Contents

DATA

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. All the data collected in a particular study are referred to as the **data set** for the study. **Table 1.1** shows a data set containing information for 25 mutual funds that are part of the Morningstar Funds500. Morningstar is a company that tracks over 7000 mutual funds and prepares in-depth analyses of 2000 of these. Their recommendations are followed closely by financial analysts and individual investors.

TABLE 1.1

Fund Name	Fund Type	NAV ($)	5YR Avg. Return(%)	Expense Ratio(%)	MorningStar Rank
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Brown Cap Small	DE	35.73	15.85	1.2	4-Star
DFA U.S. Micro Cap	DE	13.47	17.23	0.53	3-Star
Fidelity Contrafund	DE	73.11	17.99	0.89	5-Star
Fidelity Overseas	IE	48.39	23.46	0.9	4-Star
Fidelity Sel Electronics	DE	45.6	13.5	0.89	3-Star
Fidelity Sh-Term Bond	FI	8.6	2.76	0.45	3-Star
Gabelli Asset AAA	DE	49.81	16.7	1.36	4-Star
Kalmar Gr Val Sm Cp	DE	15.3	15.31	1.32	3-Star
Marsico 21st Century	DE	17.44	15.16	1.31	5-Star
Mathews Pacific Tiger	IE	27.86	32.7	1.16	3-Star
Oakmark I	DE	40.37	9.51	1.05	2-Star
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3-Star
RS Value A	DE	26.27	23.68	1.36	4-Star
T. Rowe Price Latin Am.	IE	53.89	51.1	1.24	4-Star
T. Rowe Price Mid Val	DE	22.46	16.91	0.8	4-Star
Thornburg Value A	DE	37.53	15.46	1.27	4-Star
USAA Income	FI	12.1	4.31	0.62	3-Star
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4-Star
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3-Star
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3-Star
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4-Star

ELEMENTS, VARIABLES, AND OBSERVATIONS

Elements are the entities on which data are collected. For the data set in Table 1.1 each individual mutual fund is an element: the element names appear in the first column. With 25 mutual funds, the data set contains 25 elements.

A **variable** is a characteristic of interest for the elements. The data set in Table 1.1 includes the following five variables:

- **Fund Type**: The type of mutual fund, labeled DE (Domestic Equity), IE (International Equity), and FI (Fixed Income)
- **Net Asset Value ($)**: The closing price per share.
- **5-Year Average Return (%)**: The average annual return for the fund over the past 5 years
- **Expense Ratio**: The percentage of assets deducted each fiscal year for fund expenses
- **Morningstar Rank**: The overall risk-adjusted star rating for each fund; Morningstar ranks go from a low of 1-Star to a high of 5-Stars

Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular element is called an **observation**. Referring to Table 1.1 we see that the set of measurements for the first observation (American Century Intl. Disc) is IE, 14.37, 30.53, 1.41, and 3-Star. The set of measurements for the second observation (American Century Tax-Free Bond) is FI, 10.73, 3.34, 0.49, and 4-Star, and so on. A data set with 25 elements contains 25 observations.

SAMPLES AND POPULATIONS

In statistics we make a distinction between two concepts: **A POPULATION AND A SAMPLE**:

***Population:* "*The* population *consists of the set of all measurements in which the investigator is interested. The population is also called the universe.*"**

***Sample:* "*A* sample *is a subset of measurements selected from the population. Sampling from the population is often done randomly, such that every possible sample of* n *elements will have an equal chance of being selected. A sample selected in this way is called a* simple random sample*, or just a* random sample*. A random sample allows chance to determine its elements.*"**

Example:

*Farmer Jane owns **1,264 sheep**. These sheep constitute her entire population of sheep. If 15 sheep are selected to be sheared, then these 15 represent a sample from Jane's population of sheep. Further, if the 15 sheep were selected at random from Jane's population of 1,264 sheep, then they would constitute a random sample of sheep.*

The definitions of sample and population are relative to what we want to consider. If Jane's sheep are all we care about, then they constitute a population. If, however, we are interested in all the sheep in the county, then all Jane's 1,264 sheep are a sample of that larger population (although this sample would not be random).

SCALES OF MEASUREMENT

Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio.

When the data for a variable consist of labels or names used to identify an attribute of the element, the scale of measurement is considered a **NOMINAL SCALE**. For example, referring to the data in Table 1.1, we see that the scale of measurement for the Fund Type variable is nominal because DE, IE, and FI are labels used to identify the category or type of fund. In cases where the scale of measurement is nominal, a numeric code as well as nonnumeric labels may be used. For example, to facilitate data collection and to prepare the data for entry into a computer database, we might use a numeric code by letting 1 denote Domestic Equity, 2 denote International Equity, and 3 denote Fixed Income. In this case the numeric values 1, 2, and 3 identify the category of fund. The scale of measurement is nominal even though the data appear as numeric values.

The scale of measurement for a variable is called an **ORDINAL SCALE** if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. However, the magnitude between successive values is not known, such as ratings from 1 to 5, shirt sizes, grades, or army rankings. As example, note that the Morningstar Rank for the data in Table 1.1 is ordinal data. It provides a rank from 1 to 5-Stars based on Morningstar's assessment of the fund's risk-adjusted return. Ordinal data can also be provided using a numeric code, for example, your class rank in school.

The scale of measurement for a variable is an **INTERVAL SCALE** if the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric. Scholastic Aptitude Test (SAT) scores are an example of interval-scaled data. For example, three students with SAT math scores of 620, 550, and 470 can be ranked or ordered in terms of best performance to poorest performance. In addition, the differences between the scores are meaningful. For instance, student 1 scored 620- 550 = 70 points more than student 2, while student 2 scored 550 - 470 = 80 points more than student 3.

The scale of measurement for a variable is a **RATIO SCALE** if the data have all the properties of interval data and the ratio of two values is meaningful. Variables such as distance, height, weight, and time use the ratio scale of measurement. For example, consider the cost of an automobile. A zero value for the cost would indicate that the automobile has no cost and is free. In addition, if we compare the cost of $30,000 for one automobile to the cost of $15,000 for a second automobile, the ratio property shows that the first automobile is $30,000/$15,000 = 2 times, or twice, the cost of the second automobile.

INTERVAL SCALE Vs RATIO SCALE

Interval scale and ratio scale are the two variable measurement scales where they define the attributes of the variables quantitatively. The difference between interval and ratio scales is that, while interval scales are void of absolute or true zero for example temperature can be below 0 degree Celsius (-10 or -20), ratio scales have a true zero value, for example, height or weight it will always be measured between 0 to maximum but never below 0.

