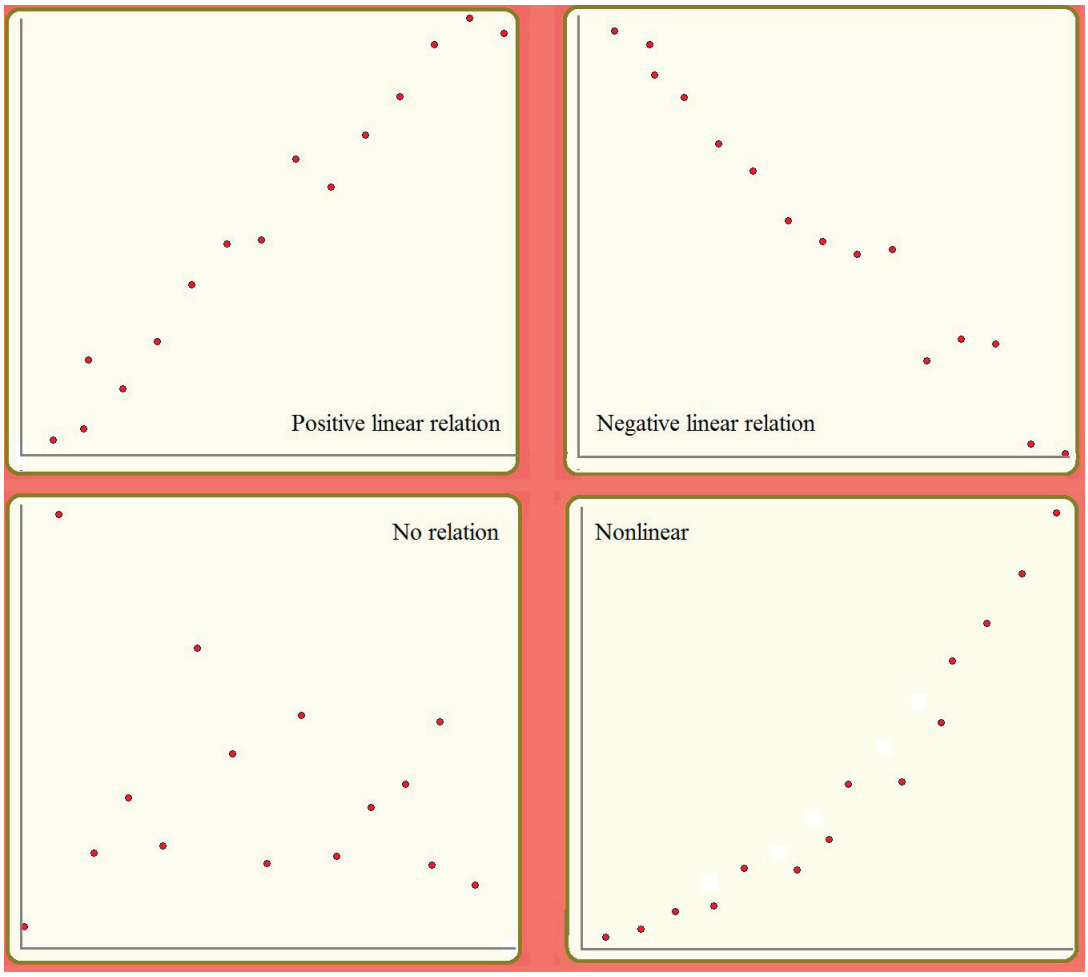


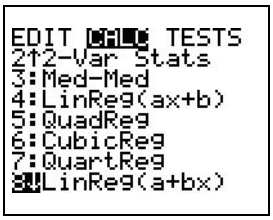
Figure 10.3: Scatter plots

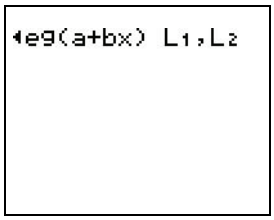
Table 10.1: Electric conductivity measurements (x, in micromoles/cm^3) and salinity level (Y, in ml)

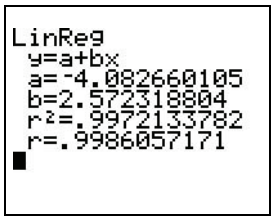
Sample No.	x	Y	$\widehat{Y}$	$(Y-\widehat{Y})^2$	Sample No.	x	Y	$\widehat{Y}$	$(Y-\widehat{Y})^2$
1	1.6	0	0.037	.0014	10	3.4	4.6	4.662	.0038
2	1.8	.45	0.551	.0102	11	3.6	5.1	5.175	.0057
3	2.0	1.1	1.066	.0012	12	3.8	5.4	5.689	.0836
4	2.2	1.7	1.579	.0147	13	3.9	6.1	5.946	.0237
5	2.4	2.1	2.093	.0000	14	4.1	6.3	6.460	.0256
6	2.6	2.7	2.606	.0088	15	4.3	7.1	6.974	.0160
7	2.8	3.2	3.120	.0064	16	4.5	7.3	7.488	.0352
8	3.0	3.5	3.634	.0180	17	4.6	7.9	7.744	.0242
9	3.2	4.3	4.148	.0232	18	4.8	8.4	8.258	.0201
								$\sum_{i=1}^{18}(Y_i-\widehat{Y}_i)^2=$	.3217

TI Calc

Calculation of the Least Squares Estimates Using TI-84:
Press [STAT], [CALC] and select [8: LinReg(a+bx)] as shown in the first box below. On the screen LinReg(a+bx), press [2nd], [L1], press comma, [2nd] [L2]; press [ENTER] to get the results as shown in the third box. Because our earlier calculation of $\widehat{b}$ and $\widehat{a}$ involved roundoff errors, they are slightly different from the estimates in the third box.







10.4 Test on the Slope Parameter

Once the model is fitted, it is desired to test if the model slope parameter is different from zero. If the slope parameter is significantly different from zero, then the fitted model is valid, and it can be used for prediction as described in Section 10.5.

Consider testing

$$H_0 : b = 0 \quad \text{vs.} \quad H_a : b \neq 0. \tag{10.2}$$

A test for the slope parameter b can be developed assuming that the random error in the model 10.1 is normally distributed with mean 0 and variance σ^2. Under this assumption, the test statistic

$$t = \sqrt{n-1}\frac{\widehat{b}s_x}{s_e}, \tag{10.3}$$

where s_x^2 is the variance of x data, and

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{n-2}}, \quad \text{with } \widehat{y}_i = \widehat{a} + \widehat{b}x_i, \quad i = 1, \ldots, n,$$

which is an estimate of σ. The $\widehat{y}_i$ is the predicted value of y_i at x_i. The test statistic

$$t = \sqrt{n-1}\frac{\widehat{b}s_x}{s_e} \text{ has } t \text{ distribution with df} = n-2.$$

Let t_0 be an observed value of t. Then the p-value for testing hypotheses in (10.2) is given by

$$P(t_{n-2} < -|t_0|) + P(t_{n-2} > |t_0|).$$

Example 10.3 continued. For this example, let us test

$$H_0 : b = 0 \quad \text{vs.} \quad H_a : b \neq 0.$$

As the x-data are stored in L1, using [STAT], [CALC] and [1 Var Stats], we find

$$s_x = 1.0112.$$

To find s_e, we first need to calculate $\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2$:

1. select [STAT], [1 Edit], move the cursor to the top of L3, press [ENTER]

2. type $(4.073 + 2.569 * L1)$, press [ENTER]. Now the numbers in L3 are the predicted values $\widehat{Y}_i$.
3. move the cursor at the top of L4, press [2nd] L2, minus and press [2nd] L3 so that L4 = L2 - L3
4. press [STAT], [CALC], [1 Var Stats], [2nd] and L4
5. The value at $\sum x^2 = 0.3217$ is the value of $\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2$.
6. Finally, $s_e = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2}{n-2}} = \sqrt{\frac{0.3217}{16}} = 0.1418$.

Thus, the observed value of test statistic is

$$t_0 = \sqrt{n-1}\frac{\widehat{b}\, s_x}{s_e} = \sqrt{17}\frac{2.569 \times 1.0112}{.1418} = 75.54.$$

The p-value is

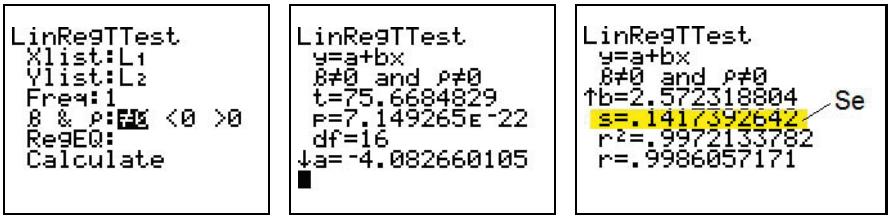
$$P(t_{16} < -75.54) + P(t_{16} > 75.54) = \text{tcdf}(-10^7, -75.54, 16) + \text{tcdf}(75.54, 10^7, 16) = 0.$$

Since the p-value is less than 0.05, we reject $H_0 : b = 0$, and conclude that the slope parameter is significantly different from zero.

TI Calc

The t test for the slope:

To find the test statistic and the p-value, press [STAT], [TESTS], and select [F: LinRegTTest]; select the options as shown in the first box below and press [ENTER] to get the results as shown in the second and third boxes.



Notice that the t statistic is **75.6685** and the p-value is **7.149265E-22 = 0**. These two values are little different from the ones we calculated earlier because earlier calculations involve roundoff errors. Also, notice that $\mathbf{s_e = 0.1417}$, practically the same as the one we obtained earlier.

10.5 Prediction Interval and CI for the Mean Response

Once the model is well fitted, we could use the model to predict the value of Y at a given x^*. The point prediction is

$$\widehat{Y}^* = \widehat{a} + \widehat{b}x^*.$$

The prediction interval for Y at a given x^* is

$$\widehat{Y}^* \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}, \tag{10.4}$$

where s_x^2 is the variance of the x-data, and $s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2$.

The above prediction interval is for Y for a single unit for which is x^* is known. For example, predicting the price of a house with the living area 2,600 square feet in a neighborhood. Sometimes,

we want to estimate the average price of all houses in the neighborhood with the same living area x^*. That is, we want to estimate $E(Y)$ at given x^*.

