CHAPTER 1

Introduction

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

Variable: A variable is **any characteristics, number, or quantity that can be measured or counted**. Age, sex, business income and expenses, country of birth, and vehicle type are examples of variables.

Categorical variable (or Qualitative 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".

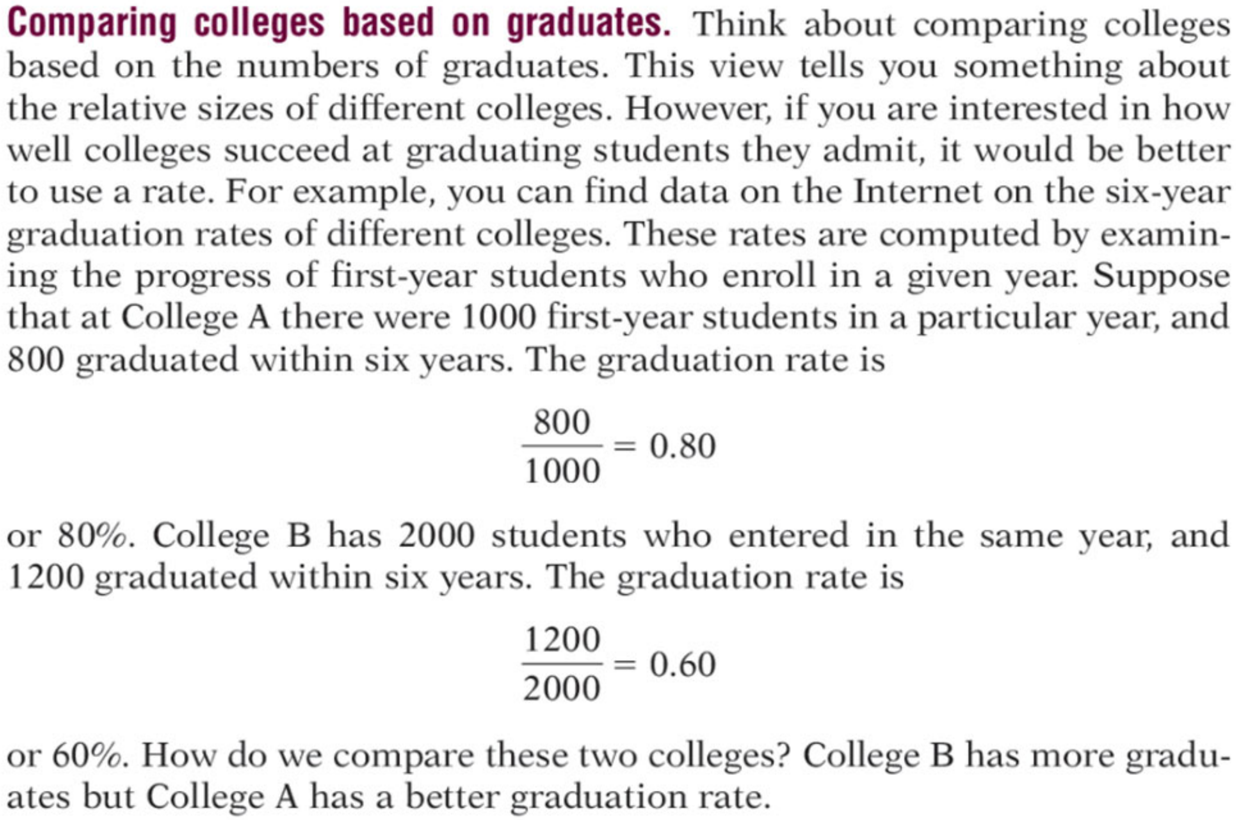
Quantitative variable: We will call any variable that takes numerical values for which arithmetic makes sense a quantitative variable.

<u>SECTION 1.1 and 1.2</u> Looking at Data/Displaying Distributions

<u>Measurement:</u> 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.

<u>EXAMPLE:</u> Example 1.6 page 6.



Comparing colleges based on graduates. Think about comparing colleges based on the numbers of graduates. This view tells you something about the relative sizes of different colleges. However, if you are interested in how well colleges succeed at graduating students they admit, it would be better to use a rate. For example, you can find data on the Internet on the six-year graduation rates of different colleges. These rates are computed by examining the progress of first-year students who enroll in a given year. Suppose that at College A there were 1000 first-year students in a particular year, and 800 graduated within six years. The graduation rate is

$$\frac{800}{1000} = 0.80$$

or 80%. College B has 2000 students who entered in the same year, and 1200 graduated within six years. The graduation rate is

$$\frac{1200}{2000} = 0.60$$

or 60%. How do we compare these two colleges? College B has more graduates but College A has a better graduation rate.

<u>Variation:</u> 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.

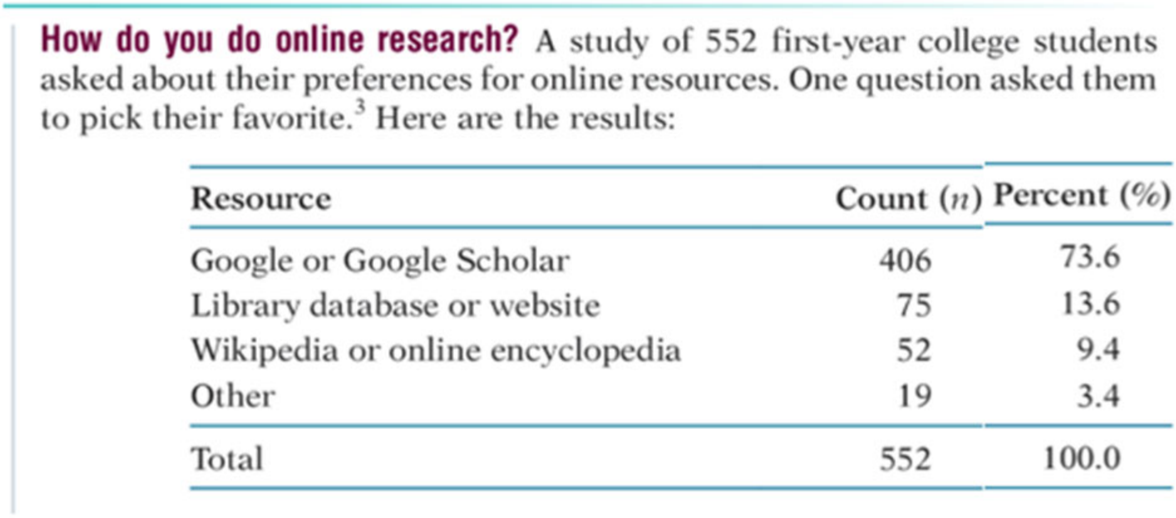
The shape of a distribution can be approximately Symmetric(each side of the center is a mirror image of the other) or Skewed(one tail extends farther from the center than the other). The number of peaks is another aspect of overall shape.