TYPES OF DATA

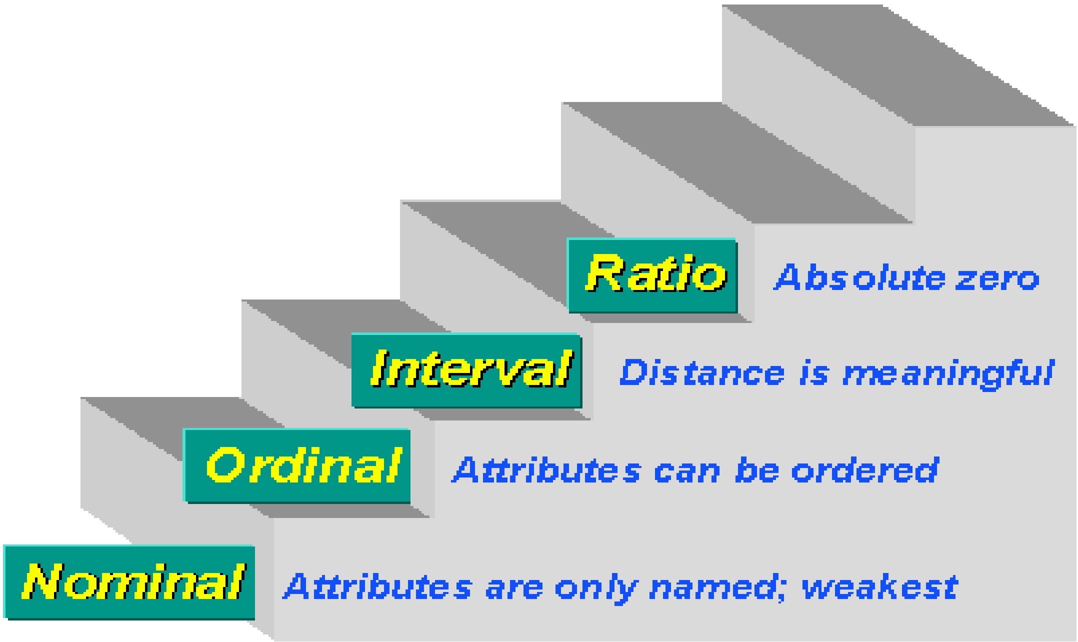
CATEGORICAL DATA (OR QUALITATIVE) DATA record qualities or characteristics about the individual, such as eye colour, gender, political party, or opinion on some issue (using categories such as agree, disagree, or no opinion). It is also called **NOMINAL DATA.** Categorical data use either the nominal or ordinal scale of measurement.

BINARY DATA is a particular kind of NOMINAL DATA with only 2 values/states (0 and 1) such as gender, medical test report (Positive Or Negative)

NUMERICAL DATA (OR QUANTITATIVE DATA) record measurements or counts regarding each individual, which may include weight, age, height, or time to take an exam; counts may include number of pets, or the number of red lights you hit on your way to work. Quantitative data are obtained using either the interval or ratio scale of measurement.

QUICK RECAP

Scale	Type of Data	Definition	Examples
Nominal	Categorical	Categorizes but does not rank	Industries, Gender, Occupation
Ordinal	Categorical	Ranked Categories. Differences between ranks not consistent	Organizational Hierarchy. Star Ratings
Interval	Numerical	Ranks Data. Differences between ranks equal. No True Zero Point.	Celsius or Fahrenheit Scale. Dates
Ratio	Numerical	Ranks Data. Differences between ranks equal. Also has a True Zero Point.	Rate of Return, Money



CONTINUOUS VS DISCRETE DATA

Quantitative data may be **discrete or continuous**. Quantitative data that measure ***how many (e.g., number of calls received in 5 minutes)*** are discrete. Quantitative data that measure ***how much (e.g., weight or time)*** are continuous because no separation occurs between the possible data values.

EXERCISE 1

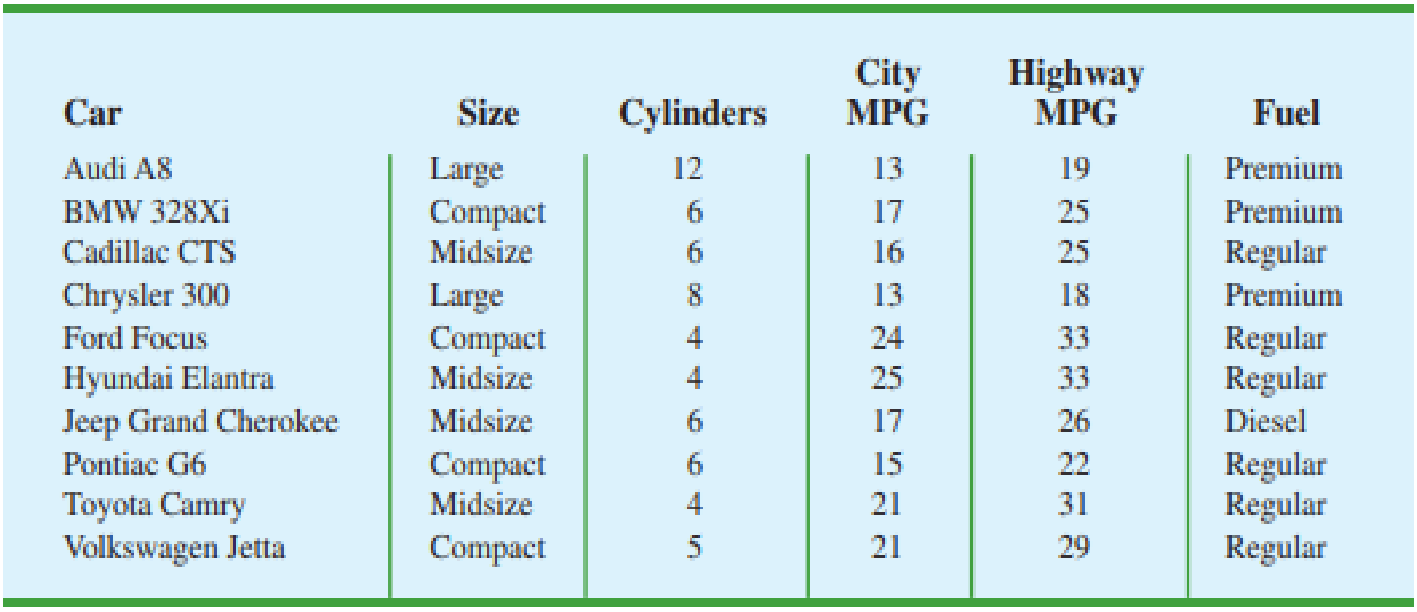
The U.S. Department of Energy provides fuel economy information for a variety of motor vehicles. A sample of 10 automobiles is shown in Table below. Data show the size of the automobile (compact, midsize, or large), the number of cylinders in the engine, the city driving miles per gallon, the highway driving miles per gallon, and the recommended fuel (diesel, premium, or regular).

a. How many elements are in this data set?

b. How many variables are in this data set?

c. Which variables are categorical and which variables are quantitative?

d. What type of measurement scale is used for each of the variables?



Car	Size	Cylinders	City MPG	Highway MPG	Fuel
Audi A8	Large	12	13	19	Premium
BMW 328Xi	Compact	6	17	25	Premium
Cadillac CTS	Midsize	6	16	25	Regular
Chrysler 300	Large	8	13	18	Premium
Ford Focus	Compact	4	24	33	Regular
Hyundai Elantra	Midsize	4	25	33	Regular
Jeep Grand Cherokee	Midsize	6	17	26	Diesel
Pontiac G6	Compact	6	15	22	Regular
Toyota Camry	Midsize	4	21	31	Regular
Volkswagen Jetta	Compact	5	21	29	Regular

