

CHAPTER 1

INTRODUCTION

- Engineers and scientists are constantly exposed to collections of facts, or data. The discipline of statistics provides methods for organizing and summarizing data, and for drawing conclusions based on information contained in the data.
- An investigator who has collected data may wish simply to summarize and describe important features of the data. This entails using methods from **descriptive statistics**.

Descriptive Statistics consists of methods for organizing and summarizing information. Some of these methods involve calculation of numerical summary measures, such as means, standard deviations, etc. Other descriptive methods are graphical in nature.

Notation: Given a data set consisting of n observations on some variable x , the individual observations will be denoted by

$$x_1, x_2, x_3, \dots, x_n$$

(though any other letter could be used in place of x).

MEASURES OF LOCATION (1.4)

Descriptive measures that indicate where the center or most typical value of a data set lies called **measures of center**.

measures of center

- sample mean
- sample median

The **sample mean** \bar{x} of observations

$$x_1, x_2, \dots, x_n$$

is given by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

The **sample median** is the number that divides the bottom 50% of the data from the top 50%. To obtain the median of a data set, we arrange the data in increasing order and then determine the middle value in the ordered list.

When the observations are denoted by x_1, x_2, \dots, x_n , we will use the symbol \tilde{x} to represent the sample median.

How to find the sample median

- If the number of observations n is odd, then the **sample median** \tilde{x} is the observation exactly in the middle of the ordered list, i.e. the $(\frac{n+1}{2})^{\text{th}}$ ordered value.
- If the number of observations is even, then \tilde{x} is the mean of the two middle observations in the ordered list, i.e. the average of $(\frac{n}{2})^{\text{th}}$ and $(\frac{n}{2} + 1)^{\text{th}}$ ordered values.

Ex.1 on p. 20: A manufacturer of electronic components is interested in determining the lifetime of a certain type of battery. A sample, in hours of life, is as follows:

123, 116, 122, 110, 175, 126, 125, 111, 118, 117

- Find the sample mean and median.
- What feature in this data set is responsible for the substantial difference between the two?

TRIMMED MEANS:

Motivation

- Note that \bar{x} and \tilde{x} are at opposite extremes of the same “family” of measures.
- After the data set is ordered, \tilde{x} is computed by throwing away as many values on each end as one can without eliminating everything (leaving just one or two middle values) and averaging what is left, whereas
- to compute \bar{x} one throws away nothing before averaging.

To paraphrase, the mean involves trimming 0% from each end of the sample, whereas for the median the maximum possible amount is trimmed from each end.

A **trimmed mean** is a compromise between \bar{x} and \tilde{x} . A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what is left over.

Example: The following is a set of algebra final exam scores:

0 58 61 63 67 69 70 71 78 80

- The sample mean =
- The sample median =
- The 10% trimmed mean of the data =

MEASURES OF VARIABILITY (1.5)

Different samples may have identical measures of center yet differ from one another in other important ways. The first sample has the largest amount of variability, the third has the smallest amount.

The sample range is the difference between the largest and the smallest sample value.

