

CHAPTER 1

Introduction

Statistics: Statistics is the science of collecting, organizing, analyzing, presenting and interpreting data.

Variable: Any characteristic of a person or thing that can be expressed as a number is called a variable. The actual number that the variable takes is called a value.
(Height, Weight, Income, Sex, e.t.c.)

Categorical variable: A categorical variable simply records into which of several categories a person or thing falls. (Sex, Political party affiliation of a person e.t.c.) Working with categorical variables we use counts or percents. For example the variable, location. Code North=0 and South=1, East=2, and West=3. We cannot meaningfully compute the "average location".

Quantitative variable: We will call any variable that takes numerical values for which arithmetic makes sense a quantitative variable.

SECTION 1.1 Displaying Distributions.

Measurement: How do we begin to examine intelligently a set of a single measured variable? A set of numbers presented in a table without some background information is meaningless. Two questions to be answered here are: (1) What variable is being measured? and (2) how it was measured?

Users of data should be aware that taking numbers at face value, without thinking about the variable measured and the process used to measure it could produce misleading results.

EXAMPLE: Example 1.5 page 5.

Variation: When we measure a variable the values will vary, either due to the experimenter or the measurement instrument or the environment.

Distribution: The pattern of variation of a variable is called its distribution. The distribution records the numerical values of the variable and how often each value occurs.

A distribution is displayed by a stemplot or by a histogram. Stemplots separate each observation into stem and leaf, while histograms are based on a frequency or relative frequency of classes of values. When examining a distribution, first locate its center. Then look at the overall shape.

The shape of a distribution can be approximately Symmetric(each side of the center is a mirror image of the other) or Skewed(one tail

extends farther from the center than the other). The number of peaks is another aspect of overall shape.

Deviations from the overall shape of a distribution include gaps and outliers(individual observations that appear not to be in accord with the remaining data).

Categorical Variables: Bar Graph and Pie Chart

Do examples 1.7, 1.8, and 1.9 pages 7-8.

Stemplots(also called stem-and-leaf plots)

Stemplots offer a quick way to picture the shape of a distribution while including the actual numerical values in the graph. A stemplot works best for small numbers of observations that are all greater than 0.

Example: Given below are the numbers of home runs for Babe Ruth hit in each of his 15 years with the New York Yankees, 1920 to 1934.

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

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

The stemplot for Babe Ruth is given to the right.

The following is a back to back stemplot comparing Ruth and McGwire.

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

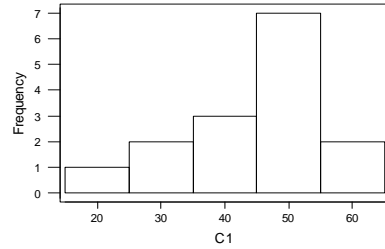
EXAMPLE - BABE RUTH

Stem-and -Leaf

```

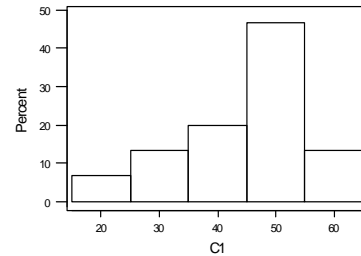
2      2 25
4      3 45
(7)   4 1166679
4      5 449
1      6 0
    
```

Histogram-Frequency



HR Count CumCnt Percent CumPct
Histogram-Percent

HR	Count	CumCnt	Percent	CumPct
22	1	1	6.67	6.67
25	1	2	6.67	13.33
34	1	3	6.67	20.00
35	1	4	6.67	26.67
41	2	6	13.33	40.00
46	3	9	20.00	60.00
47	1	10	6.67	66.67
49	1	11	6.67	73.33
54	2	13	13.33	86.67
59	1	14	6.67	93.33
60	1	15	6.67	100.00



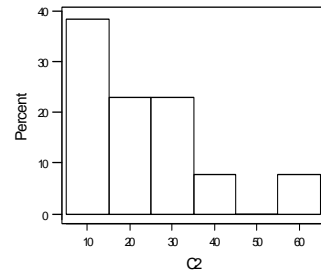
EXAMPLE- ROGER MARIS

Stem-and -Leaf

```

2      0 88
6      1 3446
(4)   2 3368
3      3 39
1      4
1      5
1      6 1
    
```

Histogram-Percent



Histograms

A histogram displays the count (or percent) of the observations that fall into each interval. We can choose any convenient number of intervals. Histograms are slower to construct by hand than are stemplots, and do not retain the actual values observed. The construction of a histogram is best shown by example.