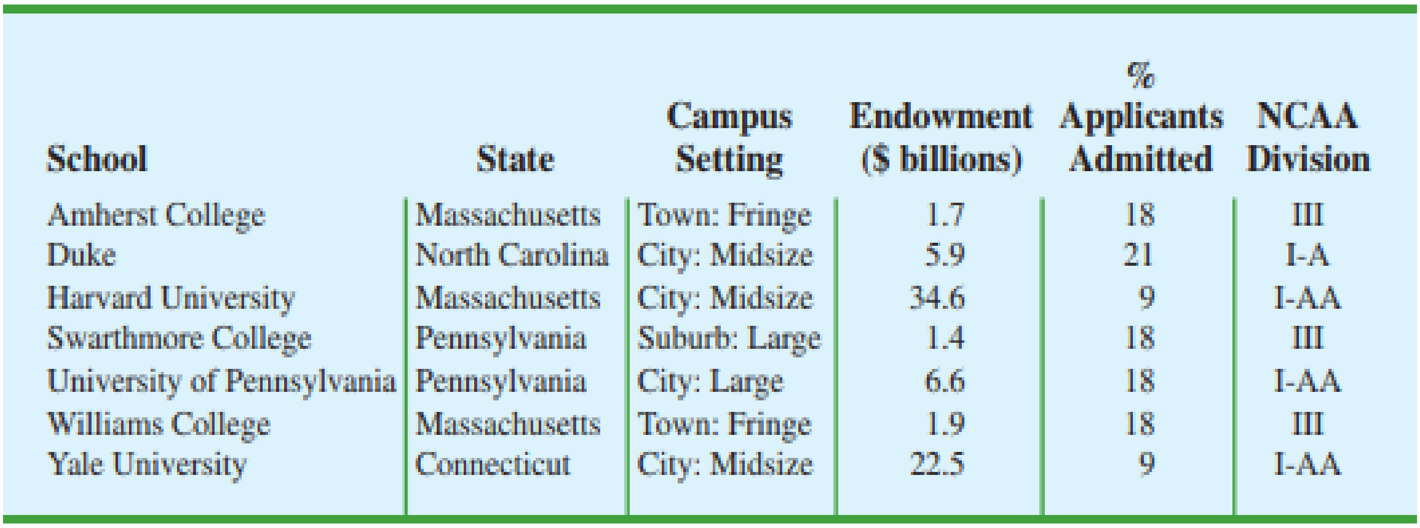
EXERCISE 2

Table below shows data for seven colleges and universities. The endowment (in billions of dollars) and the percentage of applicants admitted are shown. The state each school is located in, the campus setting, and the NCAA Division for varsity teams were obtained from the National Center of Education Statistics website, February 22, 2008.

a. How many elements are in the data set?

b. How many variables are in the data set?

c. Which of the variables are categorical and which are quantitative?



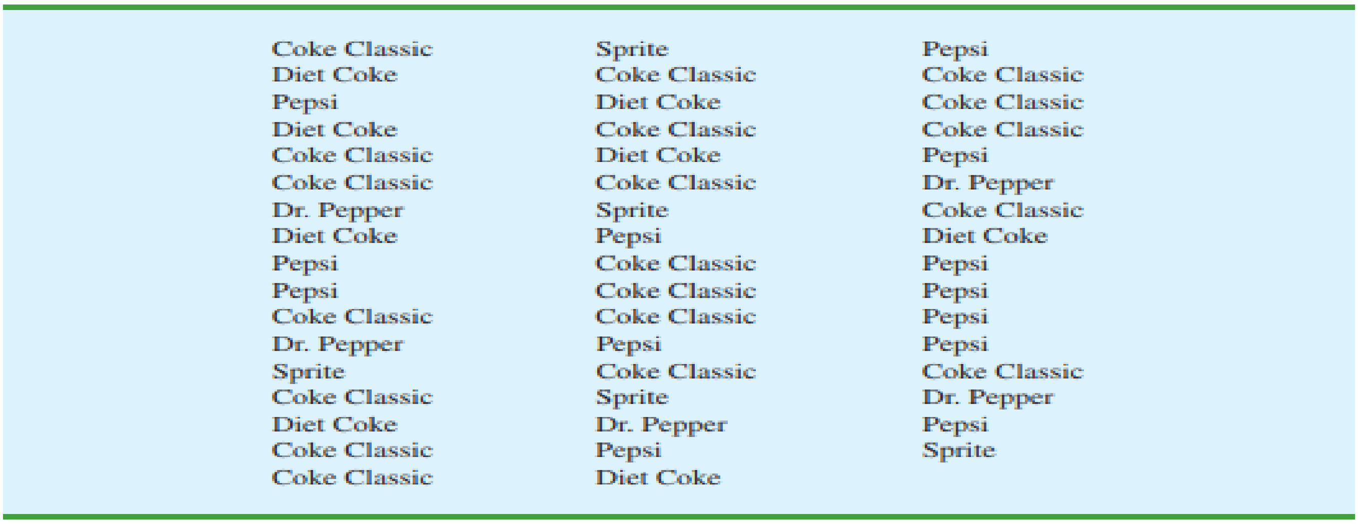
School	State	Campus Setting	Endowment ($ billions)	% Applicants Admitted	NCAA Division
Amherst College	Massachusetts	Town: Fringe	1.7	18	III
Duke	North Carolina	City: Midsize	5.9	21	I-A
Harvard University	Massachusetts	City: Midsize	34.6	9	I-AA
Swarthmore College	Pennsylvania	Suburb: Large	1.4	18	III
University of Pennsylvania	Pennsylvania	City: Large	6.6	18	I-AA
Williams College	Massachusetts	Town: Fringe	1.9	18	III
Yale University	Connecticut	City: Midsize	22.5	9	I-AA

FREQUENCY DISTRIBUTION

A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes. In other words, is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.

In Frequency distribution, we find the number of counts for a particular observation when the observations are repeated.

Let us use the following example to demonstrate the construction and interpretation of a frequency distribution for categorical data. Coke Classic, Diet Coke, Dr. Pepper, Pepsi, and Sprite are five popular soft drinks. Assume that the data in Table 2.1 show the soft drink selected in a sample of 50 soft drink purchases.



Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

To develop a frequency distribution for these data, we count the number of times each soft drink appears in Table above. Coke Classic appears 19 times, Diet Coke appears 8 times, Dr. Pepper appears 5 times, Pepsi appears 13 times, and Sprite appears 5 times. These counts are summarized in the frequency distribution below.

Soft Drink	Frequency
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	**50**

FREQUENCY DISTRIBUTION IN EXCEL

- Enter the data in column and sort it in ascending order by Data Sort
- Note the minimum & maximum values and prepare the bins of appropriate sizes e.g , suppose the data ranges from 0 to 100 then the bins will be (1-10,11-21,…,91-100).
- Now create a table with column headers 'Bins' and 'Frequency' and type in the bin ranges.
- Use the countif function to fill in the frequency column for each of the bin.

For the first bin:

countif([select data], "[< starting value of second bin]")

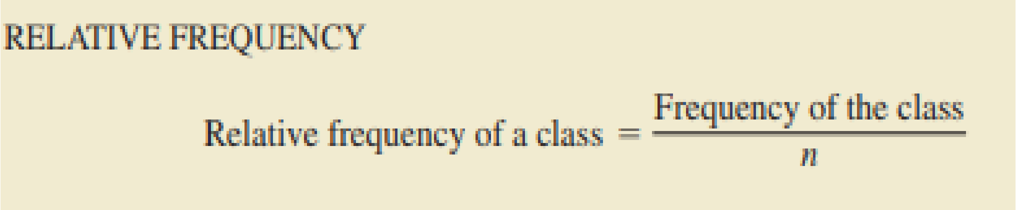
For rest of the bins:

countif([select data], "[< starting value of next highest bin]") –

countif([select data], "[< starting value of selected bin]")

RELATIVE FREQUENCY AND PERCENT FREQUENCY DISTRIBUTIONS

The **Relative Frequency** of a class equals the fraction or proportion of items belonging to a class. For a data set with n observations, the relative frequency of each class can be determined as follows:



RELATIVE FREQUENCY

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n}$$

The **Percent Frequency** of a class is the relative frequency multiplied by 100.