The confidence interval for the mean response at x^* is given by

$$\widehat{Y}^* \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}. \tag{10.5}$$

The prediction interval and the confidence intervals are valid as long as x^* is within the range of x-values.

Example 10.3 continued. For these data, we already computed

$$s_x = 1.0112,\ s_e = 0.1418 \quad \bar{x} = 3.2556.$$

Suppose we like to find a 95% prediction interval for salinity level when $x^* = 4.0$. We see in Table 10.1 that x values range from 1.6 to 4.8, and $x^* = 4.0$ is within this range, and we can use the preceding methods to find prediction interval and confidence interval. Then, noting that

$$\widehat{Y}^* = -4.073 + 2.569 \times 4 = 6.203, \quad (n-1)s_x^2 = 17 \times 1.0112^2 = 17.38$$

and the percentile

$$t_{16;.975} = \text{invT}(.975, 16) = 2.1199,$$

we find the 95% prediction interval as

$$\begin{aligned} \widehat{Y}^* \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}, &= 6.203 \pm 2.1199 \times 0.1418 \times \sqrt{1 + \frac{1}{18} + \frac{(4.0 - 3.2556)^2}{17.38}} \\ &= 6.203 \pm .313 \\ &= (5.89, 6.52). \end{aligned}$$

Thus, when the electric conductivity is 4.0, the salinity level will be somewhere between 5.89 and 6.52 with confidence 95%.

The 95% confidence interval for the mean of Y at $x^* = 4.0$ is

$$6.203 \pm 2.1199 \times 0.1418 \times \sqrt{\frac{1}{18} + \frac{(4.0 - 3.2556)^2}{17.38}} = 6.203 \pm .089 = (6.114, 6.292).$$

Thus, the mean salinity level at $x^* = 4$ is between 6.114 and 6.292 with confidence 0.95.

■ **Example 10.4** The following data represent the number of manatee deaths x, and the number of registered Florida pleasure crafts (in 10,000).

It is desired to find the linear relationship between the number of registered boats per year and the number of manatee deaths per year.

Table 10.2: Registered Florida Pleasure Craft (in 10,000) and Watercraft Related Manatee Deaths

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
No. of Boats (x)	68	68	67	70	71	73	76	81	83	84
No. of Manatee Deaths (Y)	53	38	35	49	42	60	54	67	82	78

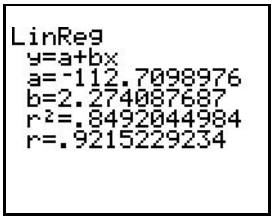
a. Construct a scatter plot and identify the relationship between Y and x.
b. Find the least square estimates of the parameters in the model $Y = a + bX + \varepsilon$.
c. Find the value of coefficient of determinant r^2. On the basis of r^2, can we conclude that the model is well fitted?
d. Interpret the meanings of the estimated slope $\widehat{b}$.
e. For testing $H_0 : b \leq 0$ vs. $H_a : b > 0$, calculate the test statistic and the p-value.
f. Find a 95% prediction interval for the number of manatee deaths when the number of boat registrations is 750,000.
g. Find a 95% confidence interval for the average number of manatee deaths when the number of boat registrations is 750,000.

Solution:

a. Enter the data in two lists, say, x in L5 and Y in L6. Select [2nd], [STAT PLOT], and select the options as shown in the following first box. Press [ZOOM] and [9: ZOOM STAT] to get the results as shown in the second box.

The scatter plot indicates that Y and X have positive linear relation.

b. Press [STAT], [CALC] and select [8: LinReg(a+bX)]; select the list L5, comma, L6. Press [ENTER] to get

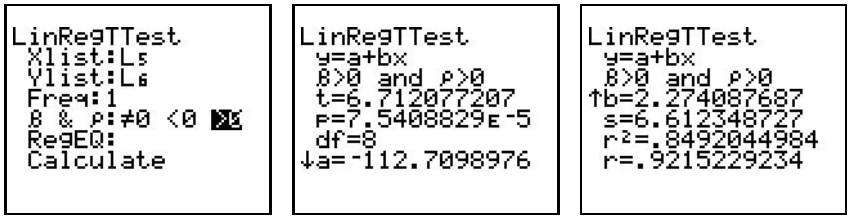


Thus, the least squares estimates are $\widehat{a} = -112.71$ and $\widehat{b} = 2.2741$.

c. The value of r^2 is .85, which is not far away from 1. We can say that the model is fairly well fitted.

d. Note that $\widehat{Y} = -112.71 + 2.27x$, where x is the number of registered boats (in 10,000). Thus, for every 10,000 increase boat registration, the number of deaths increases by 2.27.

e. Press [STAT], [TESTS], [F: LinRegTTest], and select the options as shown below.



Note that the test statistic is 6.7120 and the p-value is .00008. Since the p-value is very small, we can conclude that the slope is positive.

f. The 95% prediction interval is

$$\widehat{Y}^* \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}}.$$

The $s_e = 6.6123$ (see the third box). For x-data,

$$\bar{x} = 74.1,\ s_x^2 = 6.5056^2 = 42.3228 \quad \text{and } (n-1)s_x^2 = 9 \times 42.3228 = 380.91.$$

Noting that one unit of x represents 10,000 boats, we find

$$x^* = 750000/10000 = 75, \text{ and } \widehat{Y}^* = -112.71 + 2.2741 \times 75 = 57.85$$

The t critical value to find the 95% prediction interval is

$$t_{n-2;.975} = t_{8;.975} = \text{invT}(.975, 8) = 2.306.$$

Using these numbers in the above formula, we find the 95% prediction interval as

$$57.85 \pm 2.306 \times 6.6123 \times \sqrt{1 + \frac{1}{10} + \frac{(75-74.1)^2}{380.91}} = 57.85 \pm 16.0 = (41.85, 73.85).$$

We predict 42 to 74 manatee deaths when 750,000 boats are registered

g. The 95% confidence interval is

$$\widehat{Y}^* \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}.$$

We already calculated all the required quantities in the above formula in part **e.** Substituting these number in the preceding formula, we find

$$57.85 \pm 2.306 \times 6.6123 \times \sqrt{\frac{1}{10} + \frac{(75-74.1)^2}{380.91}} = 57.85 \pm 4.87 = (52.92, 62.72).$$

When 750,000 boats are registered, then we expect 53 to 63 manatee deaths.

■

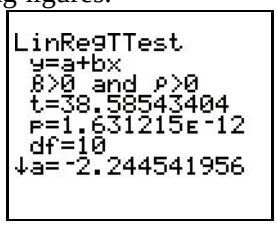
■ **Example 10.5** The following are the scores of 12 students in a math class. The midterm scores are the average of three midterm tests. The final scores are the scores (out of 100) in the final exam.

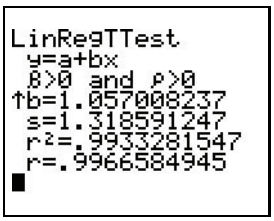
Table 10.3: Midterm and final scores of 12 students in a math class

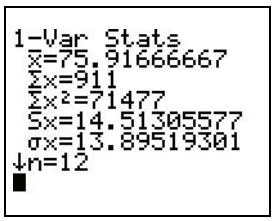
student	1	2	3	4	5	6	7	8	9	10	11	12
midterm	55	67	69	70	90	95	54	88	86	94	66	77
final	55	69	71	72	94	98	55	89	89	96	66	82

a. Construct scatter plot and identify the relationship between midterm scores and final scores.
b. Find the least square estimates of a and b.
c. Interpret the meanings of $\widehat{b}$ in the context of the fitted model.
d. For testing $H_0 : b \leq 0$ vs. $H_a : b > 0$, calculate the test statistic and the p-value, and write the conclusion.
e. Does r^2 indicate the model is well fitted? Explain.
f. What percentage of variation in final scores explained by the model prediction?
g. Find a 95% prediction interval for the final score of a student whose midterm score is 85.
h. Find a 95% confidence interval for the expected score of a student whose midterm score is 85.

Solution: To answer all the questions, it is convenient to carry out all calculations using TI-84 as shown in the following figures.







a. The scatter plot indicates that Y and x have positive linear relation.

b. The least squares estimates are $\widehat{a} = -2.2445$ and $\widehat{b} = 1.0570$.

c. The fitted model is $\widehat{Y} = -2.2445 + 1.0570X$. The estimate $\widehat{b} = 1.0570$ implies that for every one point increase in the average midterm score, the score in the final exam increases by 1.06 point.

d. The test statistic is $t = 38.59$ with the p-value of zero. So the data provide strong evidence to indicate that the slope is positive.

e. Since $r^2 = .993$ is very close to 1, we can say that the model is very well fitted.

f. The coefficient determinant $r^2 = .993$. This means that 99.3% variation in the final score is explained by the average midterm score.

g. The necessary quantities for finding the 95% PI are:

$\bar{x} = 75.92$, $s_e = 1.3186$, $(n-1)s_x^2 = 11 \times 14.513^2 = 2316.9$ $\widehat{Y}^* = -2.2445 + 1.0570 \times 85 = 87.6$,

and

$$t_{10;.975} = \text{invT}(.975, 10) = 2.228.$$

Using the formula for the PI, we find

$$87.6 \pm 2.228 \times 1.3186 \times \sqrt{1 + \frac{1}{12} + \frac{(85 - 75.92)^2}{2316.9}} = 87.6 \pm 3.1 = (84.5, 90.7).$$

Thus, we predict the final score of a student with midterm score 85 is between 85 and 91 with 95% confidence.

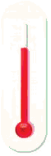
h. The 95% CI for the mean score when midterm score is 85:

$$87.6 \pm 2.228 \times 1.3186 \times \sqrt{\frac{1}{12} + \frac{(85 - 75.92)^2}{2316.9}} = 87.6 \pm 1 = (86.6, 88.6).$$

The mean score of all students with midterm score 85 is between 87 and 89 with confidence 95%.

■

■ **Example 10.6** (The cricket as a thermometer!) Dolbear[2] approximated the relationship between the number of chirps (X) per minute by snowy tree cricket and the outside temperature (Y) in Fahrenheit.



