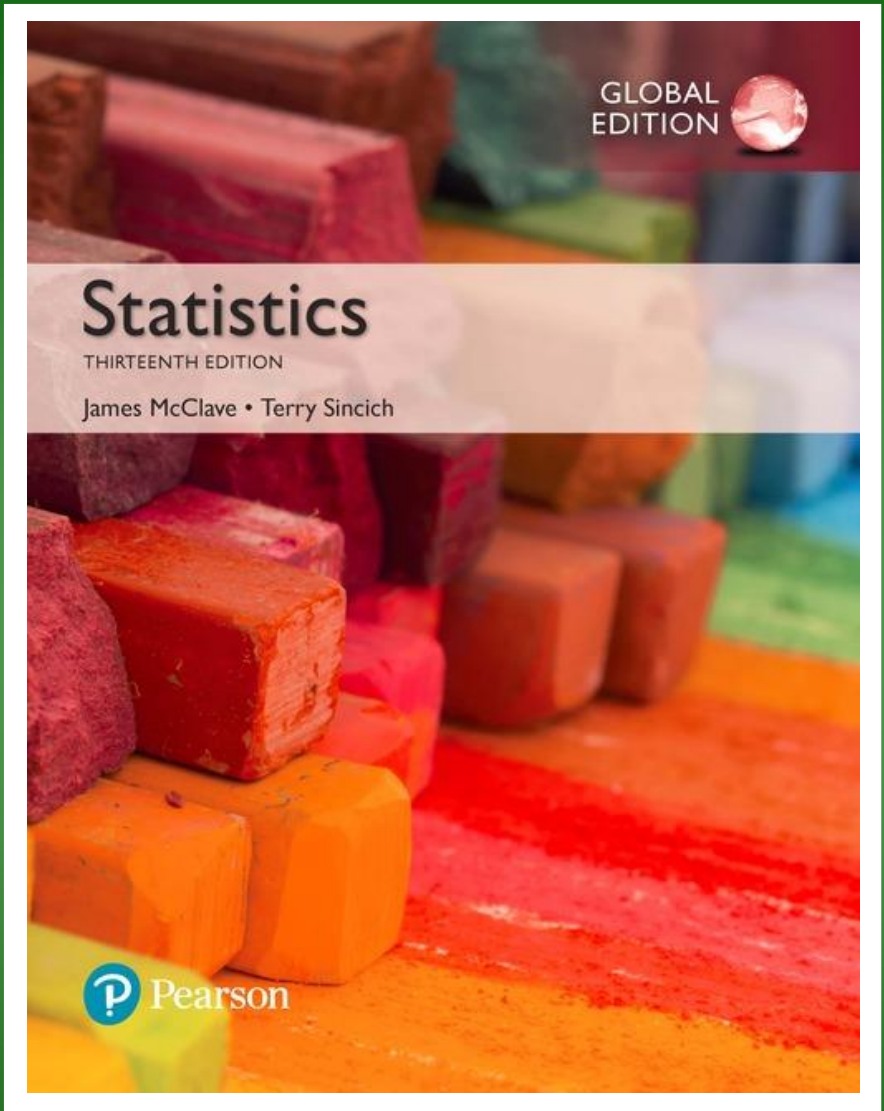


Chapter 2

Methods for Describing Sets of Data

Before we get started

- **Task:** Evaluate the math skills of a class of 1,000 first-year college students, based on their entrance exam (YKS) scores. How would you describe them?
  - YKS score
  - the variability in the scores;
  - the highest and lowest scores;
  - the "shape" of the data;
  - and whether the data set contains any unusual scores.
- Not an easy task, 1,000 scores provide **too many bits of information**.
- Some method for **summarizing** and **characterizing** the info is needed.
- Methods for describing data sets are also **essential for statistical inference**.
- Consequently, we need methods for describing a data set that let us make inferences about a population on the basis of information contained in a sample.
- We have two methods: *graphical* and *numerical*

2.1

Describing Qualitative Data

A study of aphasia published in the *Journal of Communication Disorders*

- Aphasia is the “impairment or loss of the faculty of using or understanding spoken or written language.”
- Three types of aphasia have been identified:
  - Broca’s,
  - conduction,
  - and anomic.
- The researchers wanted to determine whether one type of aphasia occurs more often than any other and, if so, how often.