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, 1.9, and 1.10 pages 9-11.

Examples 1.7 and 1.8 (Count and Percent)

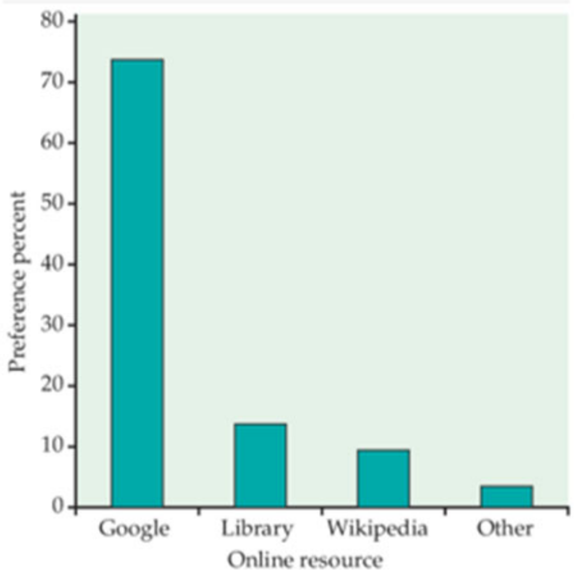


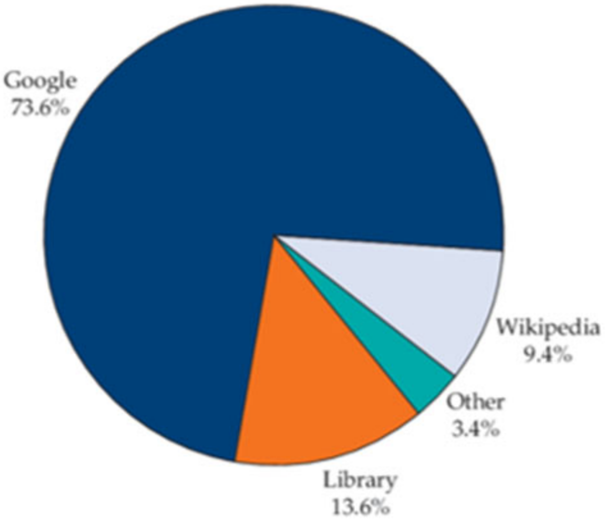
How do you do online research? A study of 552 first-year college students asked about their preferences for online resources. One question asked them to pick their favorite.[3] Here are the results:

Resource	Count (n)	Percent (%)
Google or Google Scholar	406	73.6
Library database or website	75	13.6
Wikipedia or online encyclopedia	52	9.4
Other	19	3.4
Total	552	100.0

Examples 1.9 and 1.10 (Bar Graph and Pie Chart)

Bar graph for the online resource preference data. Figure 1.2 displays the online resource preference data using a **bar graph**. The heights of the four bars show the percents of the students who reported each of the resources as their favorite.





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

The Stemplot for Babe Ruth is given to the right.

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

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

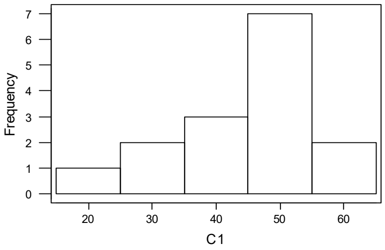
Ruth		McGwire
	0	9 9
	1	
5 2	2	2 9
5 4	3	2 2 3 9 9
9 7 6 6 6 1 1	4	2 9
9 4 4	5	2 8
0	6	5
	7	0

EXAMPLE - BABE RUTH

Stem-and -Leaf

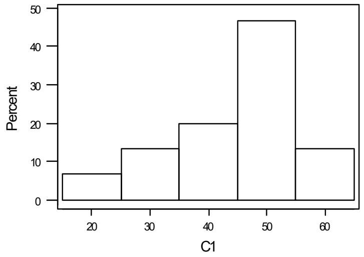
Histogram-Frequency

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



```
HR   Count CumCnt Percent CumPct
  Histogram-Percent

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
```

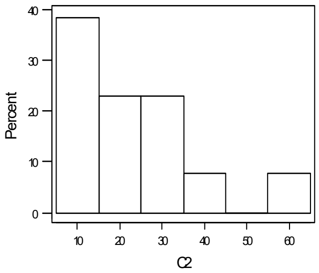


EXAMPLE- ROGER MARIS

Stem-and -Leaf

Histogram-Percent

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



Histograms

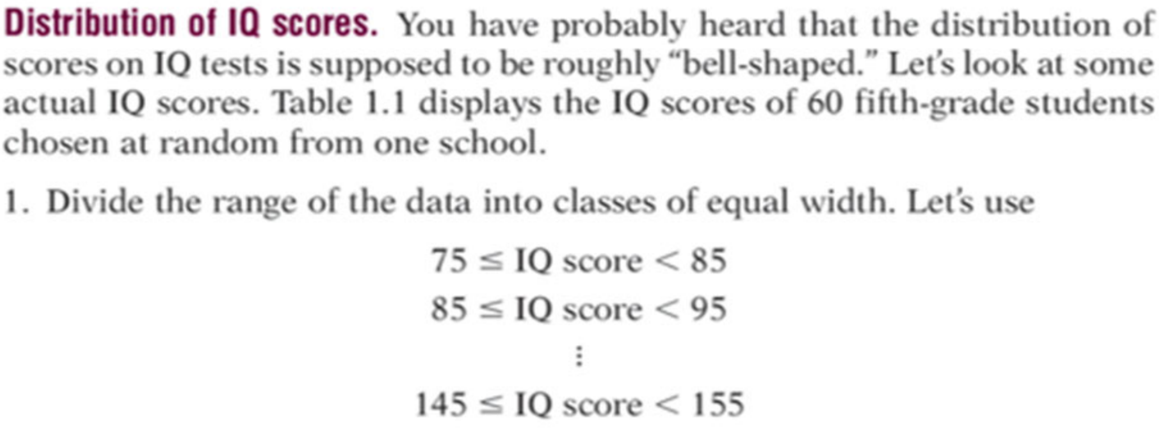
A histogram displays the count (or percent) of the observations that fall into each interval. We can choose any convenient number of intervals. Histograms are slower to construct by hand than are stemplots, and do not retain the actual values observed. The construction of a histogram is best shown by 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 1.14 page 14.

(3) **Draw the Histogram**



Distribution of IQ scores. You have probably heard that the distribution of scores on IQ tests is supposed to be roughly "bell-shaped." Let's look at some actual IQ scores. Table 1.1 displays the IQ scores of 60 fifth-grade students chosen at random from one school.