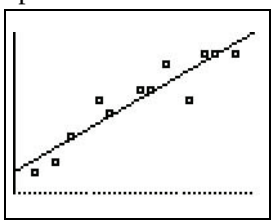
The relationship is $Y = 40 + \frac{1}{4}X$. This relation also holds approximately for field crickets. A biologist has decided to check on the formula and collected the following data on the number of chirps per minute by field crickets and the outside temperatures in Fahrenheit.

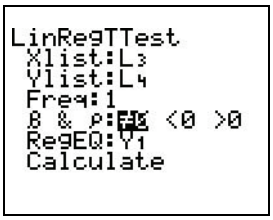
Number of chirps per minute and outside temperature												
No. of chirps per minute (X) :	122	126	129	134	136	142	144	147	151	154	156	160
Temperature (Y):	68	69	71	74	73	75	75	77	74	78	78	78

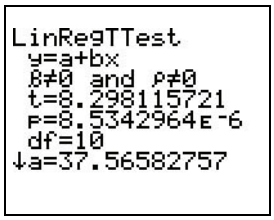
[2]Dolbear, A. (1897). The cricket as a thermometer. *The American Naturalist 31: 970-971*

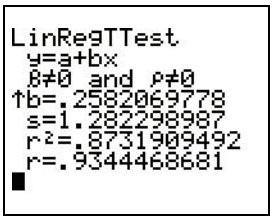
Fit a linear regression model using the above data. Is the model in agreement with the Dolbear's formula?

a. Construct scatter plot and identify the relationship between outside temperature and the number of chirps per minutes.
b. Find the least square estimates of a and b.
c. Interpret the meanings of $\widehat{b}$ in the context of the fitted model.
d. For testing $H_0 : b = .25$ vs. $H_a : b \neq .25$, calculate the test statistic and the p-value, and write the conclusion.
e. Does r^2 indicate the model is well fitted? Explain.
f. What percentage of variation in outside temperatures explained by the model prediction?
g. Find a 95% prediction interval for the outside temperature when the number of chirps per minute is score is 131.
h. Find a 95% confidence interval for the expected outside temperature when the number of chirps per minute is 131.









Solution:

a. The scatter plot above clearly indicates that a positive liner relationship exists between Y and X.
b. The least square estimates of a and b are

$$\widehat{a} = 37.6 \quad \text{and} \quad \widehat{b} = .26$$

c. $\widehat{b} = .26$. This means that for every one chirp increase in the outside temperature increase by .26°F.
d. Notice that the above result from TI-84 can be used to test only

$$H_0 : b = 0 \quad \text{vs.} \quad H_a : b \neq 0.$$

However, for the present problem it is desired to test

$$H_0 : b = .25 \quad \text{vs.} \quad H_a : b \neq .25.$$

The test statistic in this case is

$$t = \sqrt{n-1}\frac{(\widehat{b} - .25)s_x}{s_e}.$$

To calculate the test statistic, the standard deviation of the x-data is $s_x = 15.39$. Thus

$$t = \sqrt{11}\frac{(.26 - .25) \times 15.39}{1.2822} = 0.3981,$$

where s_e is from the fourth box TI-84 results. The p-value is

$$\begin{aligned} P(t_{10} < -.3981) + P(t_{10} > .3981) &= \text{tcdf}(-10^7, -.3981, 10) + \text{tcdf}(.3981, 10^7, 10) \\ &= .3495 + .3495 \\ &= .699. \end{aligned}$$

Since this p-value is quite large, we do not have enough evidence to conclude that the slope parameter is different from 0.25. In other words, the $H_0 : b = .25$ is plausible, and fitted model is $\widehat{Y} = 37.6 + 0.26X$ is not significantly different from the Dolbear's formula $Y = 40 + .25X$.

e. Since the value of $r^2 = .87$ is not far away from 1, we can say that the model is well fitted.

f. Since $r^2 = .87$, we can say about 87% of the variation in outside temperatures is explained by the model.

g. To find the prediction interval, note that

$$n = 12,\ \bar{x} = 141.75,\ s_x = 12.4252,\ (n-1)s_x^2 = 11 \times 12.4252^2 = 1698.25 \text{ and } s_e = 1.2823.$$

The critical value is $t_{n-2;.975} = t_{10;.975} = \text{invT}(.975, 10) = 2.228$. Furthermore,

$$\widehat{Y} = 37.57 + .26 \times 131 = 71.6.$$

So the 95% prediction interval at $x^* = 131$ is

$$\begin{aligned} \widehat{Y} \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}} &= 71.6 \pm 2.228 \times 1.282 \sqrt{1 + \frac{1}{12} + \frac{(131 - 141.75)^2}{1698.25}} \\ &= 71.6 \pm 3.1 \\ &= (68.5,\ 74.7) \end{aligned}$$

Thus, when the number of chirps per minutes is 131, the predict the outside temperature to be between 69 and 75°F.

h. To find the 95% confidence interval for the mean outside temperature when $x^* = 131$, we can use the calculation above as

$$\begin{aligned} \widehat{Y} \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}} &= 71.6 \pm 2.228 \times 1.282 \sqrt{\frac{1}{12} + \frac{(131 - 141.75)^2}{1698.25}} \\ &= 71.6 \pm 1.1 \\ &= (70.5,\ 72.7) \end{aligned}$$

Thus, when the number of chirps per minutes is 131, the average outside temperature will be between about 71 and 73°F.

■

■ **Example 10.7** The following data represent the list price x (in \$1,000) and the best price of 23 GMC pickup trucks.

No.	List price, x	Best price, Y	No.	List price, x	Best price, Y
1	12.4	11.2	13	22.4	19.6
2	14.3	12.5	14	19.4	16.9
3	14.5	12.7	15	15.5	14.0
4	14.9	13.1	16	16.7	14.6
5	16.1	14.1	17	17.3	15.1
6	16.9	14.8	18	18.4	16.1
7	16.5	14.4	19	19.2	16.8
8	15.4	13.4	20	17.4	15.2
9	17.0	14.9	21	19.5	17.0
10	17.9	15.6	22	19.7	17.2
11	18.8	16.4	23	21.2	18.6
12	20.3	17.7			

a. Construct scatter plot and identify the relationship between Y and X.

b. Write the fitted model and the value of r^2.

c. Is the model well fitted? Explain.

d. Interpret the meanings of $\widehat{b}$ in the context of the fitted model.

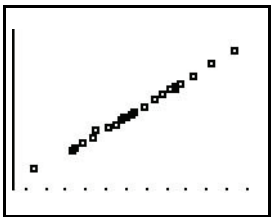
e. For testing $H_0 : b = 0$ vs. $H_a : b \neq 0$, calculate the test statistic and the p-value.

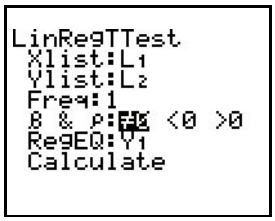
f. Predict the best price of a GMC pickup truck with the list price of \$19,500.

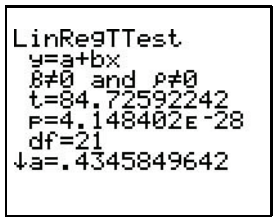
g. Find a 95% prediction interval for the best price of a GMC pickup truck with the list price of $19,500.

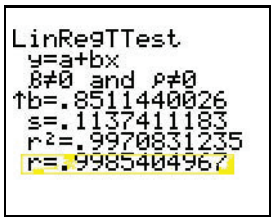
h. Find a 95% confidence interval for the expected best price of a GMC truck with the list price of $19,500.

Solution:









a. The scatter plot indicates that there is a positive linear relation between the best price and the list price.

b. The fitted model is $\widehat{Y} = .4346 + .8511x$. The r^2 is .997.

c. Since the value of r^2 is very close to 1, the model is well fitted. The fitted model can be used to predict the best price for values of x within the range of all x values.

d. From the fourth figure, we see that $\widehat{b} = 0.851$. Recall that each unit in x and Y is equal to $1,000. For one unit increase in x, Y increases by .851. In other words for every $1,000 increase in the list price, the best price increases by $851.

e. For testing $H_0 : b = 0$ vs. $H_a : b \neq 0$, the value of the t-statistic is 84.73 with the p-value of zero. So the slope parameter is significantly different from zero.

f. Noting that one unit of x represents 1,000 dollars, we find

$$x^* = 19500/1000 = 19.5,$$

and

$$\begin{aligned} \widehat{Y}^* &= \widehat{a} + \widehat{b}x^* \\ &= .435 + .851 \times 19.5 \\ &= 17.030. \end{aligned}$$

Thus, the predicted best price is $17,030.

g. The 95% prediction interval is

$$\widehat{Y}^* \pm t_{n-2;1-\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}.$$

From the fourth figure, we find $s_e = .1137$. For x-data,

$$\bar{x} = 17.465,\ s_x^2 = 2.4139^2 = 5.8269 \quad \text{and } (n-1)s_x^2 = 22 \times 5.8269 = 128.19.$$

The t critical value to find the 95% prediction interval is

$$t_{n-2;.975} = t_{21;.975} = \text{invT}(.975, 21) = 2.080.$$

Using these numbers in the above formula, we find the 95% prediction interval as

$$17.030 \pm 2.080 \times .1137 \times \sqrt{1 + \frac{1}{23} + \frac{(19.5 - 17.465)^2}{128.19}} \quad = \quad 17.030 \pm .254 = (16.776, 17.284).$$

Thus, the predicted best price of a GMC truck with the list price $19,500 is between $16,776 and $17,284.

h. The 95% confidence interval for the expected best price when the list price is \$19,500 is

$$17.030 \pm 2.080 \times .1137 \times \sqrt{\frac{1}{23} + \frac{(19.5 - 17.465)^2}{128.19}} \quad = \quad 17.030 \pm .065 = (16.965, 17.095).$$

The average best price of a truck with list price of \$19,500 is between \$16,965 and \$17,095 with confidence 95%.

■

10.6 Exercises

1. The following data represent the living area x (in 100 square ft), and the selling price Y (in \$100) for a sample of 15 houses from a new subdivision in a town. For example, the selling price of the House 1 with 2,790 ft^2 is \$457,200.