(1) **Number of Classes:** Divide the range of the data into equal width. The goal is to use enough classes to show the variation in the data but not too many so that there are only a few items in many of the classes.

NOTE: IF YOU ALREADY BUILT THE STEM-AND-LEAF DISPLAY, THEN YOU CAN USE THE NUMBER OF STEMS AS THE NUMBER OF CLASSES FOR THE FREQUENCY DISTRIBUTION.

(2) **Count the number of observations in each class.** These counts are called frequencies.

EXAMPLE: Example 1.13 page 12.

(3) **Draw the Histogram**

EXAMPLE: Look at figure 1.10 (length of phone calls) page 16

Looking at Data

Some principles have emerged from our initial experiences with data that will remain valid as we advance. These are guidelines based on experience rather than hard fast rules that must always be followed.

1. To interpret data, you must first learn something of their context: What exactly was measured? How was the measurement carried out?
2. Always examine your data. An informative picture comes first usually supplemented by some numerical calculations.
3. Look first for an overall pattern, then for deviations from that pattern, such as outliers.

HOMEWORK (section 1.1): 1.16, 1.32, 1.33, 1.35, 1.39, 1.41
pp.22-27 (Use Minitab when you can)

SECTION 1.2 Describing Distributions

Population: A population is the set of all elements of interest in a particular study.

Sample: A sample is a subset of the population.

Measuring Center

The most common measure of center is the ordinary arithmetic average, or mean. Numerical measures that are computed for the population are called **population parameters;** when they are computed for a sample, they are called **sample statistics.**

(a) **Mean:** (Measuring Center)

Sample mean \bar{x} (x-bar): $\bar{x} = \frac{\sum x_i}{n}$; n: sample size.

Population mean μ (mu): $\mu = \frac{\sum x_i}{N}$; N: Population size.

The Mean provides a good measure of the center of a symmetric distribution. Consider a sample of 5 tests scores in English.

Example 1: $X_1=20$, $X_2=80$, $X_3=30$, $X_4=70$, $X_5=100$; NOTE: $n = 5$

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^5 x_i}{5} = x_1 + x_2 + x_3 + x_4 + x_5 = \frac{20+80+30+70+100}{5} = 60$$

(b) **Median:** (Measuring Center)

(1) Arrange the sample data in an ascending order.

(2) If the number of elements in the data set(sample) is an odd number then the median is the middle number (observation). The location of the median is found by counting $(n+1)/2$ observations up from the bottom of the list.

(3) If the number is even, the median is the average of the two middle numbers. The location of the median is found by counting $(\frac{n}{2})$ and $(\frac{n}{2}+1)$ observations up from the bottom of the list.

Example 2: In example 1 we have: $X_1=20, X_2=80, X_3=30, X_4=70, X_5=100$

(1) Ascending order: 20, 30, 70, 80, 100.

(2) The sample size $n=5$ is an odd number. Hence, the Median=70.

Example 3: Consider the sample: 20, 30, 50, 70, 80, 100.

$$\text{The Median is } \frac{50+70}{2} = 60.$$

NOTE: Although the mean is the most commonly used measure of the center of the distribution, the median is a better measure when the distribution is not symmetric (skewed to the right or left) and has extreme values.

(c) **Mode:** (Measure of the location of the most frequently occurring value in the data set). Mode is the value that occurs with the greatest frequency in the data set.

Example 4: Given the following sample: 2,2,3,3,4,4,4,4,5,5,6,6,6.
The Mode is 4.

Example 5: Given the following sample: 3,3,5,5,5,6,6,7,7,7,8,8.
The sample is bimodal. 5 and 7 are the modes.

Note: Usually, we use Mode for Categorical variables since the Mean and Median cannot be used.

<u>Car</u>	<u>Model</u>	<u>Frequency</u>	
Cherv.	Cavalier	9	
Ford	Escort	14	
Ford	Taurus	8	<u>Mode:</u> Ford Escort
Honda	Accord	11	
Hyundai	Excel	8	

Percentile: A percentile is a numerical measure that also locates values of interest in the data set. A percentile provides information regarding how the data items are spread over the interval from the lowest value to the highest value.

Defn. The p^{th} percentile of a data set is a value such that at least p percent of the items take on this value or less and at least $(100 - p)$ percent of the items take on this value or more.

Step 1: Sort the data in an ascending order, that is, from the smallest to the largest.

Step 2: Find $i = \left(\frac{p}{100}\right)n$ where n is the number of data values. i is a location. x_i is the number in location i .

Step 3: If i is not an integer, round it up to the next highest integer, then p^{th} percentile = x_i .