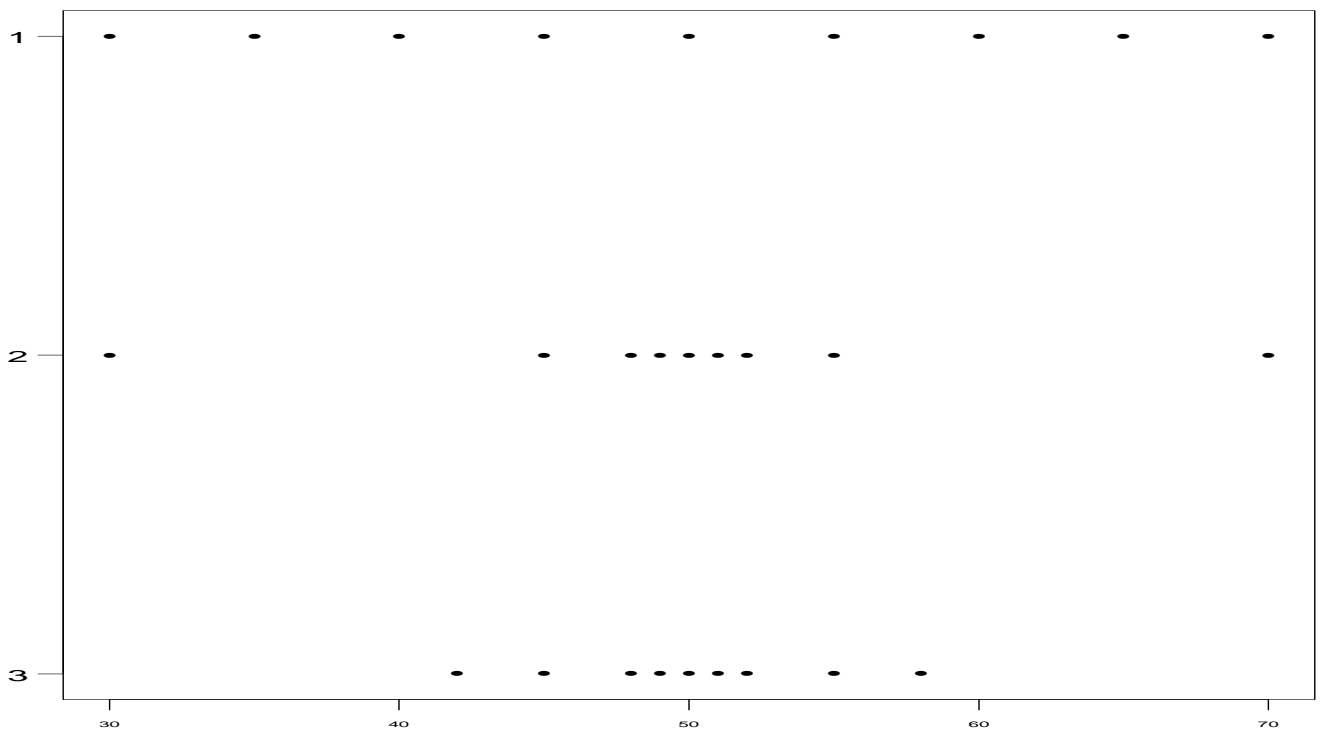


Figure 1: Samples with identical measures of center but different amounts of variability

The sample variance: The deviation from the mean are

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}.$$

A simple way to combine the deviations into a single quantity is to average them. Unfortunately, there is a major problem with this suggestion:

$$\text{sum of deviations} = \sum_{i=1}^n (x_i - \bar{x}) = 0$$

so that the average deviation is always zero.

How can we change the deviations to nonnegative quantities so the positive and negative deviations do not counteract one another when they are combined?

One possibility is to work with the absolute values of the deviations and calculate the average absolute deviation $\sum |x_i - \bar{x}|/n$.

Because the absolute value operation leads to a number of theoretical difficulties, consider instead $\sum (x_i - \bar{x})^2/n$, but for several reasons we will divide the sum of squared deviations by $n - 1$ rather than n .

The sample variance, denoted by s^2 , is given by

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

The sample standard deviation, denoted by s , is the (positive) square root of the variance: $s = \sqrt{s^2}$
An alternative expression for the numerator of s^2 is

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

The unit for s is the same as the unit for each of the x_i 's

Example:

The Bureau of Labor Statistics lists average hourly wages for 8 categories of law-related occupations. The units are dollars per hour.