Soft Drink	Frequency	RELATIVE FREQUENCY	PERCENT FREQUENCY
Coke Classic	19	0.38	38
Diet Coke	8	0.16	16
Dr. Pepper	5	0.1	10
Pepsi	13	0.26	26
Sprite	5	0.1	10
Total	**50**		

GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION

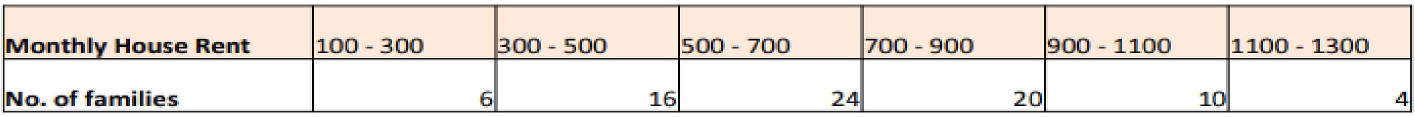
It is often useful to represent a frequency distribution by means of a diagram which makes the unwieldy data intelligible and conveys to the eye the general run of the observations. Diagrammatic representation also facilitates the comparison of two or more frequency distributions. We consider below some important types of graphic representation:

HISTOGRAM

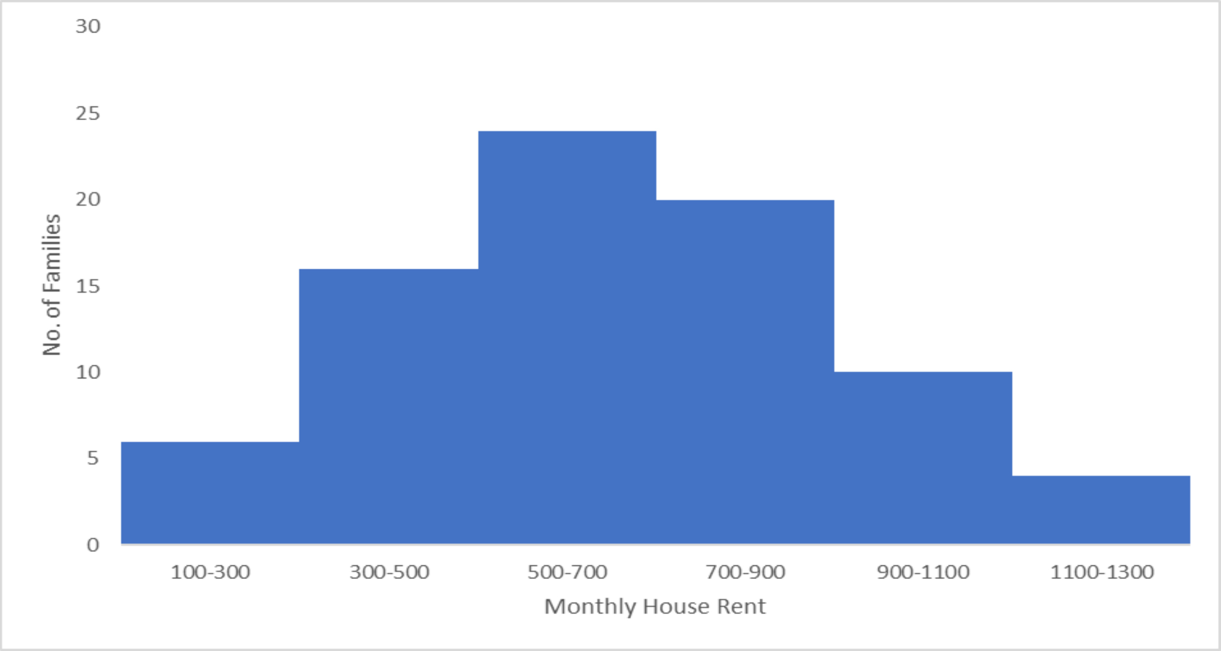
- Histogram is a graph used for the representation of the frequency distribution.
- It is series of adjacent rectangles erected on X-axis with class interval as base and Frequency on the Y-axis.
- Histograms are useful to find mode and understand the spread of the distribution (which will be discussed later).

A histogram is basically a bar graph that applies to numerical data. Because the data are numerical, the categories are ordered from smallest to largest (as opposed to categorical data, such as gender, which has no inherent order to it). To be sure each number falls into exactly one group, the bars on a histogram touch each other but don’t overlap. Each bar is marked on the x-axis (horizontal) by the values representing its beginning and endpoints. The height of each bar of a histogram represents either the number of individuals in each group (the frequency of each group) or the percentage of individuals in each group (the relative frequency of each group).

Example: Consider the following data to plot the histogram



Monthly House Rent	100 - 300	300 - 500	500 - 700	700 - 900	900 - 1100	1100 - 1300
No. of families	6	16	24	20	10	4



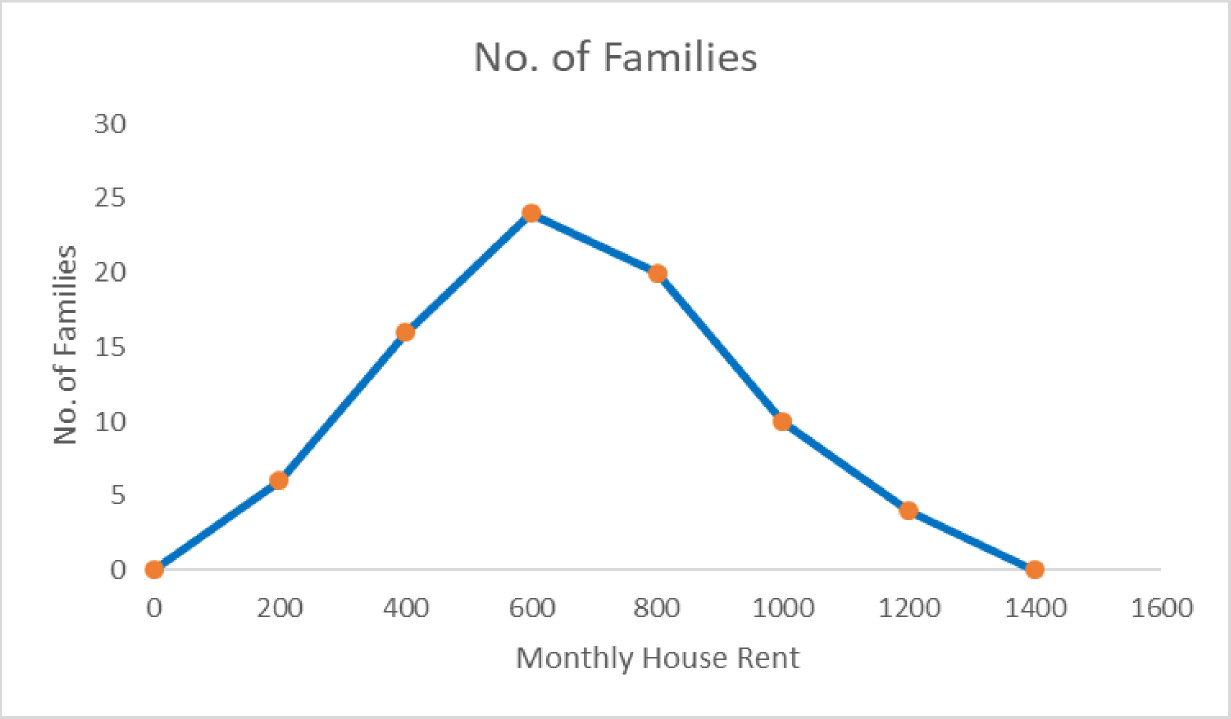
INTERPRETING A HISTOGRAM

A histogram tells you three main features of numerical data:

- ✓ How the data are distributed (symmetric, skewed right, skewed left, bell-shaped, and so on)
- ✓ The amount of variability in the data
- ✓ Where the center of the data is (approximately)

FREQUENCY POLYGON

- Frequency polygon is another way of representing the frequency distribution graphically.
- It enables us to understand the pattern in the data more clearly.
- Mid-values are taken on X-axis and frequencies are taken on Y-axis and the successive points are joined by the line segments
- To complete the polygon we obtain closed figure by taking two more classes, one preceding to first class and the other succeeding to last class. Frequency of these classes is taken to be zero.