If i is an integer, then p^{th} percentile = $\frac{x_i + x_{i+1}}{2}$.

Example 6: Given the data below, find the 50th and 90th percentiles.

26, 4, 5, 20, 6, 12, 15, 15, 15, 8, 9, 10, 14, 18, 16, 17

Soln: Step 1: Data in ascending order .

$\begin{matrix} i= & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\ x_i = & 4, & 5, & 6, & 8, & 9, & 10, & 12, & 14, & 15, & 15, & 15, & 16, & 17, & 18, & 20, & 26 \end{matrix}$

90th perc.: $i = (90/100)16 = 14.4$; $\Rightarrow x_i = x_{15} = 20$; 90th perc. = 20

50th perc.: $i = (50/100)16 = 8$; since $i=8$; then 50th perc. = $\frac{x_8 + x_9}{2} = \frac{14 + 15}{2} = 14.5$

Note: The median and the 50th percentile are the same.

Quartiles: It is often desired to divide a data set into four parts with each part containing one-fourth of the data.

$$\begin{aligned} Q_1 &= \text{First Quartile} &= & 25\% \text{ percentile} \\ Q_2 &= \text{Second Quartile} &= & 50\% \text{ percentile} \\ Q_3 &= \text{Third Quartile} &= & 75\% \text{ percentile} \end{aligned}$$

Example 7: For the data given in **Example 6**, find the first, second, and third quartiles.

Soln. $Q_1 = 8.5, \quad Q_2 = 14.5, \quad Q_3 = 16.5$

Measures of Spread.

(1) **Range:** The Range is the simplest measure of variability.

Range: The difference between the largest and the smallest values.

Example 8: Reference **Example 6**. Find the Range.

Soln. The data is 4, 5, 6, 8, 9, 10, 12, 14, 15, 15, 15, 16, 17, 18, 20, 26

$$\text{Range} = 26 - 4 = 22$$

NOTE: It's not used very often because it's influenced so much by extreme values.

(2) **The Interquartile Range (IQR):** $IQR = Q_3 - Q_1$

Note: The IQR gives the range of the middle 50% of the observations.

The Five-Number Summary

The five number summary of a data set: Min, Q_1 , Q_2 , Q_3 , and Max.

Example 9: Reference **Example 6**. Find the five-number summary.

Soln. The data is 4, 5, 6, 8, 9, 10, 12, 14, 15, 15, 15, 16, 17, 18, 20, 26

$$\text{Min} = 4, \quad Q_1 = 8.5, \quad Q_2 = 14.5, \quad Q_3 = 16.5, \quad \text{and} \quad \text{Max} = 26.$$

Boxplot : I s Buildded to Detect Outliers

1. Find Q_1 , Q_2 , Q_3 , and IQR.
2. Compute Lower Fence and Upper Fence:
Lower Inner Fence= $Q_1 - 1.5(\text{IQR})$, **Upper Inner Fence**= $Q_3 + 1.5(\text{IQR})$
Lower Outer Fence= $Q_1 - 3(\text{IQR})$, **Upper Outer Fence**= $Q_3 + 3(\text{IQR})$
3. Draw the box plot indicating the Lower an Upper fences.
4. Determine whether there are any outlier

Example 10: Use **Example 6**. Build a boxplot and check for outliers.

Defn. **Parameter:** A numerical measure for a population

Statistic: A numerical measure for a sample.

Example 11: Classify: parameter or statistic: \bar{x} , μ , σ , s

(3) **Variance:** The average squared deviation from the mean. It is used when the mean is used. That is when the distribution is symmetric.

$$\text{Population variance: } \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

An unbiased estimate of σ^2 is the sample variance.

$$\text{Sample variance: } s^2 = \frac{\sum (x_i - \bar{X})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{\sum x_i^2 - n(\bar{X})^2}{n-1}$$

Example 12:

\underline{X}_i	\underline{X}_i^2
2	4
3	9
4	16

$$\sum x_i = 9 \quad \sum x_i^2 = 29$$

$$S^2 = \frac{29 - \frac{(9)^2}{3}}{3-1} = \frac{29 - \frac{81}{3}}{2} = \frac{29 - 27}{2} = 1$$

Sample

Standard Deviation: $s = \sqrt{s^2} = \sqrt{1} = 1$.