$$30 \quad 17 \quad 36 \quad 14 \quad 17 \quad 12 \quad 15 \quad 17$$

The sample mean is

$$\bar{x} = \frac{30 + 17 + 36 + \dots + 17}{8} = \frac{158}{8} = 19.7517$$

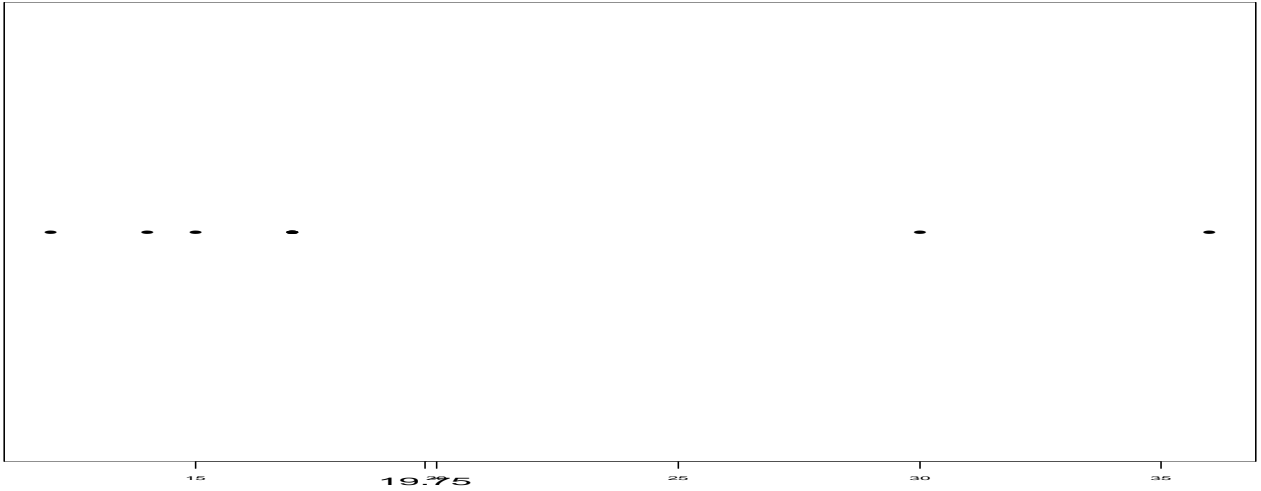


Figure 2: Average hourly wages

The deviations show how spread out the data are about their mean.

Observations	Deviations	Squared deviations
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
30	$30 - 19.75 = 10.25$	$10.25^2 = 105.0625$
17	$17 - 19.75 = -2.75$	$(-2.75)^2 = 7.5625$
36	$36 - 19.75 = 16.25$	$16.25^2 = 264.0625$
14	$41 - 19.75 = -5.75$	$(-5.75)^2 = 33.0625$
17	$17 - 19.75 = -2.75$	$(-2.75)^2 = 7.5625$
12	$12 - 19.75 = -7.75$	$(-7.75)^2 = 60.0625$
15	$15 - 19.75 = -4.75$	$(-4.75)^2 = 22.5625$
17	$17 - 19.75 = -2.75$	$(-2.75)^2 = 7.5625$
sum = 0		sum = 507.5

$$s^2 = \frac{507.5}{7} = 72.5$$

The standard deviation is

$$s = \sqrt{72.5} = 8.5147 \text{ \$ per hour}$$

DISCRETE AND CONTINUOUS DATA (1.6)

A variable is **discrete** if its set of possible values either is finite or else can be listed in an infinite sequence. A variable is **continuous** if its possible values consist of an entire interval on the number line.

GRAPHICAL METHODS AND DATA DESCRIPTION

(1.7)

Stem-and-Leaf Displays (stem-plot):

Example: Consider the following humidity readings rounded to the nearest percent:

29 44 12 53 21 34 39 25 48 23
17 24 27 32 34 15 42 21 28 37

This can also be written as

Stem	Leaf
1	2 5 7
2	1 1 3 4 5 7 8 9
3	2 4 4 7 9
4	2 4 8
5	3

Steps for Constructing a Stem-and-Leaf Display

- Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
- Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
- Write each leaf in the row to the right of its stem, in increasing order out from the stem.