1. Divide the range of the data into classes of equal width. Let's use

$$75 \le \text{IQ score} < 85$$
$$85 \le \text{IQ score} < 95$$
$$\vdots$$
$$145 \le \text{IQ score} < 155$$

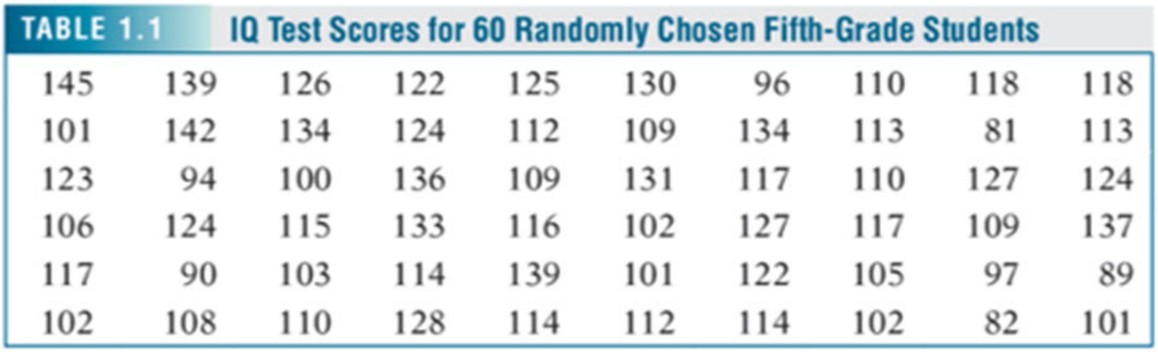
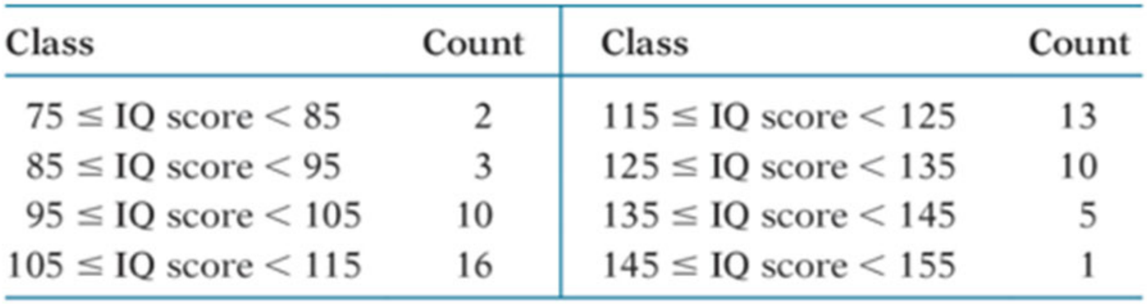


TABLE 1.1 IQ Test Scores for 60 Randomly Chosen Fifth-Grade Students

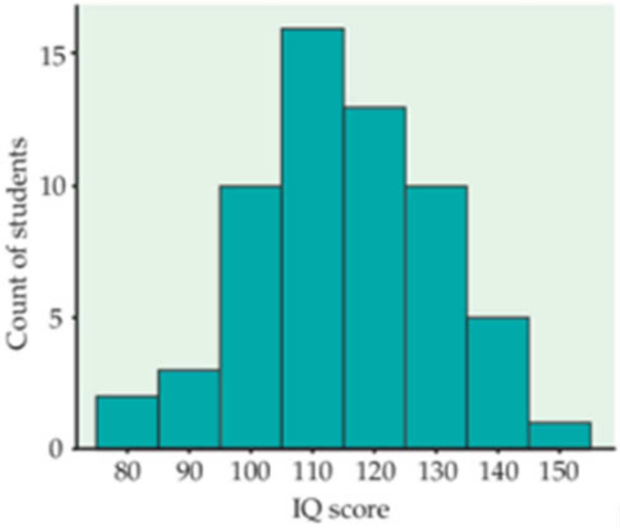
145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

Count or Frequency Distribution

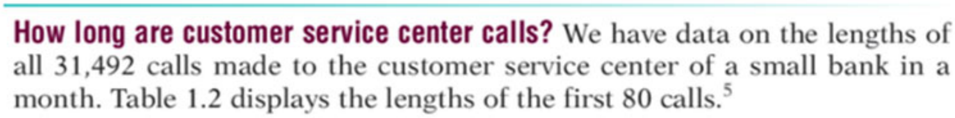


Class	Count	Class	Count
$75 \le$ IQ score < 85	2	$115 \le$ IQ score < 125	13
$85 \le$ IQ score < 95	3	$125 \le$ IQ score < 135	10
$95 \le$ IQ score < 105	10	$135 \le$ IQ score < 145	5
$105 \le$ IQ score < 115	16	$145 \le$ IQ score < 155	1

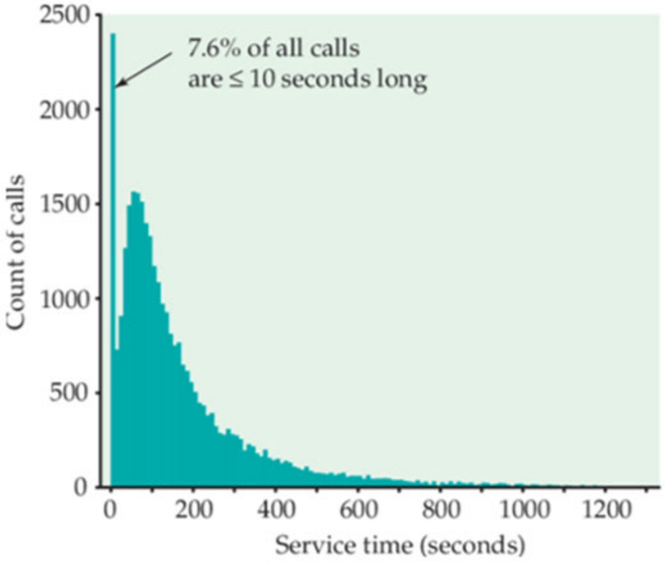
Histogram Using Count or Frequency



EXAMPLE: Look at figure 1.8 (length of phone calls) page 17



How long are customer service center calls? We have data on the lengths of all 31,492 calls made to the customer service center of a small bank in a month. Table 1.2 displays the lengths of the first 80 calls.[5]



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 and 1.2): 1.11 pp.8, 1.36, 1.39, 1.40 pp.25 (Use Minitab when you can)

SECTION 1.3 Describing Distributions with Numbers

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): $\bar{x} = \frac{\Sigma x_i}{n}$; n: sample size.

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

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

(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.

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.

Car Model	**Frequency**	
Cherv. Cavalier	9	
Ford Escort	14	
Ford Taurus	8	**Mode:** 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. Our definition of a percentile is bit inexact because there is not always a value with exactly p percent of the data at or below it.

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 .

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

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

Q_1 = First Quartile =25th percentile; Q_2 = Second Quartile =50th percentile;
Q_3 = Third Quartile =75th percentile

Q1 is the median (the middle) of the lower half of the data, and **Q3** is the median (the middle) of the upper half of the data.

(4, 5, 6, 8, 9, 10, 12, 14), | (15, 15, 15, 16, 17, 18, 20, 26).