Table 2.1

They measured the type of aphasia for a sample of 22 adult aphasics.

Table 2.1 Data on 22 Adult Aphasics

Subject	Type of Aphasia	Subject	Type of Aphasia
1	Broca's	12	Broca's
2	Anomic	13	Anomic
3	Anomic	14	Broca's
4	Conduction	15	Anomic
5	Broca's	16	Anomic
6	Conduction	17	Anomic
7	Conduction	18	Conduction
8	Anomic	19	Broca's
9	Conduction	20	Anomic
10	Anomic	21	Conduction
11	Conduction	22	Anomic

Based on Li, E. C., Williams, S. E., and Volpe, R. D. "The effects of topic and listener familiarity of discourse variables in procedural and narrative discourse tasks." *The Journal of Communication Disorders,* Vol. 28, No. 1, Mar. 1995, p. 44 (Table 1).

Definition

- For this study, the variable of interest, type of aphasia, is **qualitative** in nature.
- Qualitative data are nonnumerical in nature; thus, the value of a qualitative variable can only be classified into **categories** called **classes**.
- The possible types of aphasia—**Broca's**, **conduction**, and **anomic**—represent the classes for this qualitative variable.

Table 2.1 Data on 22 Adult Aphasics

Subject	Type of Aphasia	Subject	Type of Aphasia
1	Broca's	12	Broca's
2	Anomic	13	Anomic
3	Anomic	14	Broca's
4	Conduction	15	Anomic
5	Broca's	16	Anomic
6	Conduction	17	Anomic
7	Conduction	18	Conduction
8	Anomic	19	Broca's
9	Conduction	20	Anomic
10	Anomic	21	Conduction
11	Conduction	22	Anomic

Based on Li, E. C., Williams, S. E., and Volpe, R. D. "The effects of topic and listener familiarity of discourse variables in procedural and narrative discourse tasks." *The Journal of Communication Disorders,* Vol. 28, No. 1, Mar. 1995, p. 44 (Table 1).

A **class** is one of the categories into which qualitative data can be classified.

Definition

- We can summarize such data numerically in two ways:
- **(1)** by computing the **class frequency**—the number of observations in the data set that fall into each class

> The **class frequency** is the number of observations in the data set that fall into a particular class.

Definition

- **(2)** by computing the class relative frequency—the proportion of the total number of observations falling into each class.

The **class relative frequency** is the class frequency divided by the total number of observations in the data set; that is,

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}$$

Definition

The **class percentage** is the class relative frequency multiplied by 100; that is,

$$\text{class percentage} = (\text{class relative frequency}) \times 100$$

Figure 2.1 SPSS summary table for types of aphasia

- **10** aphasics in the study were diagnosed as suffering from anomic aphasia,
- **5** from Broca's aphasia,
- **7** from conduction aphasia.

- These numbers—10, 5, and 7—represent the class frequencies for the three classes.

Table 2.1 Data on 22 Adult Aphasics

Subject	Type of Aphasia	Subject	Type of Aphasia
1	Broca's	12	Broca's
2	Anomic	13	Anomic
3	Anomic	14	Broca's
4	Conduction	15	Anomic
5	Broca's	16	Anomic
6	Conduction	17	Anomic
7	Conduction	18	Conduction
8	Anomic	19	Broca's
9	Conduction	20	Anomic
10	Anomic	21	Conduction
11	Conduction	22	Anomic

TYPE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Anomic	10	45.5	45.5	45.5
	Brocas	5	22.7	22.7	68.2
	Conduction	7	31.8	31.8	100.0
	Total	22	100.0	100.0	

Figure 2.2 SAS bar graph for type of aphasia

- Although the summary table adequately describes the data, we often want a *graphical presentation* as well.
- Figure 2.2 shows the frequencies of the three types of aphasia in a bar graph produced with SAS.
- Note that the **height of the rectangle**, or "bar," over each class is equal to the **class frequency**.
- Optionally, the **bar heights can be** proportional to **class relative frequencies**.