MEASURES OF CENTRAL TENDENCY & DISPERSION

Let's say you are the captain of the Indian Cricket Team. We have already batted and the Australians are batting now. They require 10 runs from the last over to win. You have two good bowlers that still have an over to bowl. Who should you send to bowl the last over?

Without any extra data each is as good as the other.

SCENARIO 1

Suppose you now know that Bowler A concedes 8 runs on an average, Bowler B concedes 6 runs on an average.

Now who should we send in to bowl?

SCENARIO 2

Will a lower average always be better in such cases. Lets add some more data to what we already know. Bowler A concedes 8 runs on an average with a standard deviation of 0 runs which means he always concedes exactly 8 runs, neither more nor less (a little unrealistic). Bowler B concedes 6 runs on an average with a standard deviation of 4 runs that means he is quite erratic and may concede anything between zero to a large number.

Now who will you choose?

With a consistent eight runs you know that you have a 100% chance of winning.

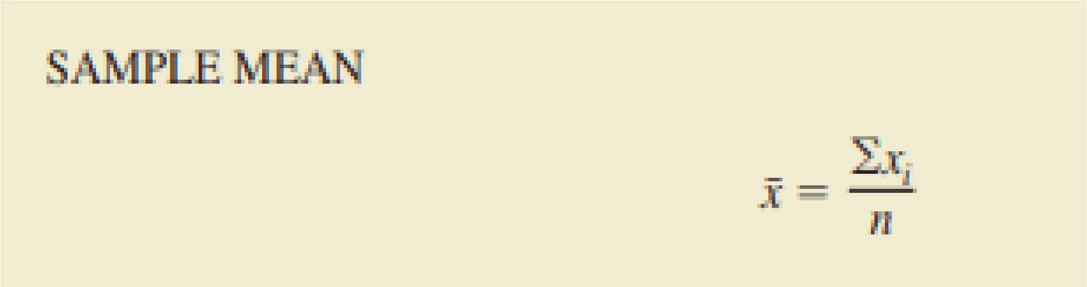
MEASURE OF CENTER – MEAN

The most common way to summarize a numerical data set is to describe where the center is. The center of a data set can be measured in different ways, and the method chosen can greatly influence the conclusions people make about the data.:

- ✓ **MEAN**
- ✓ **MEDIAN**
- ✓ **MODE**

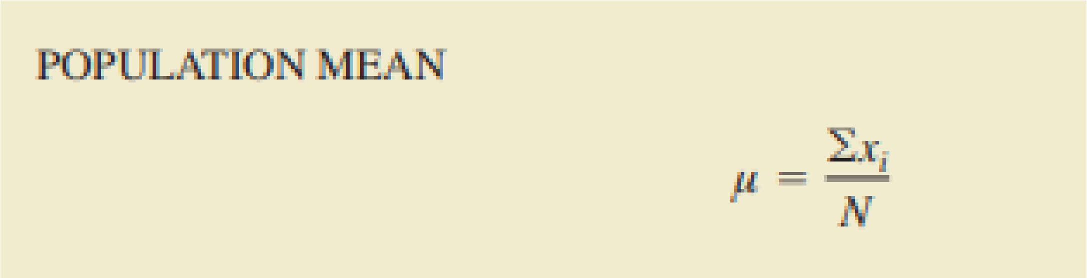
Perhaps the most important measure of location is the **MEAN,** or **AVERAGE VALUE**, for a variable. The mean provides a measure of central location for the data. If the data are for a sample, the mean is denoted by $\bar{x}$ if the data are for a population, the mean is denoted by the Greek letter **μ**.

In statistical formulas, it is customary to denote the value of variable x for the first observation by x1, the value of variable x for the second observation by x2, and so on. In general, the value of variable x for the ith observation is denoted by xi. For a sample with n observations, the formula for the sample mean is as follows.



SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x_i}{n}$$



POPULATION MEAN

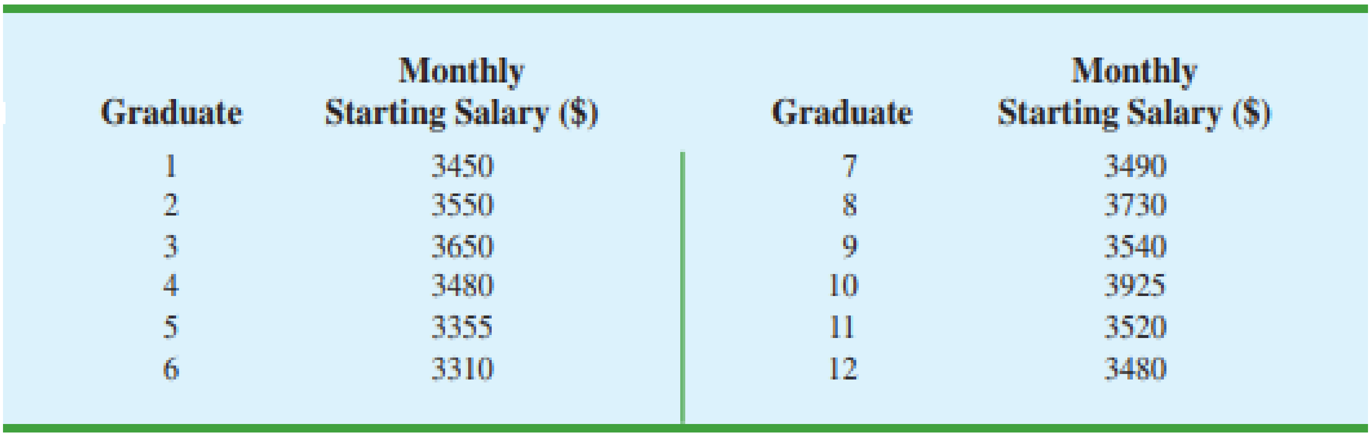
$$\mu = \frac{\Sigma x_i}{N}$$

$\overline{x}$ = Summation of Xi's Divided by n

Here is what you need to do to find the mean of a data set:

1. **Add up all the numbers in the data set.**
2. **Divide by the number of numbers in the data set, n.**

EXAMPLE:



Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

MEASURE OF CENTER – MEDIAN

The median is another measure of central location. The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value). With an odd number of observations, the median is the middle value. An even number of observations has no single middle value. In this case, we follow convention and define the median as the average of the values for the middle two observations.

The **MEDIAN** of a data set is the place that divides the data in half, once the data are ordered from smallest to largest.

To find the median of a data set:

1. Order the numbers from smallest to largest.
2. If the data set contains an odd number of numbers, the one exactly in the middle is the median.
3. If the data set contains an even number of numbers, take the two numbers that appear exactly in the middle and average them to find the median.

Let us apply this definition to compute the median class size for the sample of five college classes. Arranging the data in ascending order provides the following list:

32, 42, 46, 46, 54

Because n = 5 is odd, the median is the middle value. Thus, the median class size is 46 students. Even though this data set contains two observations with values of 46, each observation is treated separately when we arrange the data in ascending order.

Suppose we also compute the median starting salary for the 12 business college graduates. We first arrange the data in ascending order.