Q1 = (8+9)/2=8.5 and **Q3** =(16+17)/2= 16.5

Example 7: For the data given in **Example 6**, find the first, second, and third quartiles are $Q_1 = 8.5$, $Q_2 = 14.5$, $Q_3 = 16.5$.

Measures of Spread.

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

Range= Largest value – Smallest value

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

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): $\text{IQR}=Q_3-Q_1$

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

Min = 4, $Q_1 = 8.5$, $Q_2 = 14.5$, $Q_3 = 16.5$, and Max = 26.

Boxplot : I s Builded 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) **<u>Variance:</u>** The average squared deviation from the mean. It is used when the mean is used. That is when the distribution is symmetric.

Population variance: $\sigma^2 = \dfrac{\Sigma(x_i - \mu)^2}{N}$

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

Sample variance: $S^2 = \dfrac{\Sigma(x_i - \bar{X})^2}{n-1} = \dfrac{\sum x_i^2 - \dfrac{(\Sigma x_i)^2}{n}}{n-1} = \dfrac{\Sigma x_i^2 - n(\bar{X})^2}{n-1}$

<u>Example 12:</u>

X_i	X_i^2
2	4
3	9
4	16

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

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

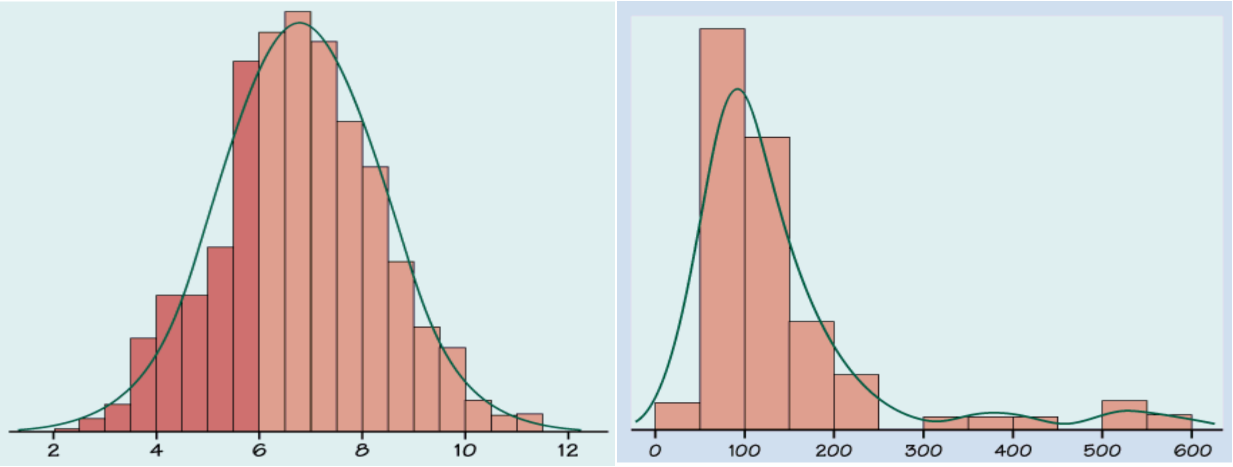
<u>Sample Standard Deviation:</u> $s = \sqrt{s^2} = \sqrt{1} = 1$.

<u>Population Standard Deviation:</u> $\sigma = \sqrt{\sigma^2}$

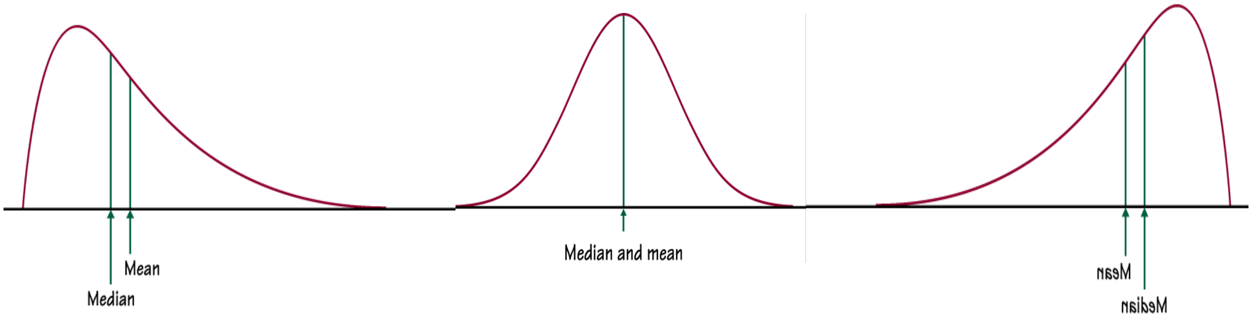
<u>Homework (section 1.3)</u> **(Use Minitab when you can):**
1.68, 1.75, 1.77, 1.79, 1.80, 187 pp.49-50

SECTION 1.4

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.



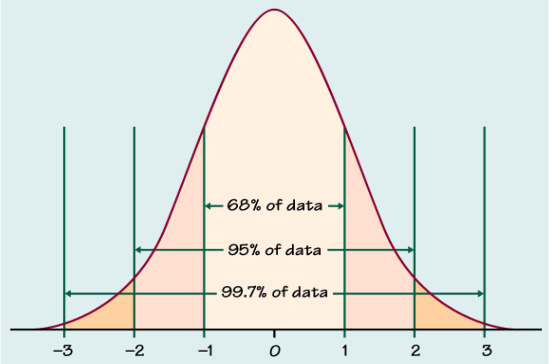
Normal Distribution

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



Standardizing a Normal Random Variable (X)

If X is a normal random variable with mean, μ , and standard deviation, σ ; i.e X~N(μ , σ) to answer probability questions we have to standardized X.

That is converting to Z. The formula to convert X to Z is: $Z = \dfrac{X-\mu}{\sigma}$.

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

Table entry for *z* is the area under the standard Normal curve to the left of *z*.