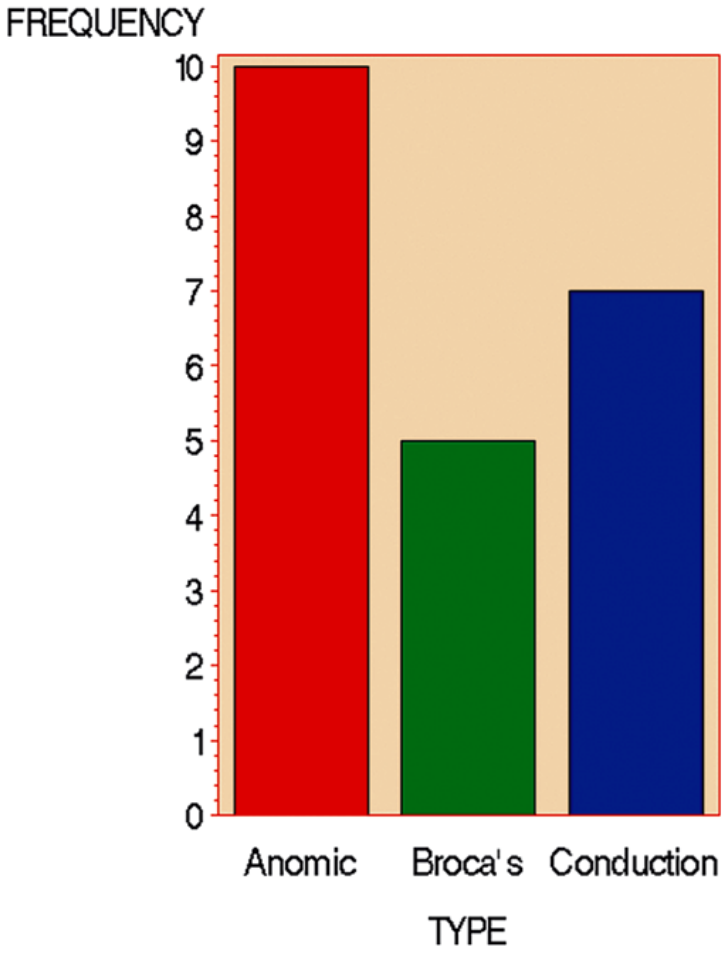


Figure 2.3 MINITAB pie chart for type of aphasia