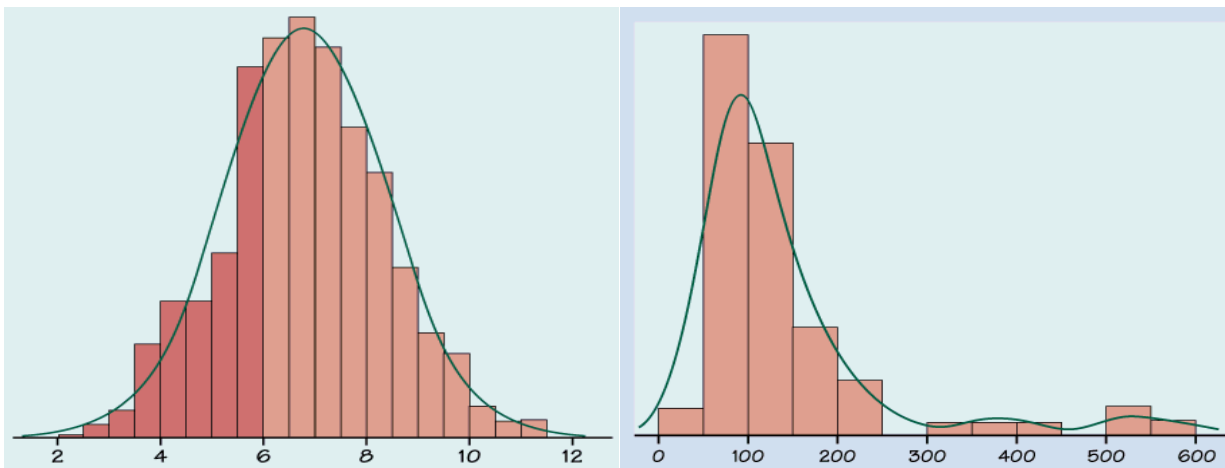
Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$

Homework (section 1.2) (Use Minitab when you can):

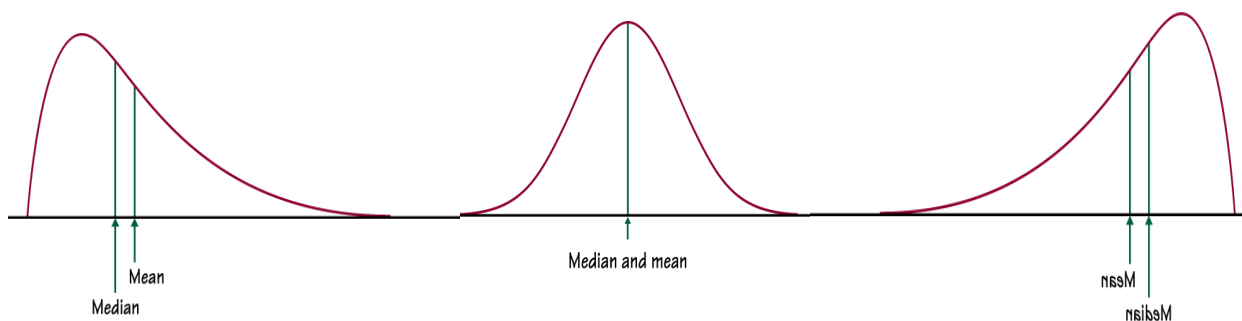
1.69, 1.75, 1.77, 1.78, 1.93 pp.47-49

SECTION 1.3

Every data set has a certain distribution. The distribution is either a Bell-Shaped Symmetric or Skewed to the right or left. How do we identify the distribution? Take a frequency Histogram or a relative frequency Histogram and draw a smooth curve over the bars. (See figures below)

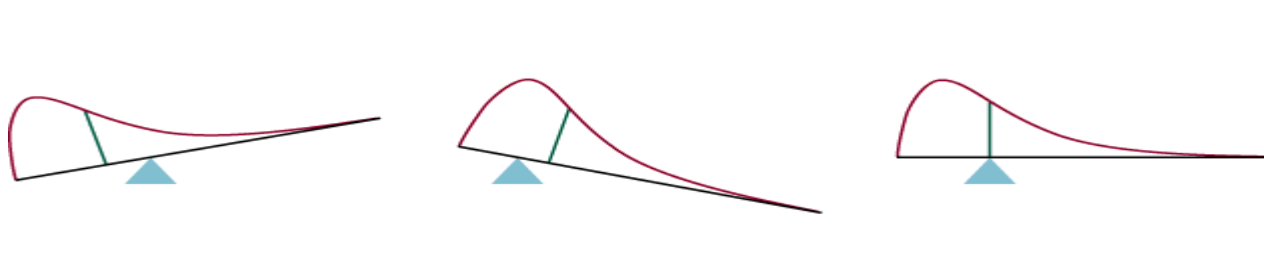


Note: The smooth curve is called density function or density curve. The area under the density curve is 1 since the sum of all relative frequencies is 1. The Bell-Shaped curves, it's a family of curves called the



Normal curves. These Normal curves are Bell-Shaped Symmetric, or Skewed to the right or left.