TABLE A Standard Normal Probabilities

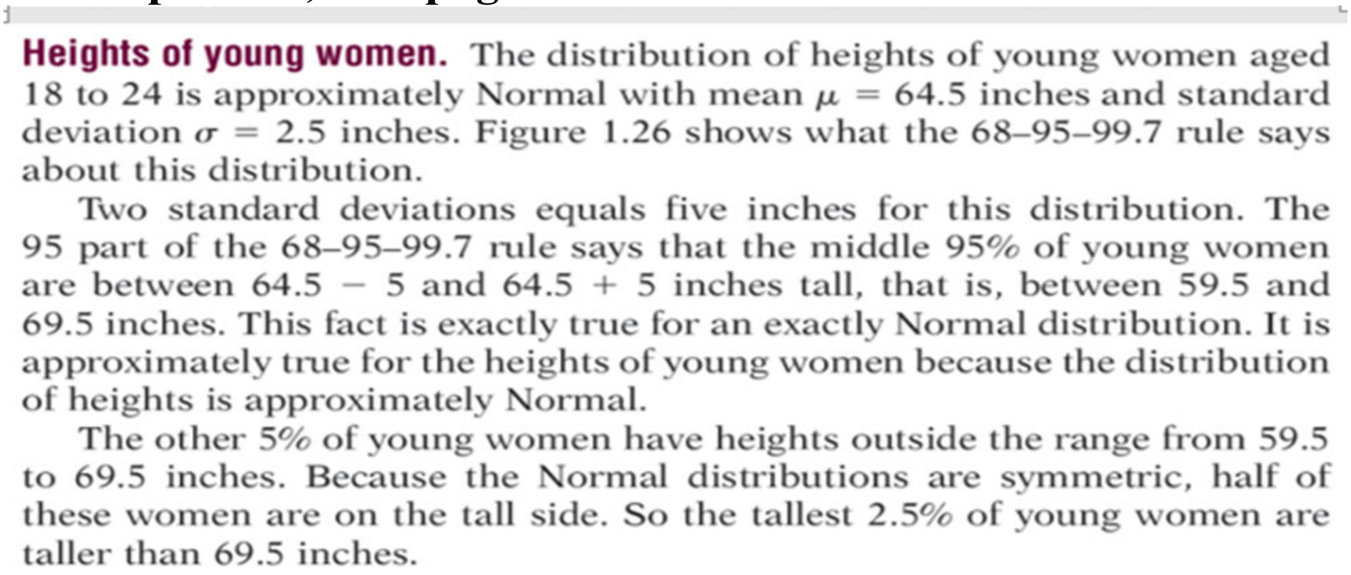
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table entry for *z* is the area under the standard Normal curve to the left of *z*.

TABLE A Standard Normal Probabilities (continued)

z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

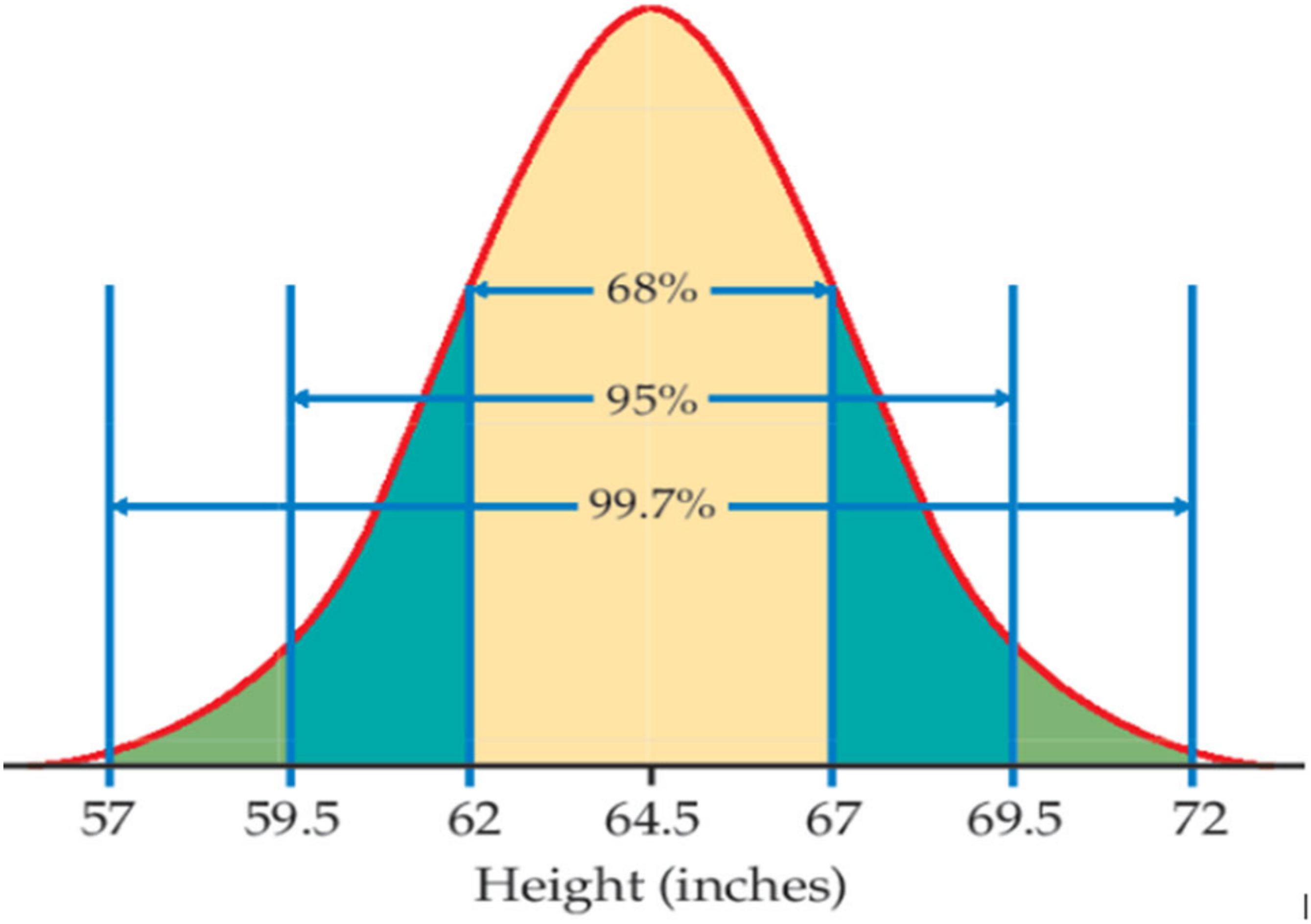
Do example 1.38, 1.39 pages 58-59.