House	area (x)	Price, Y	House	area (x)	Price, Y
1	27.90	4572	11	26.12	4305
2	24.30	4000	12	29.22	4750
3	27.45	4492	13	25.12	4114
4	25.60	4198	14	24.37	3995
5	27.70	4527	15	26.66	4359
6	24.20	4000			
7	25.50	4238			
8	24.30	3984			
9	23.40	3825			
10	24.26	3993			

   a. Construct a scatter plot and explain the relationship between Y and X.
   b. Write the fitted model and the value of r^2.
   c. On the basis of r^2, can we say that the model is well fitted? Explain.
   d. Interpret the meanings of $\widehat{b}$ in the context of the problem.
   e. Predict the selling price of a house in the neighborhood with the living area of 2,700 square feet (i.e., $x^* = 27$).
   f. By comparing the above predicted price (in part **e**) with the prices of the House 3 (2,750 ft^2) and House 15 (2,666 ft^2), can we say that the predicted price is reasonable? Explain.
   g. Find a 95% prediction interval of a house in the neighborhood with the living area of 2,700 square feet.
   h. Find a 95% confidence interval for the mean price of all houses in the subdivision with the living area of 2,700 ft^2, and interpret its meanings.

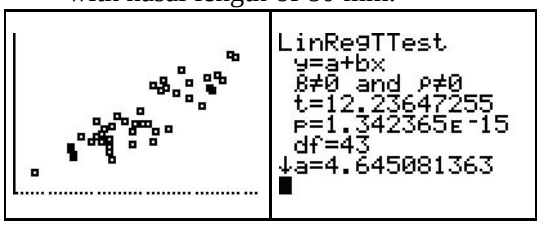
2. The data[3] in Table 10.4 are nasal lengths and nasal widths that were collected from a sample of $n = 45$ male gray kangaroos.
   a. Construct scatter plot and examine if there is any linear relationship between Y and X.
   b. Write the fitted model and the value of r^2.
   c. Is the model well fitted? Explain.
   d. Interpret the meanings of $\widehat{b}$ in the context of the problem.
   e. Is the slope parameter of the model significantly different from zero? Test using $\alpha = .05$.
   f. Predict the nasal width of a male gray kangaroo with nasal length of 80 mm.
   g. Find a 95% prediction interval of a male gray kangaroo with nasal length of 80 mm.

[3] Australian Journal of Zoology, Vol. 28, p607-613

Table 10.4: Nasal length (x, in mm) and nasal width (Y, in mm) of a sample of male gray kangaroos

No.	x	Y	No.	x	Y	No.	x	Y
1	60.9	24.1	16	83.0	28.8	31	56.2	21.6
2	62.9	22.2	17	86.4	30.6	32	58.0	22.5
3	62.0	23.3	18	63.5	23.6	33	59.6	22.0
4	56.4	20.7	19	56.5	20.4	34	59.7	21.9
5	64.5	24.7	20	82.3	27.2	35	63.6	20.1
6	49.3	18.9	21	75.5	26.8	36	55.9	21.3
7	60.6	22.6	22	71.0	27.8	37	61.5	22.8
8	66.0	24.0	23	70.1	23.8	38	69.2	23.8
9	63.0	21.5	24	80.3	25.5	39	71.0	22.1
10	67.2	23.1	25	85.5	30.8	40	73.0	28.1
11	77.8	26.3	26	83.8	28.1	41	76.3	29.2
12	61.6	22.0	27	74.0	23.4	42	68.6	25.1
13	72.7	27.1	28	67.7	23.7	43	71.7	23.1
14	81.0	28.4	29	67.5	21.7	44	73.7	27.5
15	77.8	27.9	30	62.9	21.1	45	81.6	27.5

h. Find a 95% confidence interval for the mean nasal width of all male gray kangaroos with nasal length of 80 mm.



3. The data in Table 10.5 represent the national unemployment rate for adult males x, and that of adult females Y.

Table 10.5: National unemployment rates for adult males and females

No.	males, x	females, Y
1	2.9	4.0
2	6.7	7.4
3	4.9	5.0
4	7.9	7.2
5	9.8	7.9
6	6.9	6.1
7	6.1	6.0
8	6.2	5.8
9	6.0	5.2
10	5.1	4.2
11	4.7	4.0
12	4.4	4.4
13	5.8	5.2

a. Construct a scatter plot for the data and interpret its meanings.
b. What is the value of the correlation coefficient(r) for this data set? Interpret this value in the context.
c. Find the equation of the least squares regression line.
d. What is the value of the coefficient of determination(r^2) for this data set? Interpret this

value in the context.

e. Find a 95% prediction interval for

4. Several studies reported that there is a positive correlation between heights and salaries of people in sales and management positions. Heights and salaries of 28 people from sales and management positions are given in the Table 10.6.

Table 10.6: Heights and salaries of a sample of 28 people

Height(x, inches)	Salary(Y, dollars)	Height(x, inches)	Salary(Y, dollars)
64.2	5780	69.1	6110
65	5500	70.0	5970
65	5590	70.0	6090
66.1	5320	70.0	6270
66.1	5680	70.0	6380
67.1	5730	71.1	6380
67.1	5890	71.1	6660
68.2	5700	72.9	6320
68.2	5800	73.0	6410
68.2	5910	73.2	6430
69.1	5850	73.2	6450
69.4	5890	73.9	6480
69.4	5930	75.2	6270
69.6	6000	76.1	6410

a. Construct a scatter plot for theses data and interpret its meanings.

b. Find the equation of the least squares regression line.

c. What is the value of the correlation coefficient(r) for this data set? Interpret this value in the context.

d. What is the value of the coefficient of determination(r^2) for this data set? Interpret this value in the context.

e. Would it be reasonable to use the least squares regression line in part **b** to predict the salary for a person's height of 79.5 inches? Explain why or why not?

5. The tensile strength of a cable for upper-limb prosthesis was investigated. Stainless steel cable is commonly available in three sizes (diameters): 1.19 mm, 1.59 mm and 2.38 mm. Four tests were performed for each diameter size and the results are given in Table 11.3. Let X be cable cross area (mm^2) and Y be tensile strength (KN)

Table 10.7: Cable cross area and tensile strength

Cable diameter (mm)	Cable cross area (mm^2)	Tensile strength (KN)
1.19	1.1122	1.27
1.19	1.1122	1.45
1.19	1.1122	1.43
1.19	1.1122	1.36
1.59	1.9856	2.20
1.59	1.9856	2.56
1.59	1.9856	2.38
1.59	1.9856	2.45
2.38	4.4488	4.58
2.38	4.4488	5.03
2.38	4.4488	5.67
2.38	4.4488	4.39

a. Construct a scatter diagram to illustrate these results.
b. Calculate the correlation coefficient for the data and comment on the result.
c. Obtain the least squares estimates slope and intercept parameters of the regression model of Y and X.
d. Find a 95% prediction interval for the tensile strength of a cable with cross area 3.5.
e. Find a 95% confidence interval for the mean tensile strength of cables with cross area 3.5.
f. Comment on the validity of the fitted regression equation to estimate the tensile strength for cable with cross area 6.

6. American Community Survey reported the following data on earnings in 2009 and average travel time to work for workers in the USA.

Table 10.8: Workers Earnings in 2009 and average commuting time (in minutes) to work

Earnings	mid-point	mean travel time
Less than $10,000	5,000	20.4
$10,000 to $14,999	12,500	22.1
$15,000 to $24,999	20,000	23.2
$25,000 to $34,999	30,000	24.6
$35,000 to $49,999	42,500	26.2
$50,000 to $64,999	57,500	27.8
$60,000 to $74,999	67,500	29.0
$75,000 to $99,999	87,500	30.3
$100,000 or more	100,000	30.5

11. The Chi-Square Test for Association

The χ^2-test is used to test equality of two or more than two proportions, and to test if there is association among several qualitative variables. For testing two proportions are equal, we could use the two-sample z-test or the chi-square test and both tests produce the same result as shown in the following example.

■ **Example 11.1** In 2000 the Vermont State legislature approved a bill authorizing civil unions. The vote can be broken down by gender to produce the following table. The problem of interest here is to test if voting behavior depending on the gender.

	Vote		
	Yes	No	Total
Women	35	9	44
Men	60	41	101
Total	95	50	145

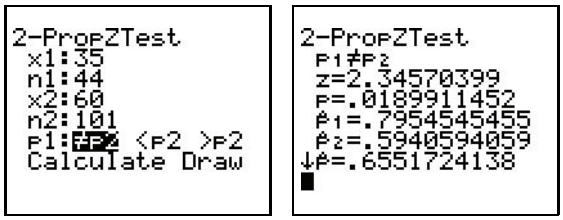
Let p_m denote the proportion of men voted "yes" and p_f denote the same among women. Suppose the voting behavior and the gender are independent (or no association between voting behavior and gender), then $p_m = p_f$. Thus, the hypotheses of interest are

$$H_0 : p_m = p_f \quad \text{vs.} \quad H_a : p_m \neq p_f,$$

where H_0 implies that there is no association between the voting behavior and gender, and H_a implies that there is some association. To test the above hypothesis, we can use the two-sample z-test as follows. Note that the data can be written as

$$X_w = 35, \quad n_w = 44, \quad X_m = 60 \quad \text{and} \quad n_m = 101,$$

where X_m denotes the number of "yes" votes among $n_m = 101$ men, and X_w denotes the number of "yes" votes among $n_w = 44$ women. Using [2 PropZTest] in TI-84, we obtain



The Z statistic is 2.3457 with the p-value .01899. Thus, we have sufficient evidence to conclude that there is association between gender and voting behavior.

The χ^2 Test

To apply the χ^2-test, we regard the data in the table as "observed frequencies" and calculate the "expected frequencies" for each cell under the null hypothesis that

"voting choice and gender are independent."

The observed frequency in the (i, j)th cell is denoted by O_{ij}. For this problem,

$$O_{11} = 35,\ O_{12} = 9,\ O_{21} = 60 \text{ and } O_{22} = 41.$$

The expected frequency in the (i, j)th cell is denoted by E_{ij}, and it can be calculated as follows.