HISTOGRAMS

A **frequency distribution** is a table that divides a set of data into a suitable number of classes (categories), showing also the number of items belonging to each class.

For the previous example

Stem	Leaf
1	2 5 7
2	1 1 3 4 5 7 8 9
3	2 4 4 7 9
4	2 4 8
5	3

we might group these data into the following distribution:

Class interval	Class midpoint	Frequency, f	Relative frequency
10 – 19	14.5	3	0.15
20 – 29	24.5	8	0.40
30 – 39	34.5	5	0.25
40 – 49	44.5	3	0.15
50 – 59	54.5	1	0.05

Class interval	Class midpoint	Frequency, f	Relative frequency
10 – 19	14.5	3	0.15
20 – 29	24.5	8	0.40
30 – 39	34.5	5	0.25
40 – 49	44.5	3	0.15
50 – 59	54.5	1	0.05

Relative frequency histogram

HISTOGRAM SHAPES

- A histogram is symmetric if the left half is a mirror image of the right half.
- A histogram is positively skewed if the right or upper tail is stretched out compared with the left or lower tail
- A histogram is negatively skewed if the stretching is to the left.

USING MINITAB

Example: Ex 7 on p. 20

The following data represent the length of life in years, measured to the nearest tenth, of 30 similar fuel pumps:

```

2.0  3.0  0.3  3.3  1.3  0.4
0.2  6.0  5.5  6.5  0.2  2.3
1.5  4.0  5.9  1.8  4.7  0.7
4.5  0.3  1.5  0.5  2.5  5.0
1.0  6.0  5.6  6.0  1.2  0.2

```

First we store the data in column C1. Then we proceed in the following manner.

Mean

- Choose Calc > Column Statistics ...
- Select the **Mean** option bottom from the Statistic field.
- Click in the Input variable text box and specify C1.
- Click OK

Median

- Using the same steps except selecting the **Median** option bottom instead of the Mean option bottom, we obtain the median

Stem-and-Leaf

- Choose Graph > Stem-and-Leaf
- Specify C1 in the Variable text box.
- Click in the Increment text box and type 1 (This tells Minitab that the distance between the smallest possible number on one line and the smallest possible number on the next line should be 1)
- Click OK

Histogram

- Choose Graph > Histogram
- Specify C1 in the X text box for Graph 1
- Click OK

FUNDAMENTAL SAMPLING DISTRIBUTION

AND DATA DESCRIPTION: CH 8

Data Display and Graphical Methods

Quartiles: The quartiles of a data set divide it into quarters, or four equal parts. The first quartile, is the number that divides the bottom 25% of the data from the top 75% ; the second quartile is the median; and the third quartile, is the number that divides the bottom 75% of the data from the top 25%. The first and third quartiles are the 25th and 75th percentiles, respectively

The interquartile range, denoted IQR, is the difference between the first and third quartiles; that is, Roughly speaking, the IQR gives the range of the middle 50% of the observations.

Outliers: are observations that fall well outside the overall pattern of the data. We can use quartiles and the IQR to identify potential outliers. Outliers are numbers that lie, respectively, 1.5 IQRs below the first quartile and 1.5 IQRs above the third quartile

Box and Whisker Plot or Boxplot

We will go over Example 8.3 on p. 202 using minitab.

Comparative Boxplots

A comparative or side-by-side boxplot is a very effective way of revealing similarities and differences between two or more data sets.

We will go over Ex 3. from ch 1 using minitab.

COMPARATIVE STEM-PLOTS

A study at 35 large city high schools given the following back-to-back stem-plot of the percentages of students who say they have tried alcohol.

School Year 1995 – 96	0	School Year 1998 – 99
	1	
2	1	
9	2	
6 6 3	3	1 8
4 3 1 1	4	0 2 9
9 9 8 6 5 3 2 2 0	5	3 3 4 6
9 8 7 6 6 5 1 1	6	1 2 2 2 7
5 4 3 2 2	7	0 1 3 3 5 5 6 7 8 9
5 4 0	8	2 3 4 5 8 8 9
0	9	0 1 1 2

Which of the following does not follow from the above data?