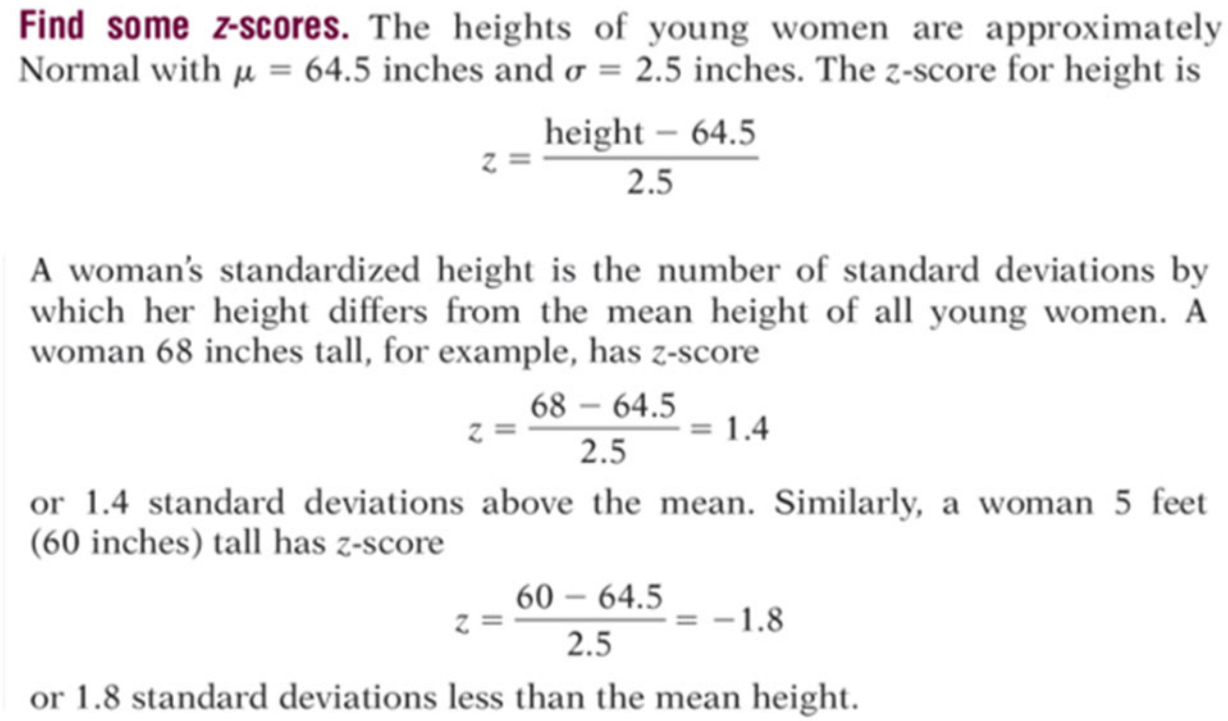


Heights of young women. The distribution of heights of young women aged 18 to 24 is approximately Normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches. Figure 1.26 shows what the 68–95–99.7 rule says about this distribution.

Two standard deviations equals five inches for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of young women are between 64.5 − 5 and 64.5 + 5 inches tall, that is, between 59.5 and 69.5 inches. This fact is exactly true for an exactly Normal distribution. It is approximately true for the heights of young women because the distribution of heights is approximately Normal.

The other 5% of young women have heights outside the range from 59.5 to 69.5 inches. Because the Normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches.





Find some *z*-scores. The heights of young women are approximately Normal with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. The z-score for height is

$$z = \frac{\text{height} - 64.5}{2.5}$$

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 68 inches tall, for example, has z-score

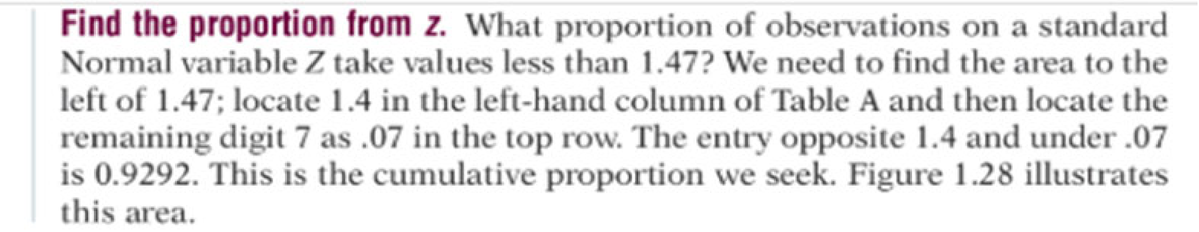
$$z = \frac{68 - 64.5}{2.5} = 1.4$$

or 1.4 standard deviations above the mean. Similarly, a woman 5 feet (60 inches) tall has z-score

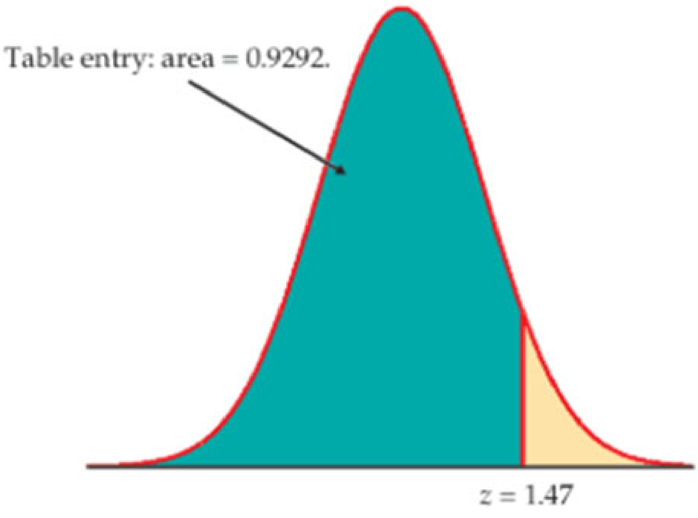
$$z = \frac{60 - 64.5}{2.5} = -1.8$$

or 1.8 standard deviations less than the mean height.