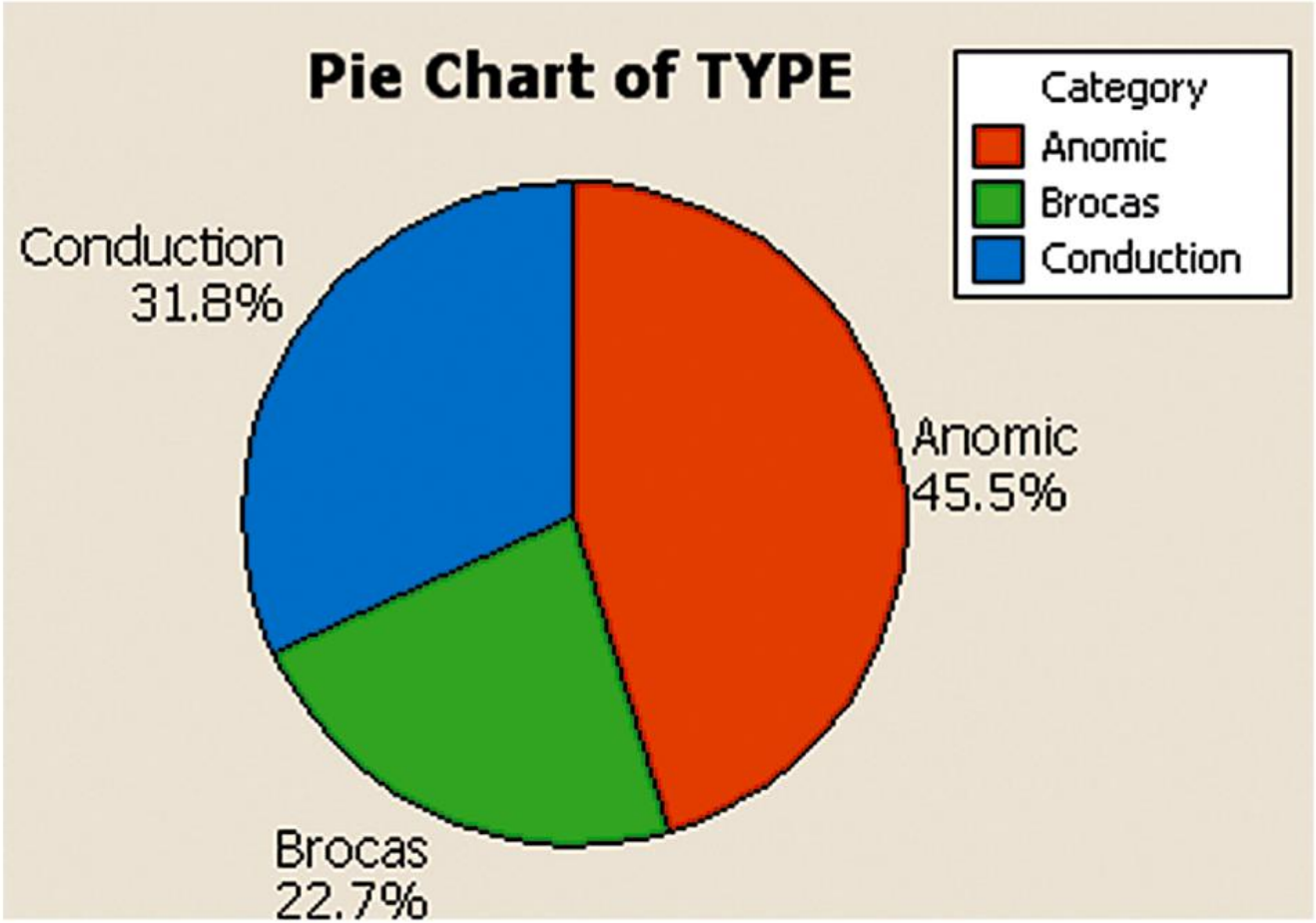
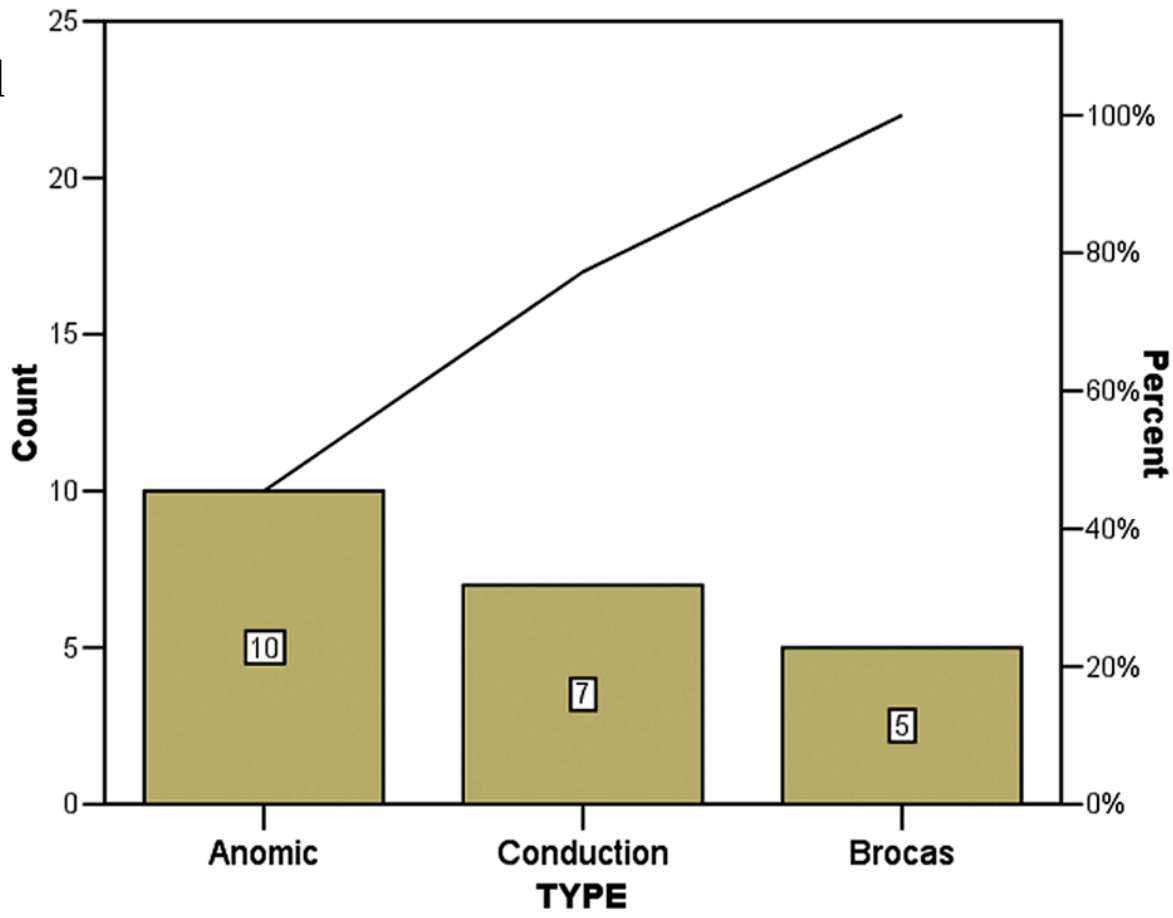