3310 3355 3450 3480 3480 **3490 3520** 3540 3550 3650 3730 3925
Middle Two Values

Because n = 12 is even, we identify the middle two values: 3490 and 3520. The median is the average of these values.

Median = (3490 + 3520)/2 = 3505

MEAN Vs MEDIAN

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers
- Thus, the median or a trimmed mean is also commonly used. Trimmed here means outliers are removed.
- Mean can only be used with numeric whereas median works with both ordinal and numeric

For instance, suppose that one of the graduates had a starting salary of $10,000 per month. If we change the highest monthly starting salary from $3925 to $10,000 and recompute the mean, the sample mean changes from $3540 to $4046. The median of $3505, however, is unchanged, because $3490 and $3520 are still the middle two values. With the extremely high starting salary included, the median provides a better measure of central location than the mean. We can generalize to say that whenever a data set contains extreme values, the median is often the preferred measure of central location.

Which measure of center should you use, the mean or the median? It depends on the situation, but reporting both is never a bad idea. When the mean and median are not close to each other in terms of their value, it’s a good idea to report both and let the reader interpret the results from there. Also, as a rule, be sure to ask for the median if you are only given the mean.

MEASURE OF CENTER – MODE

- The mode is the value that occurs with greatest frequency.
- For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data, but it can be used on any data type.

Situations can arise for which the greatest frequency occurs at two or more different values. In these instances, more than one mode exists. If the data contain exactly two modes, we say that the data are **BIMODAL**. If data contain more than two modes, we say that the data are **MULTIMODAL**. In multimodal cases the mode is almost never reported because listing three or more modes would not be particularly helpful in describing a location for the data.

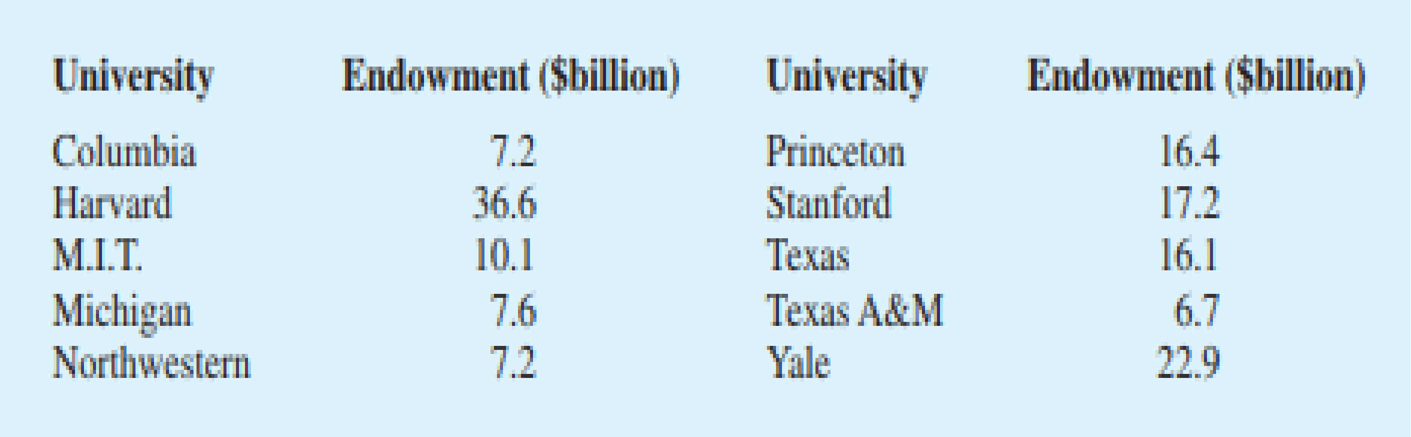
EXERCISE

Endowment income is a critical part of the annual budgets at colleges and universities. A study by the National Association of College and University Business Officers reported that the 435 colleges and universities surveyed held a total of $413 billion in endowments. The 10 wealthiest universities are shown below. Amounts are in billions of dollars.

a. What is the mean endowment for these universities?

b. What is the median endowment?

c. What is the mode endowment?



University	Endowment ($billion)	University	Endowment ($billion)
Columbia	7.2	Princeton	16.4
Harvard	36.6	Stanford	17.2
M.I.T.	10.1	Texas	16.1
Michigan	7.6	Texas A&M	6.7
Northwestern	7.2	Yale	22.9

MEASURES OF DISPERSION

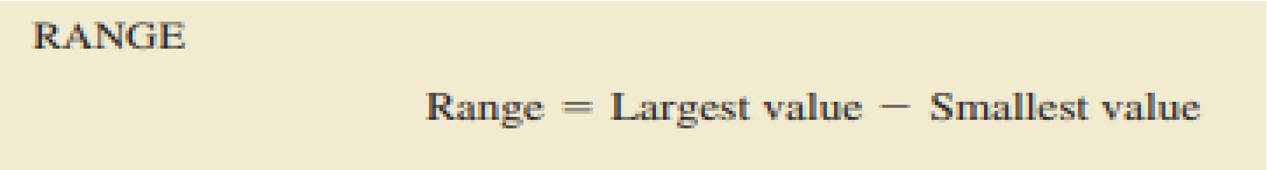
The measures of dispersion indicate the spread of the data or how closely the data is concentrated around the center value.

Variability is what the field of statistics is all about. Results vary from individual to individual, from group to group, from city to city, from moment to moment. Variation always exists in a data set, regardless of which characteristic you're measuring, because not every individual will have the same exact value for every characteristic you measure. Without a measure of variability you can't compare two data sets effectively. What if in both sets two sets of data have about the same average and the same median? Does that mean that the data are all the same? Not at all. For example, the data sets 199, 200, 201, and 0, 200, 400 both have the same average, which is 200, and the same median, which is also 200. Yet they have very different amounts of variability. The first data set has a very small amount of variability compared to the second.

We turn now to a discussion of some commonly used measures of variability.

RANGE

The simplest measure of variability is the **RANGE.**



RANGE

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

PERCENTILES

A percentile provides information about how the data are spread over the interval from the smallest value to the largest value. For data that do not contain numerous repeated values, the pth percentile divides the data into two parts.

Approximately p percent of the observations have values less than the pth percentile; approximately (100 - p) percent of the observations have values greater than the pth percentile.

The pth percentile is formally defined as follows.

The pth percentile is a value such that at least p percent of the observations are less than or equal to this value and at least (100

---

p) percent of the observations are greater than or equal to this value.

If your exam score is at the 90th percentile, for example, that means 90% of the people taking the exam with you scored lower than you did.

CALCULATING THE pTH PERCENTILE

Step 1. Arrange the data in ascending order (smallest value to largest value).

Step 2. Compute an index **i**, where **p** is the percentile of interest and **n** is the number of observations.

$$i=(p/100)*n$$

Step 3.

(a) If **i** is not an integer, round up. The next integer greater than **i** denotes the position of the **pth** percentile.

(b) If **i** is an integer, the **pth** percentile is the average of the values in positions **i** and **i+1**.

EXAMPLE:

As an illustration of this procedure, let us determine the **85th percentile** for the starting salary data above:

Step 1:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Step 2:

$i = (85/100)*12=10.2$

Step 3:

Because **i** is not an integer, round up. The position of the 85th percentile is the next integer greater than 10.2, **the 11th position**. Returning to the data, we see that the 85th percentile is the data value in the 11th position, or 3730.

EXAMPLE:

As another illustration of this procedure, let us consider the calculation of the 50th percentile for the starting salary data. Applying step 2, we obtain $i=(50/100)*12=6$

Because i is an integer, step 3(b) states that the 50th percentile is the average of the sixth and seventh data values; thus the 50th percentile is **(3490+3520)/2 = 3505**. Note that the 50th percentile is also the median. It is the point in the data where 50% of the data fall below that point and 50% fall above it.