- (A) In general, the percentage of students trying alcohol seems to have increased from 1995–96 to 1998–99
- (B) The median alcohol percentage among the 35 schools increased from 1995 – 96 to 1998 – 99.
- (C) The spread between the the lowest and highest alcohol percentages decreased from 1995–96 to 1998–99.
- (D) For both school years in most of the 35 schools, most of the students said they had tried alcohol.
- (E) The percentage of students trying alcohol increased in each of the schools between 1995 – 96 and 1998 – 99.

Review

1. A sample was taken of the salaries of four employees from a large company. The following are their salaries (in thousands of dollars) for this year.

33 31 24 36

The variance of their salaries is

- a. 5.1
- b. 26
- c. 31

2. Consider the following Stem-and-Leaf plot.

1		6
2		2 4 8 9
3		0 1 1 2 3 6 7 8
4		0 5 8
5		0 1 8
6		1

The mean of the data represented in this stem-plot is

- a. 34.5
- b. 37
- c. cannot be computed from the information given.

3. In a class of 25 students, 22 students had grades between 71 and 80 (both endpoints included), and three students had grades between 91 and 100 (both endpoints included). For these data,

- a. the median must be between 71 and 80.
- b. the mean must be between 71 and 80.
- c. both of the above.

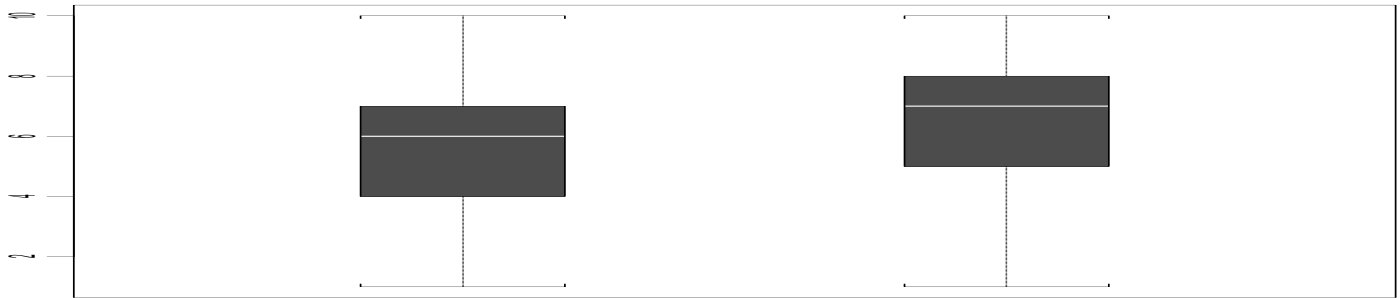


Figure 3: Physical Flexibility Rating

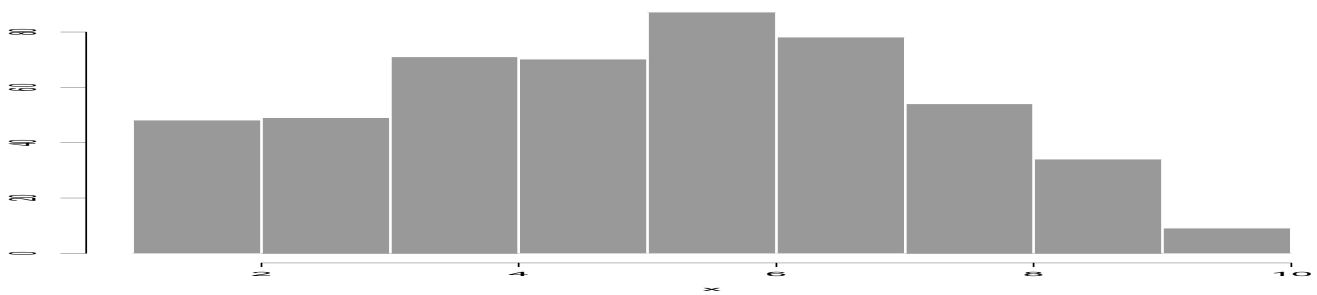


Figure 4: Middle-Aged Men

4. Five hundred randomly selected middle-aged men and five hundred randomly selected young adult men were rated on a scale from 1 to 10 on their physical flexibility, with 10 being the most flexible. Their rating appear in the frequency table below. For example, 17 middle-aged men had a flexibility rating of 1.

Physical Flexibility Rating	Frequency of Middle-aged Men	Frequency of Young Adult Men
1	17	4
2	31	17
3	49	29
4	71	39
5	70	54