Figure 2.4 SPSS Pareto diagram for type of aphasia

- This rearrangement of the bars in a bar graph is called a **Pareto diagram**.
- One goal of a Pareto diagram (named for the Italian economist Vilfredo Pareto) is to make it **easy to locate the "most important" categories**—those with the largest frequencies.



Definition

Summary of Graphical Descriptive Methods for Qualitative Data

Bar Graph: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage.

Pie Chart: The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the class relative frequency.

Pareto Diagram: A bar graph with the categories (classes) of the qualitative variable (i.e., the bars) arranged by height in descending order from left to right.

Example

- **Problem** A group of cardiac physicians in southwest Florida has been studying a new drug designed to reduce blood loss in coronary bypass operations. Blood loss data for 114 coronary bypass patients (some who received a dosage of the drug and others who did not) are saved in the BLOOD file. Although the drug shows promise in reducing blood loss, the physicians are concerned about possible side effects and complications. So their data set includes not only the qualitative variable DRUG, which indicates whether or not the patient received the drug, but also the qualitative variable COMP, which specifies the type (if any) of complication experienced by the patient. The four values of COMP are (1) redo surgery, (2) post-op infection, (3) both, or (4) none.

Figure 2.5 SAS summary tables for DRUG and COMP

Summary tables for DRUG and COMP. Interpret the results.

The FREQ Procedure

DRUG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	57	50.00	57	50.00
YES	57	50.00	114	100.00

COMP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
BOTH	6	5.26	6	5.26
INFECT	15	13.16	21	18.42
NONE	79	69.30	100	87.72
REDO	14	12.28	114	100.00

Figure 2.6 MINITAB side-by-side bar graphs for COMP by value of DRUG

Interpret the results.