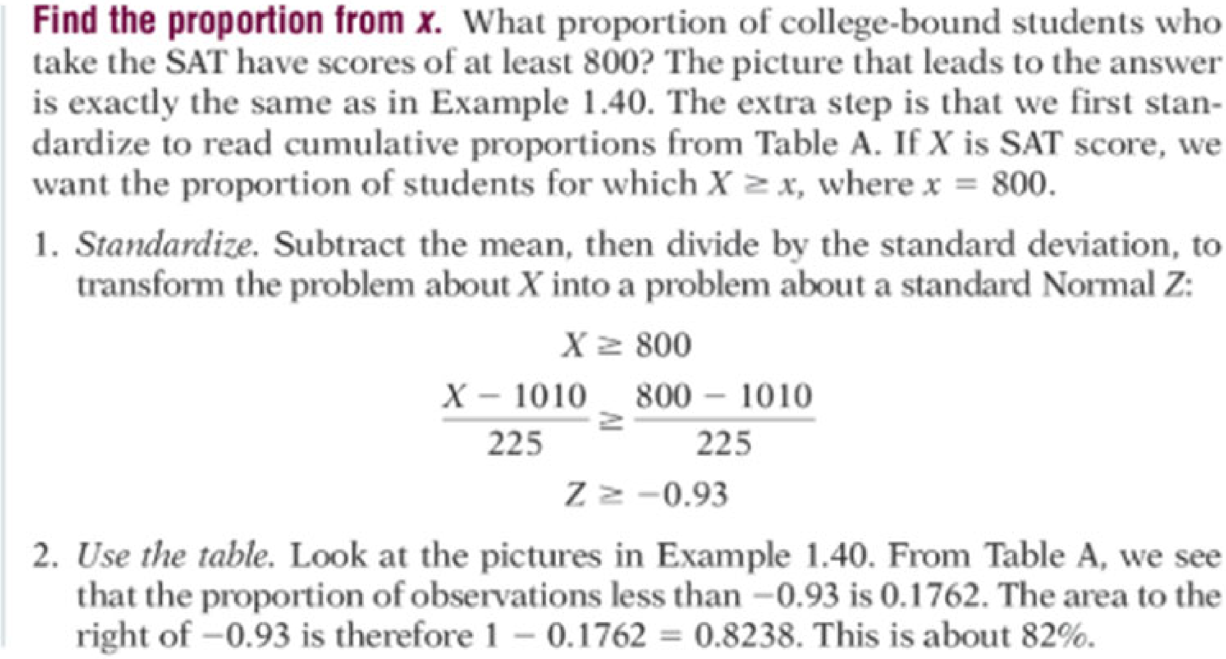
Example 1.42 pp 63



Find the proportion from *z*. What proportion of observations on a standard Normal variable Z take values less than 1.47? We need to find the area to the left of 1.47; locate 1.4 in the left-hand column of Table A and then locate the remaining digit 7 as .07 in the top row. The entry opposite 1.4 and under .07 is 0.9292. This is the cumulative proportion we seek. Figure 1.28 illustrates this area.



Example 1.43 pp 63



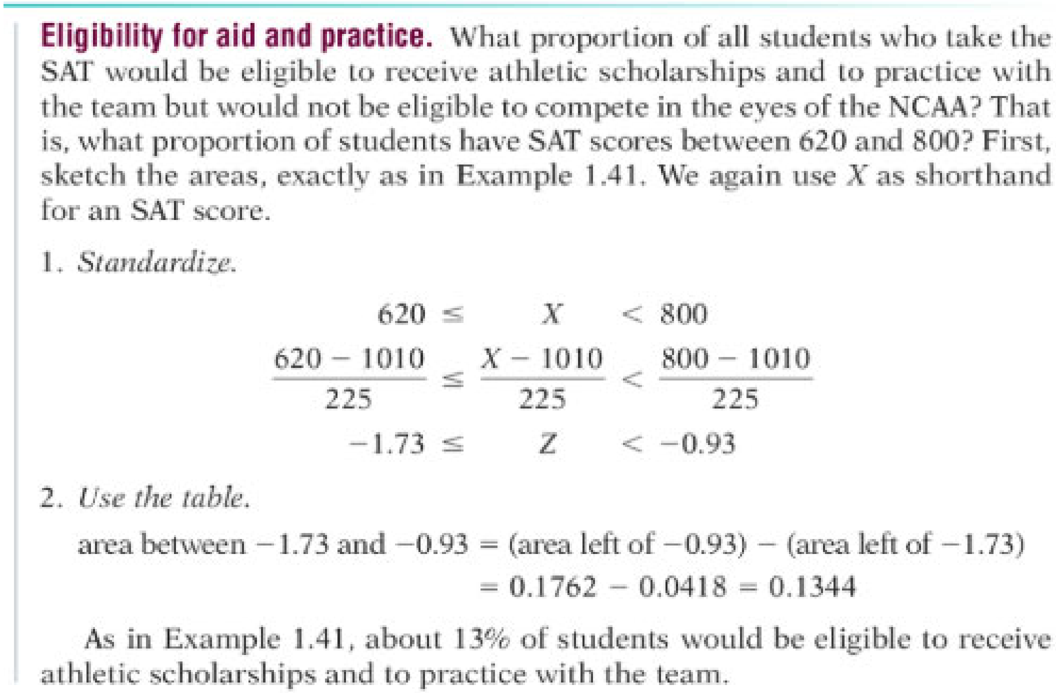
Find the proportion from *x*. What proportion of college-bound students who take the SAT have scores of at least 800? The picture that leads to the answer is exactly the same as in Example 1.40. The extra step is that we first standardize to read cumulative proportions from Table A. If X is SAT score, we want the proportion of students for which $X \geq x$, where $x = 800$.

1. *Standardize.* Subtract the mean, then divide by the standard deviation, to transform the problem about X into a problem about a standard Normal Z:

$$X \geq 800$$
$$\frac{X - 1010}{225} \geq \frac{800 - 1010}{225}$$
$$Z \geq -0.93$$

2. *Use the table.* Look at the pictures in Example 1.40. From Table A, we see that the proportion of observations less than -0.93 is 0.1762. The area to the right of -0.93 is therefore $1 - 0.1762 = 0.8238$. This is about 82%.

Example 1.44 pp 64



Eligibility for aid and practice. What proportion of all students who take the SAT would be eligible to receive athletic scholarships and to practice with the team but would not be eligible to compete in the eyes of the NCAA? That is, what proportion of students have SAT scores between 620 and 800? First, sketch the areas, exactly as in Example 1.41. We again use X as shorthand for an SAT score.

1. *Standardize.*

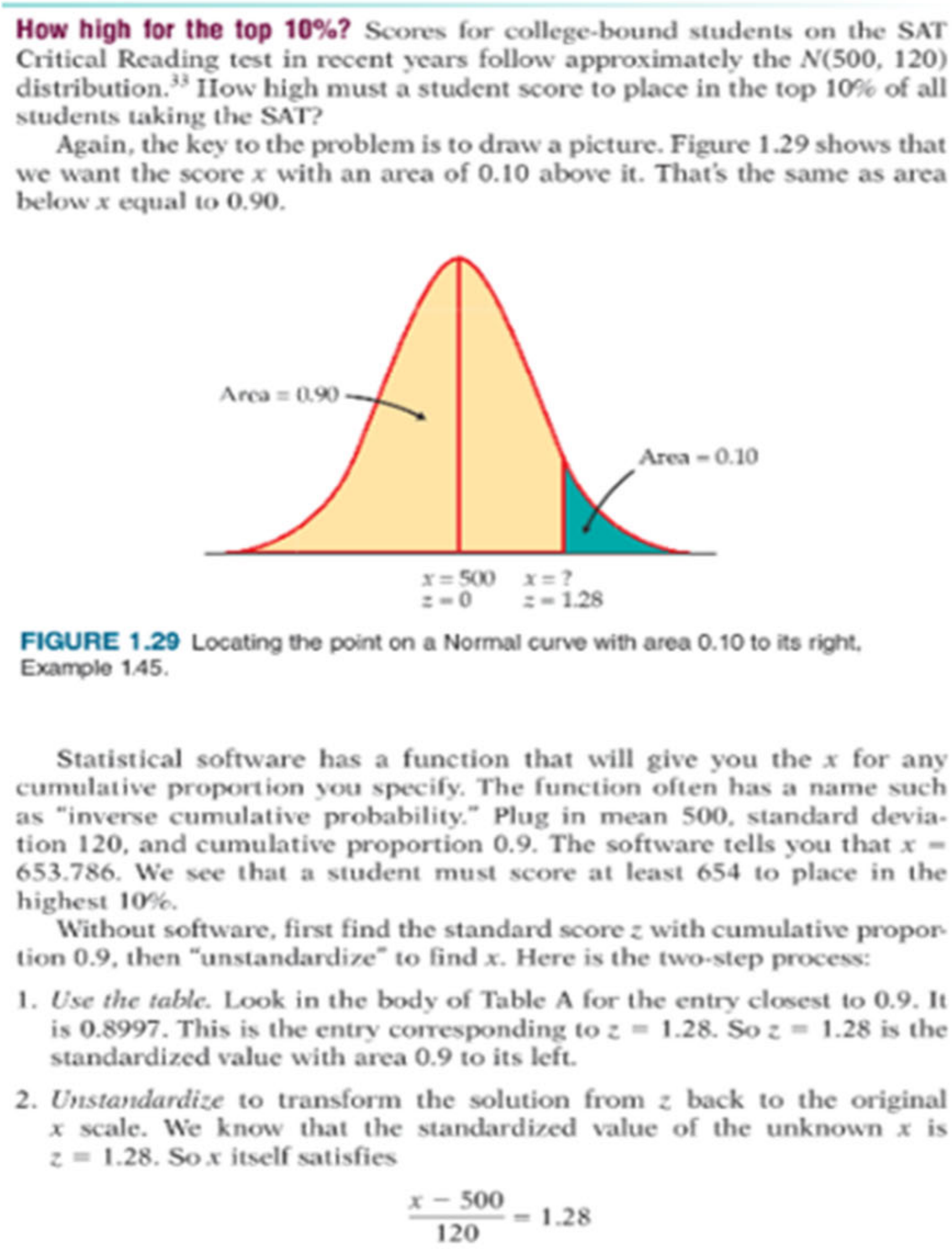
$$\begin{array}{ccccc} 620 & \leq & X & < & 800 \\ \frac{620 - 1010}{225} & \leq & \frac{X - 1010}{225} & < & \frac{800 - 1010}{225} \\ -1.73 & \leq & Z & < & -0.93 \end{array}$$