6	87	69
7	78	83
8	54	93
9	34	73
10	9	39

(a) Display these data graphically so that the flexibility of middle-aged men and young adult men can be easily compared.

(b) Write a few sentences comparing the flexibility of middle-aged men with the flexibility of young adult

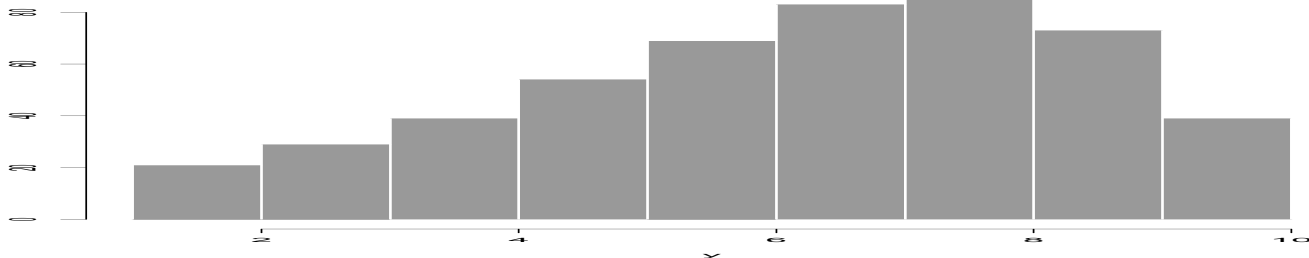


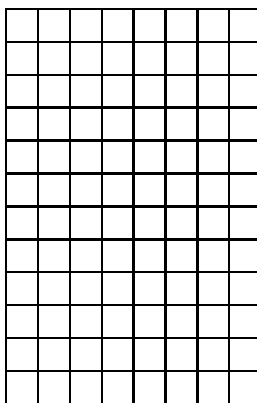
Figure 5: Young Adult Men

5. A consumer advocate conducted a test of two popular gasoline additives, A and B. There are claims that the use of either of these additives will increase gasoline mileage in cars. A random sample of 30 cars was selected. Each car was filled with gasoline and the cars were run under the same driving conditions until the gas tanks were empty. The distance traveled was recorded for each car. Additive A was randomly assigned to 15 of the cars and additive B was randomly assigned to the other 15 cars. The gas tank of each car was filled with gasoline and the assigned additive. The cars were again run under the same driving conditions until the tanks were empty. The distance traveled was recorded and the difference in the distance with the additive minus the distance without the additive for each car was calculated.

The following table summarizes the calculated differences. Note that negative values indicate less distance was traveled with the additive than without the additive.

Additive	Values Below Q_1	Q_1	Median	Q_3	Values above Q_3
A	-10, -8, -2	1	3	4	5, 7, 9
B	-5, -3, -3	-2	1	25	35, 37, 40

(a) On the grid below, display parallel boxplots (showing outliers if any) of the differences of the two additives.



(b) Which additive, A or B, would you recommend if the goal is to increase gas mileage in the highest proportion of cars? Explain your choice.

(c) Which additive, A or B, would you recommend if the goal is to have the highest mean increase in gas mileage? Explain your choice.