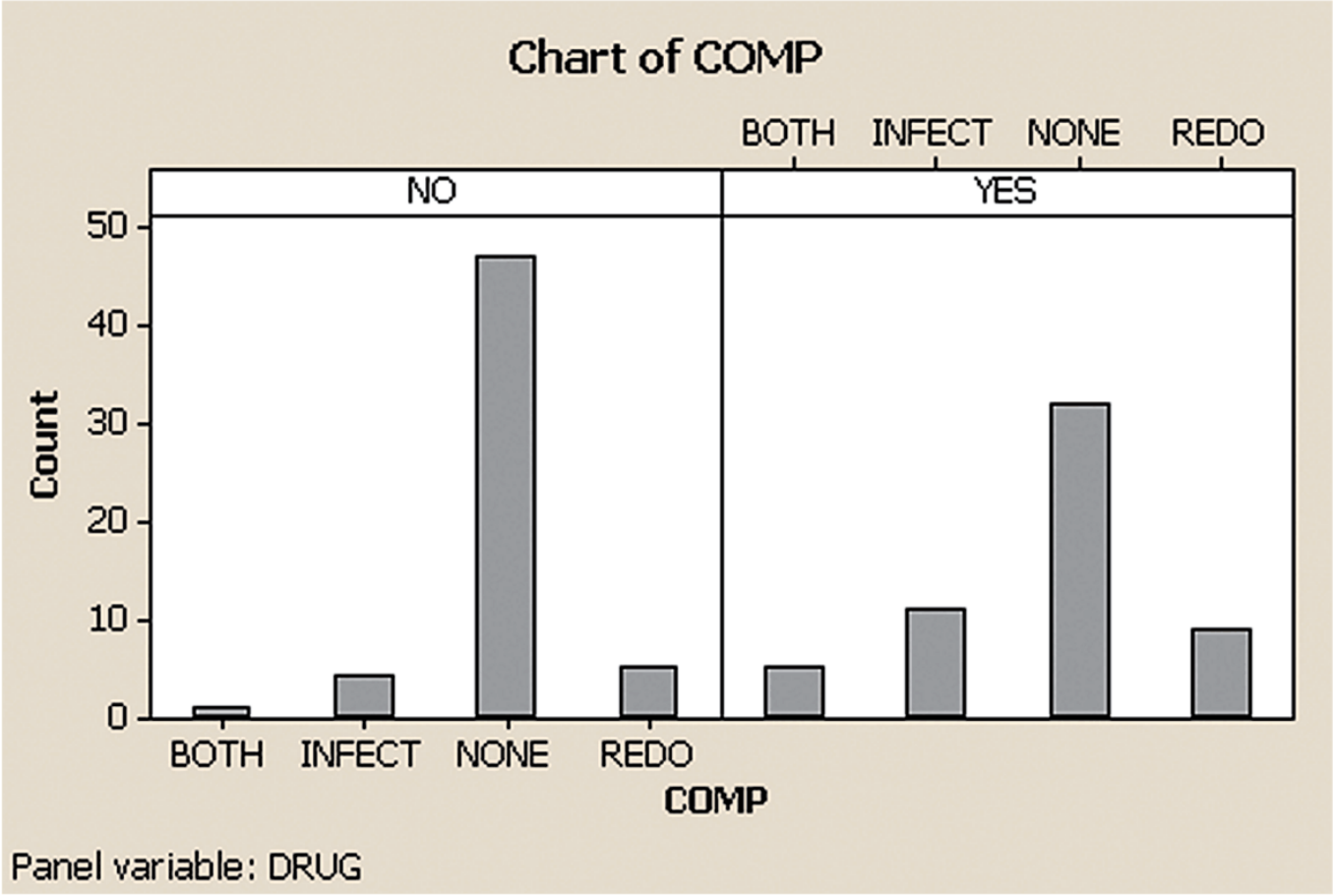


Figure 2.7 SPSS summary tables for COMP by value of drug

Interpret the results.

COMP

DRUG			Frequency	Percent	Valid Percent	Cumulative Percent
NO	Valid	BOTH	1	1.8	1.8	1.8
		INFECT	4	7.0	7.0	8.8
		NONE	47	82.5	82.5	91.2
		REDO	5	8.8	8.8	100.0
		Total	57	100.0	100.0	
YES	Valid	BOTH	5	8.8	8.8	8.8
		INFECT	11	19.3	19.3	28.1
		NONE	32	56.1	56.1	84.2
		REDO	9	15.8	15.8	100.0
		Total	57	100.0	100.0	

Solution

- The top table in Figure 2.5 is a summary frequency table for DRUG. Note that exactly half (57) of the 114 coronary bypass patients received the drug and half did not. The bottom table in Figure 2.5 is a summary frequency table for COMP. The class percentages are given in the Percent column. We see that about 69% of the 114 patients had no complications, leaving about 31% who experienced either a redo surgery, a post-op infection, or both.
- Figure 2.6 is a MINITAB side-by-side bar graph of the data. The four bars in the left-side graph represent the frequencies of COMP for the 57 patients who did not receive the drug; the four bars in the right-side graph represent the frequencies of COMP for the 57 patients who did receive a dosage of the drug. The graph clearly shows that patients who did not receive the drug suffered fewer complications. The exact percentages are displayed in the SPSS summary tables of Figure 2.7. Over 56% of the patients who got the drug had no complications, compared with about 83% for the patients who got no drug.
- So, should we continue using the drug or not? Can we make a decision based on inference?

Should we continue using the drug?

- Although these results show that the drug may be **effective in reducing blood loss**, Figures 2.6 and 2.7 imply that **patients on the drug may have a higher risk of incurring complications**. But before using this information to make a decision about the drug, the physicians will need to provide **a measure of reliability for the inference**. That is, the physicians will want to know whether the difference between the percentages of patients with complications observed in this sample of 114 patients **is generalizable to the population** of all coronary bypass patients.