- There are 95 "yes" votes out of 145 people.
- The proportion of women in all voters is $\frac{44}{145}$.
- If there is no association between the gender and voting behavior, the expected number of "yes" votes for women should be

$$\begin{aligned} \text{total number of "yes" votes} &\times \text{proportion of women among all voters} \\ &= 95 \times \frac{44}{145} \\ &= \frac{95 \times 44}{145} \\ &= 28.83 \\ &= \text{expected frequency in (1,1) cell, denoted by } E_{11}. \end{aligned}$$

- If there is no association between the gender and voting behavior, the expected number of "yes" votes for men should be

$$\begin{aligned} \text{total number of "yes" votes} &\times \text{proportion of men among all voters} \\ &= 95 \times \frac{101}{145} \\ &= \frac{95 \times 101}{145} \\ &= 66.17 \\ &= \text{expected frequency in (2,1) cell, denoted by } E_{21}. \end{aligned}$$

Notice that

$$\begin{aligned} E_{11} &= \frac{\text{the first row total} \times \text{the first column total}}{\text{Grand Total}} \\ &= \frac{44 \times 95}{100} = 28.83, \end{aligned}$$

and

$$E_{21} = \frac{\text{the 2nd row total} \times \text{the first column total}}{\text{Grand Total}} = \frac{101 \times 95}{100} = 66.17.$$

Similarly, we can compute the expected frequencies in other cells as

$$E_{12} = \frac{44 \times 50}{145} = 15.17 \quad \text{and} \quad E_{22} = \frac{101 \times 50}{145} = 34.83.$$

Notice that

$$\begin{aligned}\text{sum of the expected frequencies} &= 28.83+66.17+15.17+34.83\\ &= 145\\ &= \text{sum of the observed frequencies}.\end{aligned}$$

Description of the Chi-square Test

The χ^2-test is used to see if there is any association among several categorical variables such as eye color and hair color or nationality and eye color. The data are collected from a sample of n individuals, and the response of each individual is classified into one of the $R \times C$ cells as shown below.

	Attribute 1				
Attribute 2	Level 1	Level 2	...	Level c	Total
Level 1	O_{11}	O_{12}	...	O_{1c}	R_1
Level 2	O_{21}	O_{22}	...	O_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
Level r	O_{11}	O_{12}	...	O_{1c}	R_r
Total	C_1	C_2	...	C_c	Grand Total

In general,

$$\text{the expected frequency in the } (i,j)\text{th cell} = E_{ij} = \frac{\text{the } i\text{th row total} \times \text{the } j\text{th column total}}{\text{Grand Total}},$$

Let r denote the number of rows and c denote the number of columns. The chi-square statistic, denoted by χ^2-stat, is calculated as

$$\chi^2\text{−stat} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

which has an approximate chi-square distribution with degrees of freedom $r-1 \times c-1$. The p-value of the test is given by

$$P(\chi^2_{(r-1)\times(c-1)} > \chi^2\text{−stat}) = \chi^2\text{cdf}\left(\chi^2\text{−stat}, 10^7, (r-1)\times(c-1)\right),$$

where χ^2cdf(., .,) is the function in TI-84, which can be accessed by [2nd], [DISTR],[8: χ^2cdf].

Example 11.1 continued. Returning back to the example, we shall rewrite the table with the expected frequencies (in the square brackets) as shown in the following table:

	Vote		
	Yes	No	Total
Women	35[28.83]	9[15.17]	44
Men	60[66.17]	41[34.83]	101
Total	95	50	145

Thus, the chi-square statistic is

$$\chi^2-\text{stat} = \frac{(35-28.83)^2}{28.83} + \frac{(9-15.17)^2}{15.17} + \frac{(60-66.17)^2}{66.17} + \frac{(41-34.83)^2}{34.83} = 5.5025.$$

To find the p-value, the degrees of freedom for the chi-square distribution is

$$(2-1)\times(2-1) = 1.$$

The p-value is

$$P(\chi_1^2 > 5.5025) = \chi^2\text{cdf}(5.5025, 10^7, 1) = \mathbf{0.01899},$$

the same p-value that we obtained using the two-proportion z-test. ■

Remark

Two-proportion z test vs. Chisquare test: If there are only two groups with a single variable of interest, then the association between the groups and the variable can be tested either using two-proportion z-test or the χ^2-test. If there are two groups with more than one variables of interest, or three or more groups, then only the χ^2-test can be used as shown in the following example.

■ **Example 11.2** A sample 593 people was classified on the basis of their hair colors and eye colors as shown in Table 11.1 This table includes 4 rows for hair color and 4 columns for eye color. The data are referred to as the **observed frequencies**. The hypotheses of interest here are

Table 11.1: Observed frequencies on eye colors and hair colors

Hair Color	**Eye Color**				
	Brown	Blue	Hazel	Green	Total
Black	68[40.25]	20[39.16]	15[16.94]	5[11.66]	108
Brown	120[106.96]]	84[104.06]	54[45.01]	29[30.97]	287
Red	26[26.46]	17[25.74]	14[11.13]	14[7.66]	71
Blond	7[47.33]	94[46.05]	10[19.92]	16[13.71]	127
Total	221	215	93	64	593

$$H_0: \text{There is no association between eye and hair colors} \quad \text{vs.} \quad H_a: \text{some association},$$

equivalently,

$$H_0: \text{eye and hari colors are indepedent} \quad \text{vs.} \quad H_a: \text{they are dependent}.$$

Denoting the ith row total by R_i and jth column total by C_j, we find the expected frequency as

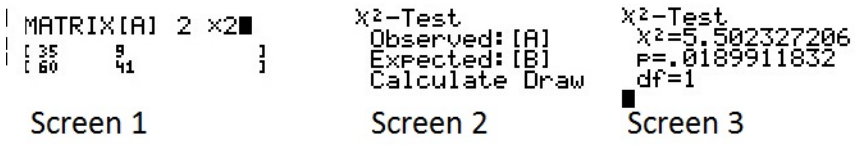
$$E_{ij} = \frac{R_i \times C_j}{\text{Grand Total}}.$$

The χ^2 statistic and the p-value can be calculated using TI-84 as shown in the next page. For this problem,

$$\chi^2-\text{stat} = 138.5 \quad \text{with p-value} = 0.$$

Thus, the data provide strong evidence to indicate there is association between eye colors and hair colors. ■

■ **Example 11.3** (Example 11.1 continued.) We shall find the χ^2 statistic and the p-value using TI-84:



Screen 1 Screen 2 Screen 3

Notice that the χ^2 statistic is 5.5023 and the p-value is 0.01899, and they are in agreement with those we obtained earlier. ■

In the following, we use TI-84 to calculate the p-value for Example 11.1. Also, notice that TI-84 calculates expected frequencies, and you DO NOT have to enter expected frequencies. After entering the observed frequencies in "Matrix A", press [STAT], [TESTS], [χ^2 test], and [Calculate] to get the p-value.

Calculation of χ^2 statistic and p-value using TI-84	
Press [2nd], [MATRIX], [EDIT] and select the matrix [A] Enter the order of the matrix 2 x 2 to indicate that there are 2 rows and 2 columns	NAMES MATH EDIT 1:[A] 2×1 2:[B] 3:[C] 4:[D] 5:[E] 6:[F] 7↓[G]
Enter the numbers in the 1st row, 2nd row ...	MATRIX[A] 2 ×2 [35 8] [60 41] 2,2=41
Press [2nd], [TESTS], [χ^2-Test] Select [Calculate], press [ENTER]	χ²-Test Observed:[A] Expected:[B] Calculate Draw
The χ^2 statistic is 138.49 The p-value is 2.11484E-25 = 0 df = $(4-1)\times(4-1)=9$	χ²-Test χ²=5.502327206 p=.0189911832 df=1

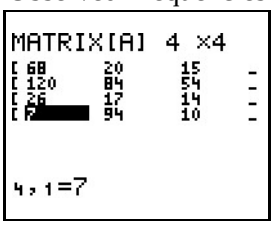
TI Calc

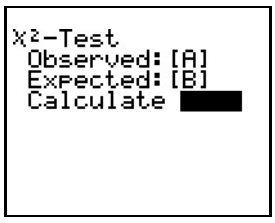
Deleting a matrix from the memory

1. Press the [2nd] key and the [+] key on the TI-84
2. Scroll to [Mem Mgmt/Del]
3. Press the [ENTER] key
4. Press [5] to select [Matrix] and press the [ENTER] key
5. Scroll to each matrix and press [DEL]. This will clear the matrix from the memory.

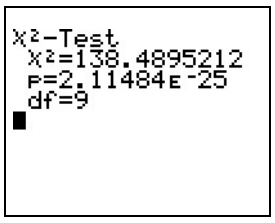
Example 11.2 continued. The p-value for this example can be calculated (using TI-84) as follows. Enter the observed frequencies in matrix A, select [χ^2 test], and select [Calculate].

Observed Frequencies





Results



■ **Example 11.4** The Mediterranean diet, rich in vegetables, fruits, and grains, is healthier can lower risk of heart disease. To compare the Mediterranean diet with the low-fat diet recommended by the American Heart Association (AHA), 605 survivors of a heart attack, were randomly assigned to the AHA diet and a Mediterranean-type diet. The researchers collected information on number of deaths from cardiovascular causes by heart attack, strokes, as well as nonfatal heart-related episodes. The results[1] are given in the following table. Test if types of diet and outcomes are associated at the level 0.05.

Table 11.2: Observed frequencies on diet and outcomes

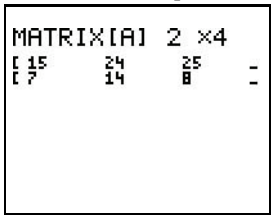
Diet	**Outcomes**				
	Cancer	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15[11.02]	24[19.03]	25[16.53]	239[256.42]	303
Mediterranean	7[10.98]	14[18.97]	8[16.47]	273[255.58]	302
Total	22	38	33	512	605

Solution:

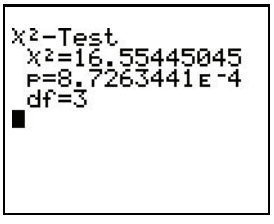
The hypotheses of interest are

H_0 : types of diet and health outcomes are independent vs. H_a : they are dependent.