QUARTILES

It is often desirable to divide data into four parts, with each part containing approximately one-fourth, or 25% of the observations. Figure 3.1 shows a data distribution divided into four parts. The division points are referred to as the quartiles and are defined as:

Q1 = first Quartile or 25th Percentile

Q2 = Second Quartile or 50th Percentile or MEDIAN

Q3 = Third Quartile of 75th Percentile

The computations of quartiles Q1 and Q3 require the use of the rule for finding the 25th and 75th percentiles. These calculations follow.

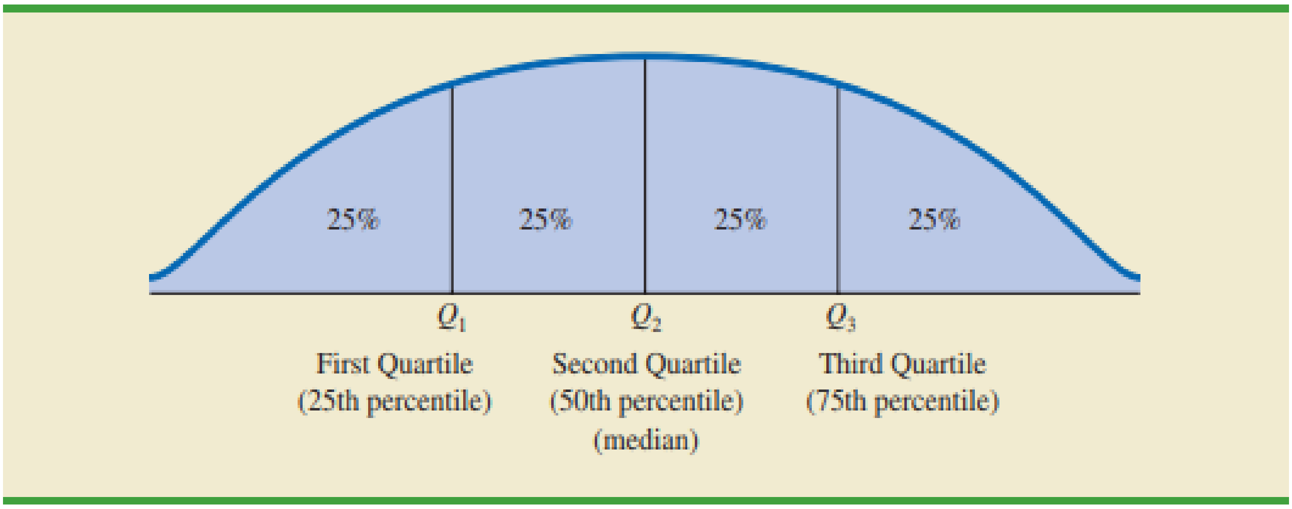
Q1 = (25/100)*12 = 3

Because i is an integer, step 3(b) indicates that the first quartile, or 25th percentile, is the average of the third and fourth data values; thus, Q1= (3450+3480)/2=3465.

Q3 = (75/100)*12 = 9

Because i is an integer, step 3(b) indicates that the first quartile, or 75th percentile, is the average of the ninth and tenth data values; thus, Q3= (3550+3650)/2=3600.

LOCATION OF THE QUARTILES



EXERCISE:

The cost of consumer purchases such as single-family housing, gasoline, Internet services, tax preparation, and hospitalization were provided in The Wall-Street Journal (January 2, 2007).

Sample data typical of the cost of tax-return preparation by services such as H&R Block are shown below.

120 230 110 115 160 130 150 105 195 155 105 360 120 120 140 100 115 180 235 255

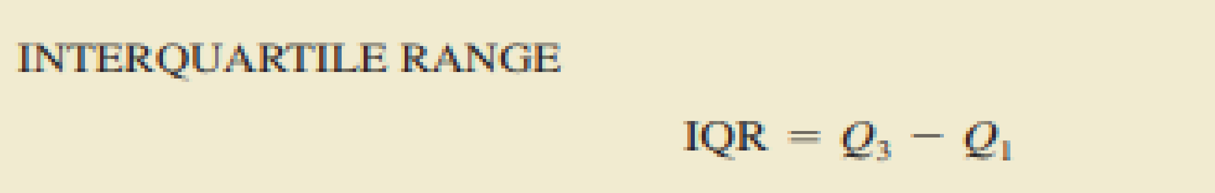
a. Compute the mean, median, and mode.

b. Compute the first and third quartiles.

c. Compute and interpret the 90th percentile.

INTER-QUARTILE RANGE

A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR). This measure of variability is the difference between the third quartile, Q3, and the first quartile, Q1. In other words, the interquartile range is the range for the middle 50% of the data.



INTERQUARTILE RANGE

$$\mathrm{IQR} = Q_3 - Q_1$$

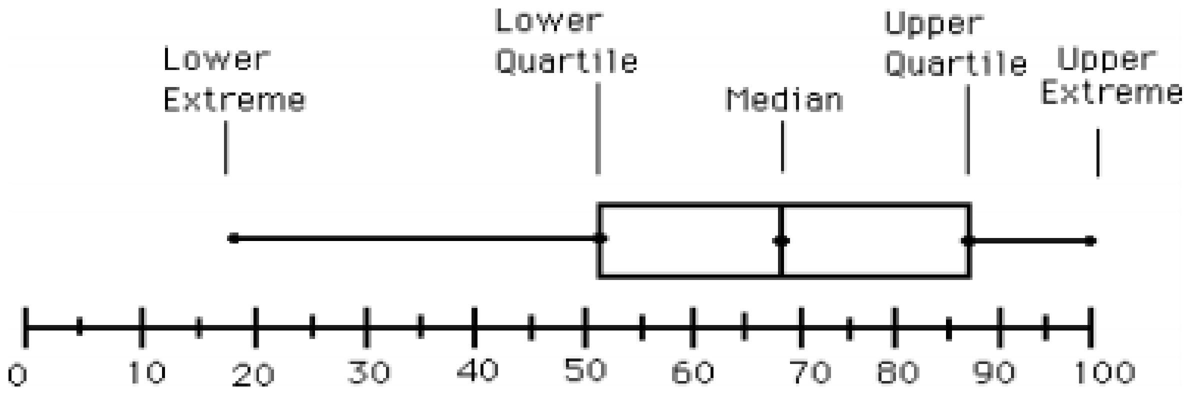
The IQR equals Q3 – Q1 and reflects the distance taken up by the innermost 50% of the data. If the IQR is small, you know there is much data close to the median. If the IQR is large, you know the data are more spread out from the median.

FIVE NUMBER SUMMARY

The five-number summary is a set of **five descriptive statistics** that divide the data set into four equal sections. The five numbers in a five number summary are:

1. The minimum (smallest) number in the data set.
2. The 25th percentile, aka the first quartile, or Q1.
3. The median (or 50th percentile).
4. The 75th percentile, aka the third quartile, or Q3.
5. The maximum (largest) number in the data set.