The mean is the point at which the curve would balance if it made of solid material.



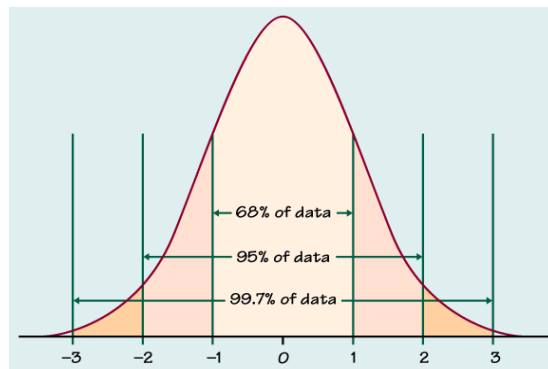
Normal Distribution

The Normal distributions are symmetric, single-peaked, bell-shaped density curves. The nice thing about the Normal distribution is that you can completely describe it by knowing the mean, μ , and the standard deviation, σ .

Empirical Rule: Applies to all bell-shaped curves.

68% fall within 1σ of the mean, μ .
 95% fall within 2σ of the mean, μ .
 99.7% fall within 3σ of the mean, μ .

If $\mu=0$ and $\sigma =1$ see figure to the right.



Do example 1.34 page 57.

Standardizing a Normal Random Variable (X)

If X is a normal random variable with mean, μ , and standard deviation, σ ; i.e $X \sim N(\mu, \sigma)$ to answer probability questions we have to standardized X . That is converting to Z . The formula to convert X to Z is:

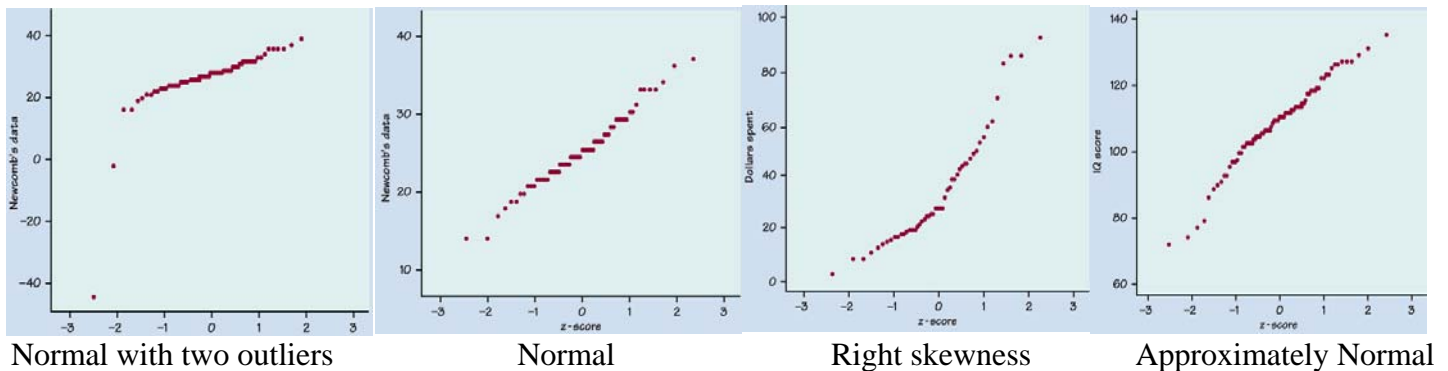
$$Z = \frac{X - \mu}{\sigma}$$
 The random variable Z is called the standard normal with mean, $\mu = 0$ and standard deviation, $\sigma = 1$ see the above figure.

Do examples 1.38, 1.39, 1.40, 1.41 pp 61-64.

Normal Quantile Plots

To see if a distribution fits the normal curve, we have to compute the quantiles for our data set and plot them against the quantiles of the standard normal distribution. If the plot is a straight line, then the data set has a normal distribution. Hence, all the rules about the normal curves apply to this data set.

Next are some Normal quantiles plots.



Left Skewness: The smallest observations fall to the left of the line drawn by the other points.

Right Skewness: The largest observations fall to the right of the line drawn by the other points.

Homework (section 1.3)

1.116, 1.117, 1.118, 1.119, 1.120, 1.122, 1.126, 1.127, 1.128, 1.129, 1.130, 1.131, 1.133, 1.135, 1.138, 1.139, 1.144, 1.145, 1.150, 1.151 pp 70-73.