Observed Frequencies



Results



[1] De Longerill, M., Salen, P., Martin, J., Monjaud, I., Boucher, P., Mamelle, N. (1998). Mediterranean Dietary pattern in a Randomized Trial. Archives of Internal Medicine, 158, 1181-1187.

The χ^2 statistic is 16.55, and the p-value is .0009. Since the p-value is very small, we reject the null hypothesis, and conclude that types of diet and outcomes are dependent. ■

■ **Example 11.5** Nationality and eye color for a sample of 4,848 people from European countries are given in the following table. Using the data, test if there is any association between nationality and eye color.

	Nationality					
Eye Color	English	German	Irish	Italian	French	Total
Blue	709	539	185	22	145	1600
Brown	661	462	145	145	229	1642
Green	335	222	86	19	55	717
Hazel	346	256	66	39	94	801
Other	45	13	13	6	11	88
Total	2096	1492	495	231	534	4848

Solution: The hypotheses of interest are

$$H_0 : \text{no association between eye color and nationality vs. } H_a : \text{some association}$$

The

$$\chi^2 - \text{statistic} = 165.17 \quad \text{with the p-value of } \ 7.711...10^{-27} = 0.$$

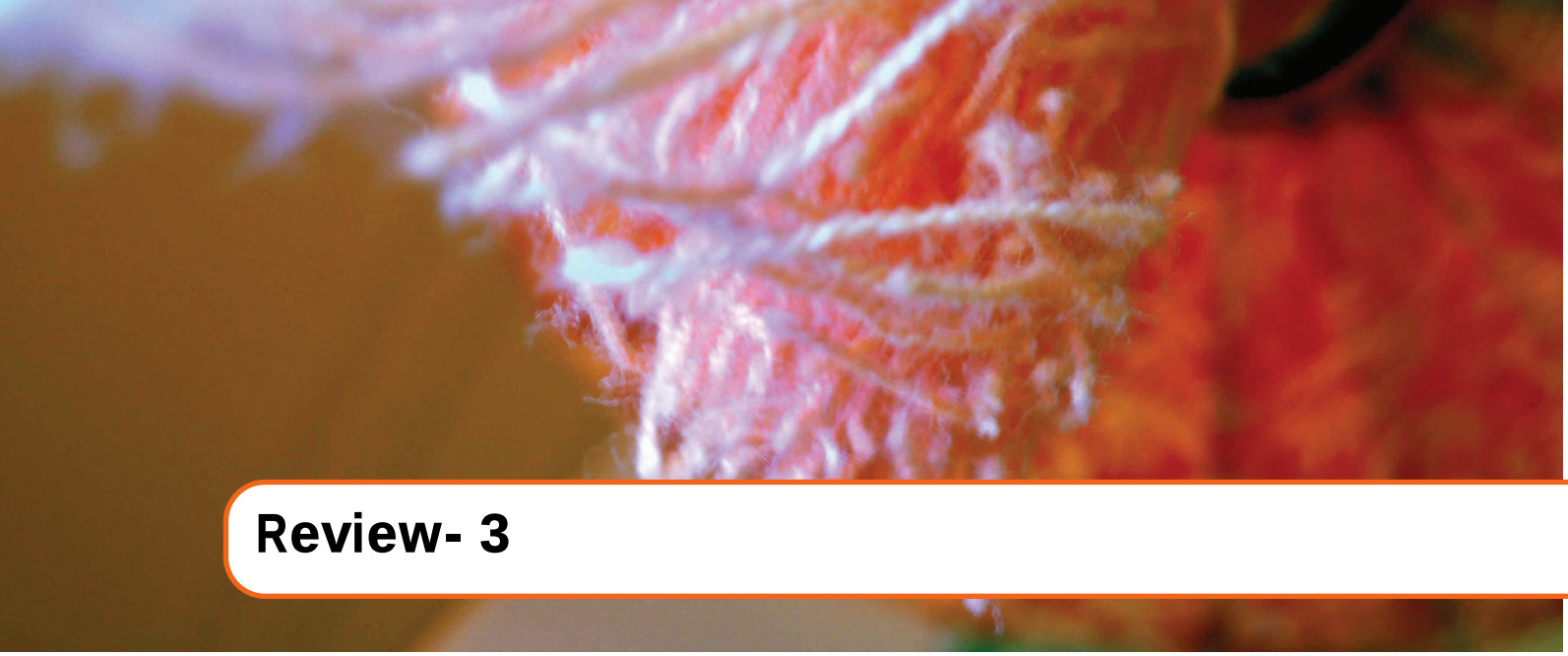
■

11.1 Exercises

1. A group of 8,027 people was classified on the basis of their gender and periodontal status as shown in the following table. It is desired to test if there is any association among periodontal status and gender.

	periodontal status			
gender	healthy	gingivitis	perio	Total
male	1143	929	937	3009
female	2607	1490	921	5018
Total	3750	2419	1858	8027

   a. State the null and alternative hypotheses of interest for this problem.
   b. Compute the chi-square statistic and the p-value.
   c. Write your conclusion based on the p-value? [2+5+3=10pts]



Review- 3

One Proportion	Hypothesis Test: [STAT], [TESTS] [1-PropZTest] Conf Interval: [STAT], [TESTS], [1-PropZint...]
Two Proportions	Hypothesis Test: [STAT], [TESTS] [2-PropZTest] CI for $p_1 - p_2$: [STAT], [TESTS] [2-PropZInt...]
Mean	Hypothesis Test: [STAT], [TESTS] [T-Test] CI for Mean: [STAT], [TESTS] [TInterval]
Comparing Two Means (independent samples)	Hypothesis Test: [STAT], [TESTS] [2-SampTTest] Conf Interval: [STAT], [TESTS], [2-SampTInt...]
Matched Pair (dependent samples)	Apply one-sample t method for the differences

Some Important Results

1. If a test rejects H_0 at the level of significance .05, then the test should also reject H_0 at any level of significance $>$ .05.
2. If H_a is right-sided (i.e., $>$) and the test statistic is large, then the p-value should be small.
3. If H_a is left-sided (i.e., $<$) and the test statistic is large negative, then the p-value should be small.
4. The maximum probability of rejecting the null hypothesis when it is actually true is the level of significance α.
5. The width of a CI increases with increasing confidence level.
6. The margin of error of a CI is one half of the width of the CI.
7. The margin of error increases with increasing confidence level.
8. If a CI for the difference of means, say, $\mu_1 - \mu_2$ is of the form

   * $(+,+)$ then $\mu_1 > \mu_2$
   * $(-,+)$ then μ_1 and μ_2 are not significantly different
   * $(-,-)$ then $\mu_1 < \mu_2$
9. The p-value is a probability, and so

$$0 \leq \text{p} - \text{value} \leq 1.$$

10. The correlation coefficient r is a measure of association between two variables, and

$$-1 \leq r \leq 1.$$

11. The correlation coefficient is not affected by units of measurements. For example, the correlation between (height and weight) expressed in (inches and pounds) is the same when the data are transformed from (inches, pounds) to (centimeter, kilograms).
12. In a simple linear regression model,

$$y = a + bx,$$

a is the intercept (the value of y at $x = 0$) and b is the slope. For every one unit increase x, y increases by b units.

Some Problems

1. Two very large herds are managed under different husbandry systems. Random sample of 68 animals from the first herd, and a random sample of 55 animals from the second, were selected as sentinel groups just before the rainy season began. The attack rate for a common wet-season complaint was recorded for each group as shown in the following table.

	Herd 1	Herd 2
Sample Size	$n_1 = 68$	$n_2 = 55$
Number of infected animals	$X_1 = 25$	$X_2 = 11$

We wish to investigate whether there is a difference in the attack rates under the two husbandry systems. Test appropriate hypotheses at the level of 0.05.

a. Identify the parameters of interest for this testing problem.

b. What is the appropriate test?

c. State the null and alternative hypotheses.

d. Calculate the value of the test statistic and the p-value.

e. Write the conclusion of the test.

f. Does the conclusion of the test in part **e** hold at the level .01? Explain.

2. Many people believe that they are allergic to penicillin, but have not had allergy testing. These people are often given alternative antibiotics prior to surgery to ward off infection. As antibiotic choices are limited, treatment alternatives may be more toxic, more expensive and less effective. According to a study presented at the American College of Allergy, Asthma and Immunology (ACAAI) Annual Scientific Meeting 2014, 384 people who believed they were allergic to penicillin were given penicillin skin testing, and 361 tested negative for penicillin allergy. On the basis of this information, it is desired to estimate the true proportion of people who believe that they are allergic to penicillin when they are actually not.
   a. Identify the parameter of interest in the above problem.

   b. What interval estimation method is appropriate for the above problem?

   c. Find a 95% confidence interval for the parameter in part **a**

   d. Interpret the meanings of the CI in part **b**..

   e. What is the margin of error in the 95% confidence interval?

3. In general, males are taller than females in the same age group. The following results (in inches) were collected from a sample 156 females and a sample of 145 males from a university.

	Males	Females
Sample Size	$n_1 = 145$	$n_2 = 156$
Mean height $\bar{x}$	68.5	64.25
SD s	3.12	2.91

   It is desired to estimate the difference between mean heights of male and female students in the university.
   a. Identify the parameters of interest in the above problem.

   b. Find a 95% confidence interval for the difference between the parameters in part **a**

   c. Interpret the meanings of the CI in part **b**.

   d. On the basis of the CI in part **c**, can we say that

   e. What is the margin of error in the 95% confidence interval?

4. Suppose that a 95% CI for the mean difference $\mu_1 - \mu_2$ based on two independent samples indicates that $\mu_1 > \mu_2$. Would a 90% CI based on the same samples indicate the same? What about a 99% CI?

5. Suppose that a 90% CI for the mean difference $\mu_1 - \mu_2$ based on two independent samples includes zero. Would a 95% CI based on the same samples indicate μ_1 and μ_2 are significantly different? Explain.

6. Consider testing $H_0 : p_1 \leq p_2$ vs. $H_a : p_1 > p_2$, where p_1 and p_2 denote the proportions of an attribute in two populations. A test based on independent samples indicated that $p_1 > p_2$ at the level .05. Can we conclude the same at the level of .10? What about at the level .01?