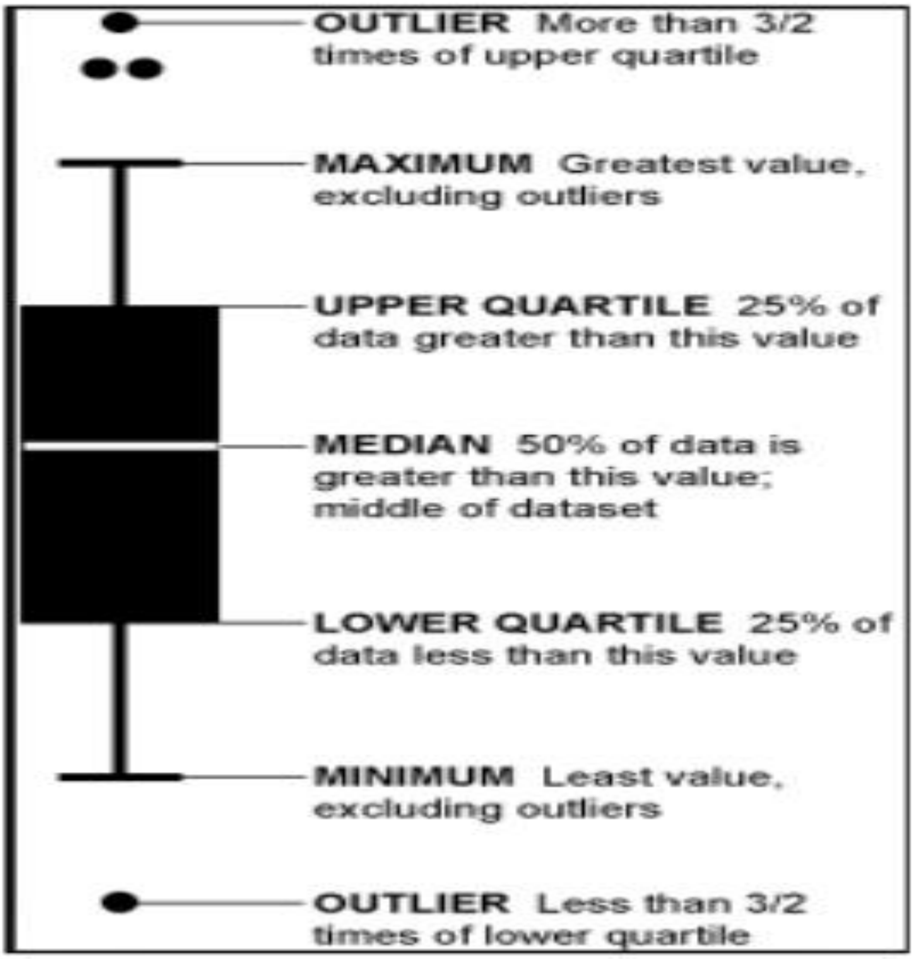
The purpose of the five-number summary is to give **descriptive statistics for center, variability, and relative standing** all in one shot. **T<u>he measure of center in the five-number summary is the median</u>, <u>and the first quartile, median, and third quartiles are measures of relative standing</u>.**



BOX PLOTS

A box plot is a graphical summary of data that is based on a five-number summary. A key to the development of a box plot is the computation of the median and the quartiles, Q1 and Q3. The interquartile range, IQR = Q3 - Q1, is also used.

- ✓ Data is represented with a box
- ✓ The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- ✓ The median is marked by a line within the box
- ✓ Whiskers: Two lines outside the box extended to Minimum and Maximum
- ✓ Outliers: Points beyond a specified outlier threshold, plotted individually


OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

VARIANCE

The variance is a measure of variability that utilizes all the data. The variance is based on the difference between the value of each observation (xi) and the mean. The difference between each xi and the mean ($\bar{x}$ for a sample, **μ** for a population) is called ***a deviation about the mean***. For a sample, a deviation about the mean is written (xi - $\bar{x}$); for a population, it is written (xi - μ). In the computation of the variance, the deviations about the mean are squared. If the data are for a population, the average of the squared deviations is called the population variance. The population variance is denoted by the Greek symbol **σ^2**. For a population of N observations and with μ denoting the population mean, the definition of the population variance is as follows:

POPULATION VARIANCE

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

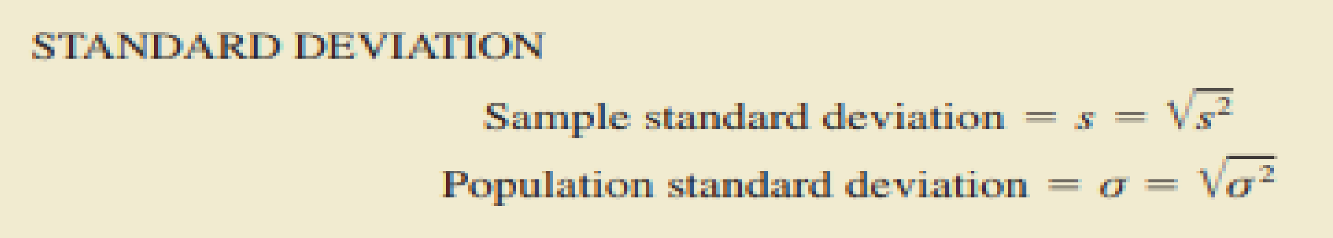
In most statistical applications, the data being analyzed are for a sample. When we compute a sample variance, we are often interested in using it to estimate the population variance σ2. If the sum of the squared deviations about the sample mean is divided by **n - 1**, and not n, the resulting sample variance provides an unbiased estimate of the population variance. For this reason, the sample variance, denoted by s2, is defined as follows:

SAMPLE VARIANCE

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

STANDARD DEVIATION

The standard deviation is defined to be the positive square root of the variance. Following the notation, we adopted for a sample variance and a population variance, we use **s** to denote the sample standard deviation and **σ** to denote the population standard deviation. The standard deviation is derived from the variance in the following way.



STANDARD DEVIATION

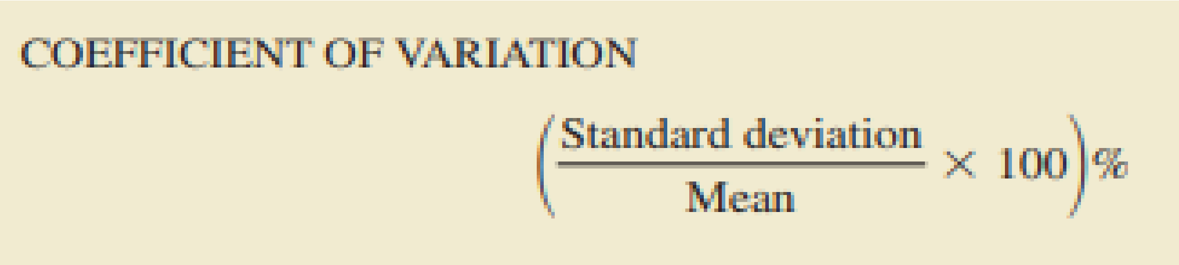
$$\text{Sample standard deviation} = s = \sqrt{s^2}$$

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2}$$

In other words, the standard deviation is measured in the same units as the original data. For this reason, the standard deviation is more easily compared to the mean and other statistics that are measured in the same units as the original data.

CO-EFFICIENT OF VARIANCE

In some situations, we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the coefficient of variation and is usually expressed as a percentage.



COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100\right)\%$$

For the class size data, we found a sample mean of 44 and a sample standard deviation of 8. The coefficient of variation is [(8/44) * 100] % = 18.2%. In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean. For the starting salary data with a sample mean of 3540 and a sample standard deviation of 165.65, the coefficient of variation, [(165.65/3540) * 100] % = 4.7%, tells us the sample standard deviation is only 4.7% of the value of the sample mean. In general, the coefficient of variation is a useful statistic for comparing the variability of variables that have different standard deviations and different means.

EXERCISE:

A bowler's scores for six games were 182, 168, 184, 190, 170, and 174. Using these data

as a sample, compute the following descriptive statistics:

a. Range c. Standard deviation

b. Variance d. Coefficient of variation