2. *Use the table.*

$$\begin{aligned} \text{area between } -1.73 \text{ and } -0.93 &= (\text{area left of } -0.93) - (\text{area left of } -1.73) \\ &= 0.1762 - 0.0418 = 0.1344 \end{aligned}$$

As in Example 1.41, about 13% of students would be eligible to receive athletic scholarships and to practice with the team.

Example 1.45 pp 65



How high for the top 10%? Scores for college-bound students on the SAT Critical Reading test in recent years follow approximately the $N(500, 120)$ distribution.[33] IIow high must a student score to place in the top 10% of all students taking the SAT?

Again, the key to the problem is to draw a picture. Figure 1.29 shows that we want the score x with an area of 0.10 above it. That's the same as area below x equal to 0.90.

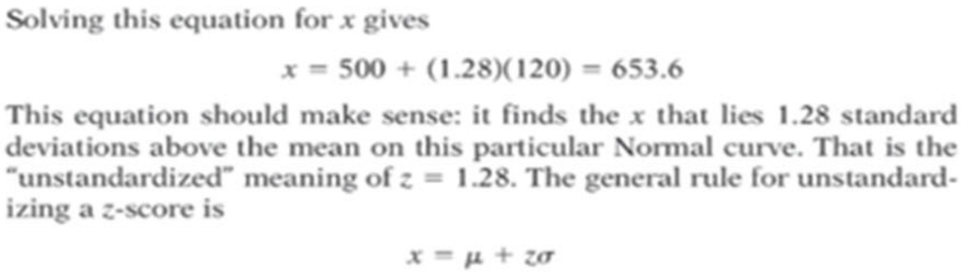
FIGURE 1.29 Locating the point on a Normal curve with area 0.10 to its right, Example 1.45.

Statistical software has a function that will give you the x for any cumulative proportion you specify. The function often has a name such as "inverse cumulative probability." Plug in mean 500, standard deviation 120, and cumulative proportion 0.9. The software tells you that $x = 653.786$. We see that a student must score at least 654 to place in the highest 10%.

Without software, first find the standard score z with cumulative proportion 0.9, then "unstandardize" to find x. Here is the two-step process:

1. *Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to $z = 1.28$. So $z = 1.28$ is the standardized value with area 0.9 to its left.
2. *Unstandardize* to transform the solution from z back to the original x scale. We know that the standardized value of the unknown x is $z = 1.28$. So x itself satisfies

$$\frac{x - 500}{120} = 1.28$$



Solving this equation for x gives

$$x = 500 + (1.28)(120) = 653.6$$

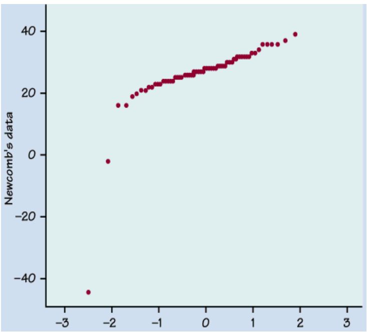
This equation should make sense: it finds the x that lies 1.28 standard deviations above the mean on this particular Normal curve. That is the "unstandardized" meaning of $z = 1.28$. The general rule for unstandardizing a z-score is

$$x = \mu + z\sigma$$

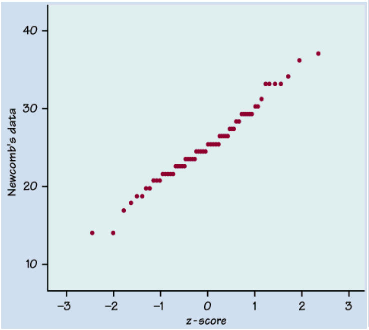
Normal Quantile Plots

To see if a distribution fits the normal curve, we must 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.

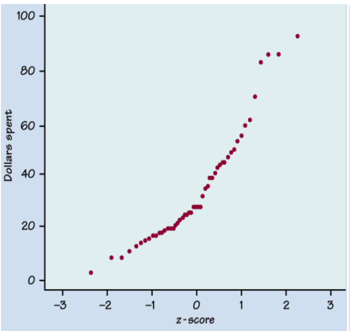
Next are some Normal quantiles plots.



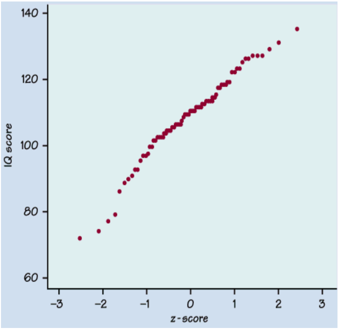
Approximately Normal with 2 outliers



Approximately Normal



Right skewness



Approximately Normal

<u>Left Skewness:</u> The smallest observations fall to the left of the line draw by the other points.

<u>Right Skewness:</u> The largest observations fall to the right of the line draw by the other points.

Homework (section 1.4)

1.112, 1.113, 1.114, 1.115, 1.118, 1.119, 1.120, 1.121, 1.122, 1.123, 1.125, 1.126, 1.127, 1.130, 1.131, 1.134, 1.136, pp. 71-73.