7. Consider the following data that represent the heart rates for six people before and half an hour after drinking two cups of coffee.

After:	83	66	77	74	75	71
Before:	78	64	70	71	70	68

   It is desired to test if drinking coffee raises the heart rate.

   a. What is an appropriate test?

   b. Identify the parameter to be tested.

   c. State the null and alternative hypotheses.

   d. Find the test statistic and the p-value.

   e. Write the conclusion of the test.

8. A 95% CI for the mean of a population on the basis of a sample is

   $$32.4 \pm 2.56.$$

   Should the margin of error of a 90% CI based on the same sample be greater than 2.7? Explain. Can we say the ME of a 99% CI based on the same sample should be greater than 2.56?

9. In order to estimate the true annual household income in a neighborhood, a sample of 10

households was selected, and the incomes (in $1,000) were recorded as follows.

```
59.77 62.22 63.88 63.60 60.13 58.96 65.99 62.14 70.32 60.84
```

It is desired to estimate the mean annual income in the entire neighborhood.

a. What interval estimation method is appropriate?

b. What is the parameter of interest?

c. Find a 95% CI for the the mean.

d. Interpret the meanings of the CI in part **c.**

10. Suppose that a 95% CI for the mean difference $\mu_1 - \mu_2$ is $(1.2, 3.4)$.
    a. Does this CI indicate that $\mu_1 > \mu_2$? Explain.

    b. Does the 90% CI based on the same data indicate the same? Explain.

    c. Between the 90 and 95 percent CIs, which one has the smaller ME? Explain.

11. The quality control inspector in a manufacturing company likes to check if the percentage of defective items produced during the night shift is higher than that produced in day shift. In a sample of 120 items produced in night shifts 6% were found to be defective, and in a sample of 140 items produced in day shifts 4% were found to be defective.
    a. What is the appropriate test for the above problem?
    b. State the null and alternative hypotheses.
    c. What is the p-value?
    d. Can we conclude that the percentage of defective items produced in night shifts is higher than that in day shifts at the level .05? Explain.
12. A 95% CI for the mean of a population is given by $(12.5, 14.6)$. Identify the following statements as TRUE or FALSE.
    a. We are 95% confident that the population mean is between 12.5 and 14.6.

    b. We can conclude that the true mean is between 12.5 and 14.6.

    c. The sample mean is between 12.5 and 14.6 with probability .95.

    d. The probability that the population mean is between 12.5 and 14.6 is .95

13. According to the credit rating company Equifax, credit limits on newly issued credit cards have increased between Jan 2011 and May 2011. The following data were collected to estimate the mean difference.

Jan 2011	May 2011
$n_1 = 400$	$n_2 = 500$
$\bar{x}_1 = \$2635$	$\bar{x}_2 = \$2887$
$s_1 = \$365$	$s_2 = \$412$

a. What interval estimation method is appropriate for this problem?

b. Find a 95% CI for the difference between the mean credit limits in Jan 2011 and May 2011.

c. Does the CI indicate that the mean credit limit has been increased since Jan 2011? Explain.

d. State the assumptions under which your conclusion is valid.

14. The manufacturer of a gasoline additive claims that the use of this additive increase gasoline mileage. A random sample of 10 cars were driven one week without additive, and then for one week with the gasoline additive with the following data: It is desire to estimate the

mileage data

cars	1	2	3	4	5	6	7	8	9	10
no additive	24	24	23	25	25	25	28	28	24	26
additive	28	27	27	28	30	28	25	28	27	25

mean difference in mileage of gasoline without additive and with additive.

a. What interval estimating method is appropriate for this problem?

b. Find a 95% CI for the mean difference in mileage.

c. What is the average difference?

d. On the basis of the CI in part **b**, can we conclude that the gasoline additive increase the mean mileage? Explain.

15. The tensile strength of a cable for upper-limb prosthesis was investigated. Stainless steel cable is commonly available in three sizes (diameters): 1.19 mm, 1.59 mm and 2.38 mm. Four tests were performed for each diameter size and the results are given in Table 11.3. Let X be cable cross area (mm^2) and Y be tensile strength (KN)

Table 11.3: Cable cross area and tensile strength

Cable diameter (mm)	Cable cross area (mm^2)	Tensile strength (KN)
1.19	1.1122	1.27
1.19	1.1122	1.45
1.19	1.1122	1.43
1.19	1.1122	1.36
1.59	1.9856	2.20
1.59	1.9856	2.56
1.59	1.9856	2.38
1.59	1.9856	2.45
2.38	4.4488	4.58
2.38	4.4488	5.03
2.38	4.4488	5.67
2.38	4.4488	4.39

a. Construct a scatter plot X and Y and interpret its meanings.

b. Calculate the correlation coefficient for the data and comment on the result.

c. Obtain the least squares estimates slope and intercept parameters of the regression model of Y and X.

d. Find the coefficient of determinant r^2 and interpret its meanings.

e. Write the fitted model $\widehat{Y} = \widehat{a} + \widehat{b}X$.

f. Interpret the meanings of the estimated slope $\widehat{b}$.

g. Predict the tensile strength of a cable with cross area 3.5.

h. Find a 95% prediction interval for the tensile strength of a cable with cross area 3.5.

i. Find a 95% confidence interval for the mean tensile strength of cables with cross area 3.5.

j. Comment on the validity of the fitted regression equation to estimate the tensile strength for cable with cross area 6.

16. A study of traffic violations and drivers who use cell phones while driving produced the following data: Test if there is any association between traffic violation and cell phone use

	No. of violations in a year		Total
Cell phone used	60	260	320
Cell phone not used	45	405	450
Total	105	665	770

while driving at the level .05. State the hypotheses.

17. **Comparison Between Two Means**

There are two methods for comparing means, one is for independent samples and one is for matched pair samples (or dependent samples).

Independent Samples: Let $(\bar{x_1}, s_1)$ be (mean, std deviation) based on a sample from a normal population with mean μ_1 and std deviation σ_1. Let $(\bar{x_2}, s_2)$ be (mean, std deviation) based on a sample from a normal population with mean μ_2 and std deviation σ_2. We like to test, for example,

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 \neq \mu_2.$$

The test statistic is given by

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

To compute the test statistic and the p-value using TI calculator:

Select STAT $\rightarrow$ TESTS $\rightarrow$ 2-SampTTest;
select [Stats] for input if the means and std deviations are given,
else select [Data] for input;
enter the means, std deviations, and sample sizes;
choose $\neq \mu_2$ (depends on the H_a);
choose "No" for pooled, and calculate.

For example, let $\bar{x_1} = 3.4, s_1 = 1.2, n_1 = 10, \bar{x_2} = 2.9, s_2 = 1.0$, and $n_2 = 20$. Suppose we want to test

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_a : \mu_1 > \mu_2.$$

To compute the p-value and the test statistic:
Select STAT $\rightarrow$ TESTS $\rightarrow$ 2-SampTTest;
select [Stats] for input;
enter the means, std deviations, and sample sizes.
Choose $> \mu_2$; choose "No" for pooled, and calculate to get
test statistic t = 1.135 and p-value p = .137.

Since the p-value is not less than 0.05, we do not reject the null hypothesis. That is, no evidence to indicate that μ_1 is greater than μ_2.

To compute a 95% confidence interval for $\mu_1 - \mu_2$:
Select STAT $\rightarrow$ TESTS $\rightarrow$ 2-SampTint...;

Select [Stats] for input; enter the means, std deviations, and C-Level .95; choose "No" for pooled, and calculate to get

$$(-.436, 1.436).$$

Note that the CI is $(-,+)$, which implies that the population means μ_1 and μ_2 are not significantly different.

18. **Matched Pair or Dependent Samples**

Here the samples are from related individuals (husband and wife, twin or samples are collected from the same individuals at two different time periods). We have to use the one-sample t test or t-interval for the differences of the samples.

sample 1	x_1	x_2	...	x_n
sample 2	y_1	y_2	...	y_n
difference	d_1	d_2	...	d_n

Apply the one-sample t-test or t-interval using the differences $d_1, ..., d_n$. For example, if the CI for the mean difference includes zero, then we conclude that the means are not significantly different.

19. Suppose that a 95% CI for the mean difference $\mu_1 - \mu_2$ is $(1.2, 3.4)$.
    - **a.** Does this CI indicate that $\mu_1 > \mu_2$? Explain.
    - **b.** Does the 90% CI based on the same data indicate the same? Explain.
    - **c.** Between the 90 and 95 percent CIs, which one has the smaller ME? Explain.
20. A random sample of 100 electronic components was tested to estimate the average lifetime. The sample mean is 125 hours with the standard deviation of 5 hours. It is desired to estimate the mean life hours of all electronic components produced by the manufacturer.
    - **a.** What is the appropriate interval estimating method?
    - **b.** Find a 95% CI for the mean.
    - **c.** What is the ME of your CI?
    - **d.** On the basis of the 95% CI, can we conclude that the mean life hours is at least 115 hours? Explain.
21. A 95% CI for the mean of a population is given by $(12.5, 14.6)$. Identify the following statements as TRUE or FALSE.
    - **a.** We are 95% confident that the population mean is between 12.5 and 14.6.
    - **b.** We can conclude that the true mean is between 12.5 and 14.6.
    - **c.** The sample mean is between 12.5 and 14.6 with probability .95.
    - **d.** The probability that the population mean falls in the interval (12.5, 14.6) is .95
22. The following display presents the results of a hypothesis test on a mean.

t-test
$\mu \neq 127$
t = 1.50821
p = .0734
$\bar{X} = 131.6$
$S = 20$
$n = 43$

- **a.** What are the null and alternative hypotheses?
- **b.** What is the value of the test statistic?

c. What is the p-value?
d. Do you reject the null hypothesis at $\alpha = .05$ level?
e. Do you reject the null hypothesis at $\alpha = .10$ level?

Answers to Problems

Chapter 3

Exercise 3.2–3.3

3.3.1 $1-.3-.4=.3$

3.3.2 **a.** $P(E_6)+P(E_7)=1-.1-.05-.15-.15-.25=.3$. So $P(E_6)=.3/2=.15$ and $P(E_7)=.15$

b. $P(A)=.1+.15+.15=.4$

c. $P(B)=.05+.15+.2+.15=.55$

d. B, because $P(B)>P(A)$

3.3.3 **a.** $\{(1,1,1),(1,2,1),(2,1,1),(2,2,1),(1,1,2),(1,2,2),(2,1,2),(2,2,2)\}$

b. $\frac{3}{8}$

c. $\frac{4}{8}=\frac{1}{2}$

3.3.4 **a.** The sample space is

$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$

b. $\frac{4}{24}=\frac{1}{6}$

c. $\frac{18}{24}=\frac{3}{4}$.

3.3.5 **a.** $\{RR,RB,BR\}$; **b.** $\frac{1}{3}$; **c.** $\frac{2}{3}$.

Exercise 3.4

3.4.1 For each student, there are 5 possibilities. So $5^7=78125$.

3.4.2 $\frac{12}{52}=\frac{3}{13}$

3.4.3 **a.** $\frac{1}{2}$; **b.** $\frac{1}{2}$; **c.** $\frac{1}{6}$

3.4.4 **a.** $\frac{1}{52C4}=\frac{1}{270725}$; **b.** $\frac{4}{52C4}=\frac{4}{270725}$.

3.4.5 $2^{10}=1,024$

3.4.6 $5! = 120$

3.4.7 **a.** $\frac{9!}{(9-3)!} = \frac{9!}{6!} = 504$; **b.** $\frac{8!}{(8-2)!} = 56$.

3.4.8 **a.** $\frac{10!}{(10-5)!} = 30,240$; **b.** $\frac{1}{30240}$

3.4.9 **a.** $40C4 = 91,390$; **b.** $(15C2) \times (25C2) = 31,500$

Exercise – Chapter 3

3.6.1 **a.** No. There are outcomes common to A, B and C.
b. No. There are outcomes not in $(A \cup B \cup C)$. The union does not exhaust the sample space.
c. $P(A|B) = \frac{4}{11}$; $P(B|A) = \frac{4}{9}$.
d. $P((A \cap C)|B) = \frac{2}{11}$; $P((A \cap B)|C) = \frac{2}{11}$.
e. $P(A^c \cup B^c) = P((A \cap B)^c) = \frac{22}{26}$.

3.6.2 **a.** .45; **b.** .96; **c.** .96

3.6.3 **a.** $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.1}{.4} = \frac{1}{4}$.
b. $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.1}{.3} = \frac{1}{3}$.
c. Since the events A and C are disjoint, $P(A|C) = 0$.

3.6.4 **a.** Total number of outcomes in the sample space is $2^{10} = 1,024$. Only one outcome includes all heads, so the probability is $\frac{1}{1024}$.
b. All outcomes includes at least one head, except the one $(TT...T)$. The probability is $\frac{1023}{1024}$.

3.6.5 Since the outcomes of all four rolls are independent,

$$P(7,7,7,7) = P(7)P(7)P(7)P(7) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{1296}.$$

3.6.6 **a.** Let A denote the event of observing 7. There are six pairs (6,1),(5,2),...,(1,6), each with sum 7, and so the probability of observing 7 is $P(A) = \frac{6}{36} = \frac{1}{6}$.
b. The number of outcomes with sum 7 in the conditional sample space is 6. Of these 6 outcomes only two outcomes, namely, (2,5) and (5,2) differ by 3. So the conditional probability is $\frac{2}{6} = \frac{1}{3}$.

3.6.7 $.5 \times .7 = .35$

3.6.8 **a.** $P\left(\text{commit suicide}\middle|\text{Cinese Woman}\right) = 1.4P\left(\text{commit suicide}\middle|\text{Cinese man}\right)$
b. $P\left(\text{developing Alzheimer's disease}\middle|\text{hispanic}\right) = 1.5P\left(\text{developing Alzheimer's disease}\middle|\text{white}\right)$
c. E_1 = the event that engine 1 fails, E_2 = event that engine 2 fails. $P(E_1 \cap E_2) = .0001$.
d. $P(gettingadmission) = .6$; $P(scholarship|admission) = .15$.

3.6.9 **a.** $\frac{305}{636} = .480$.
b. $\frac{208}{636} + \frac{153}{636} - \frac{90}{636} = \frac{271}{636} = .426$.
c. L = event of returning a laptop; H = event of returning an equipment for hardware problems.

$$P(L^c \cap H^c) = P((L \cup H)^c) = 1 - P(L \cup H) = .311$$

d. $\frac{78}{208}$

3.6.10 **a.** $\frac{1}{4}$; **b.** $\frac{1}{7}$

3.6.11 **a.** $\frac{2400}{36300} = .661$; **b.** .278; **c.** $1 - .278 = .722$

3.6.12 $\frac{.75 \times .5}{.75 \times .5 + .25 \times .5} = .75$

3.6.13 $\frac{.6 \times .25}{.6 \times .25 + .5 \times .4 + .35 \times .35} = .823$

3.6.14 quad $\frac{.06 \times .04}{.06 \times .04 + .04 \times .96} = .059$

3.6.15 Let A = HIV infected; B = the test is positive. $P(A \cap B) = .0038$, $P(B) = .1532$ and $P(A|B) = .025$.

Chapter 8

Section 8.3

1. **a.** $H_0 : \mu \leq 69.3$ vs. $H_a : \mu > 69.3$.
b. $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{69.9 - 69.3}{3/\sqrt{100}} = 2.0$
c. p-value = $P(t_{99} > 2.0) = \text{tcdf}(2.0, 10^7, 99) = .0241$.

d. Since the p-value is less than .05, we conclude that on average the male business executives are taller than the general male population.

e. Since we used the level of significance of 0.05, the maximum probability of rejecting the null hypothesis when it is true is 0.05.

Chapter 11

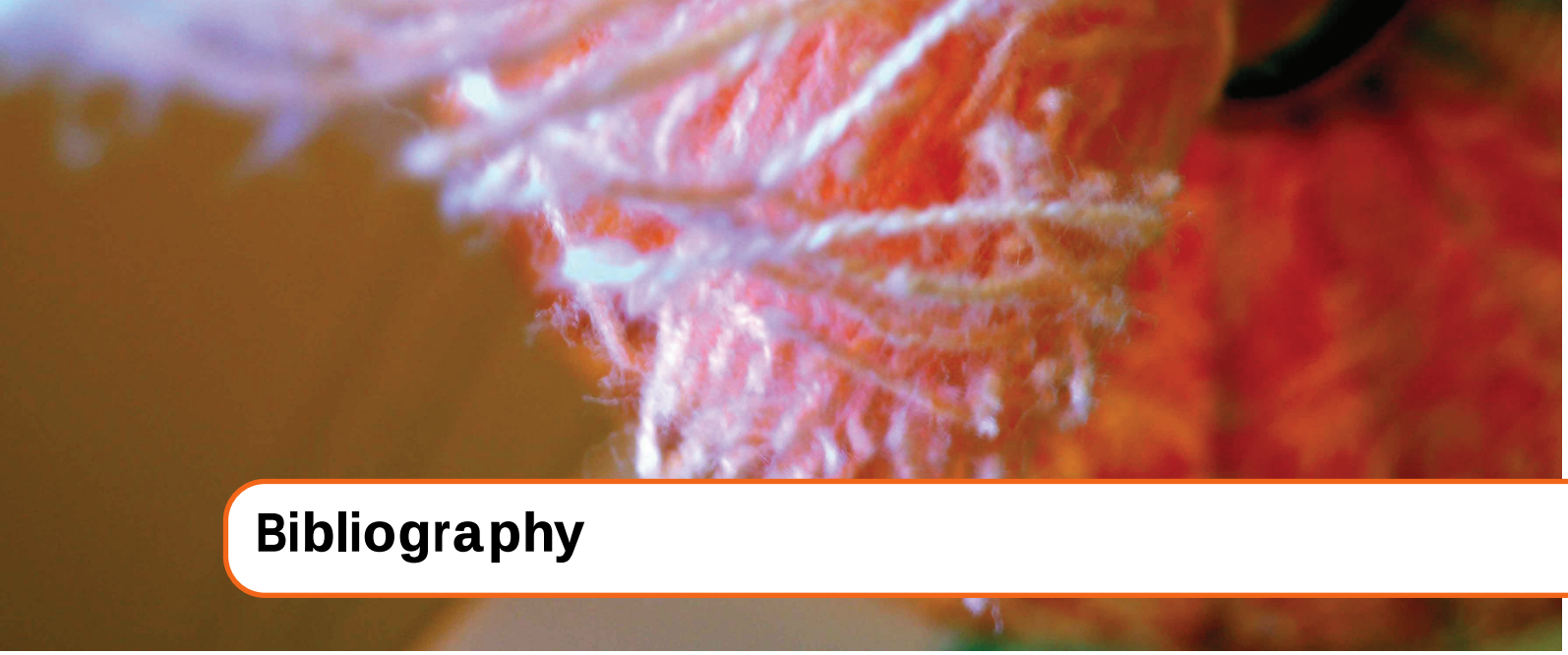
1. The $\chi^2 -$ statistic $= 165.17$ with the p-value of $7.711...10^{-27} = 0$.

Solutions to Test 2

1. (a) The sample proportion is $\frac{120}{500} = 0.24$
   (b) (.203, .277)
   (c) $\frac{.277-.203}{2} = .037$
   (d) Yes, because $0.203 > .20$
2. (a) $(-.48, 4.48)$
   (b) $\frac{4.48-(-.48)}{2} = 2.48$
   (c) No, because $(-,+)$
3. (a) $H_0 : \mu_M = \mu_W$ vs. $H_a : \mu_M > \mu_W$
   (b) 1.708
   (c) .049
   (d) Reject H_0; yes.
4. (a) $H_0 : p_H = p_L$ vs. $H_a : p_H > p_L$
   (b) 3.705
   (c) p-value = .0001
   (d) Since the p-value < .05, reject H_0; proportion of high income voters support the increase in sales tax is more than that of low income voters.
   (e) .05
5. (a) F
   (b) F
   (c) T
   (d) F
   (e) T

11.2 Problems

Problem 11.1 What is the average airspeed velocity of an unladen swallow?



Bibliography