Statistics in Action
Interpreting Pie Charts for the Body image Data

- In the Body Image: An International Journal of Research (Jan. 2010) study, Brown University researchers measured several qualitative (categorical) variables for each of 92 body dysmorphic disorder (BDD) patients: Gender (M or F), Comorbid Disorder (Major Depression, Social Phobia, Obsessive Compulsive Disorder—OCD, or Anorexia/Bulimia Nervosa), and Dissatisfied with Looks (Yes or No).
- Pie charts and bar graphs can be used to summarize and describe the responses for these variables.

Figure SIA2.1a MINITAB pie charts for Gender, Disorder, and Dissatisfies with Looks

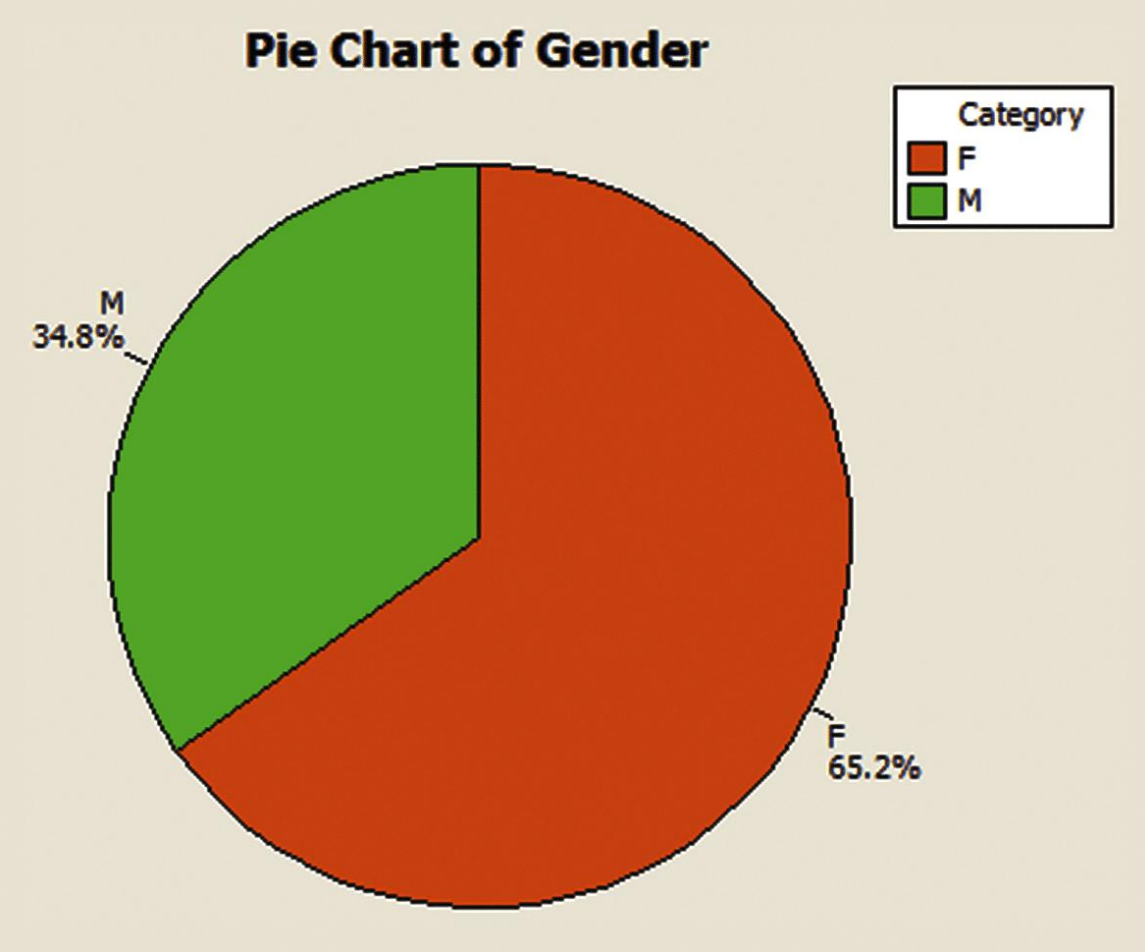


Figure SIA2.1b MINITAB pie charts for Gender, Disorder, and Dissatisfies with Looks

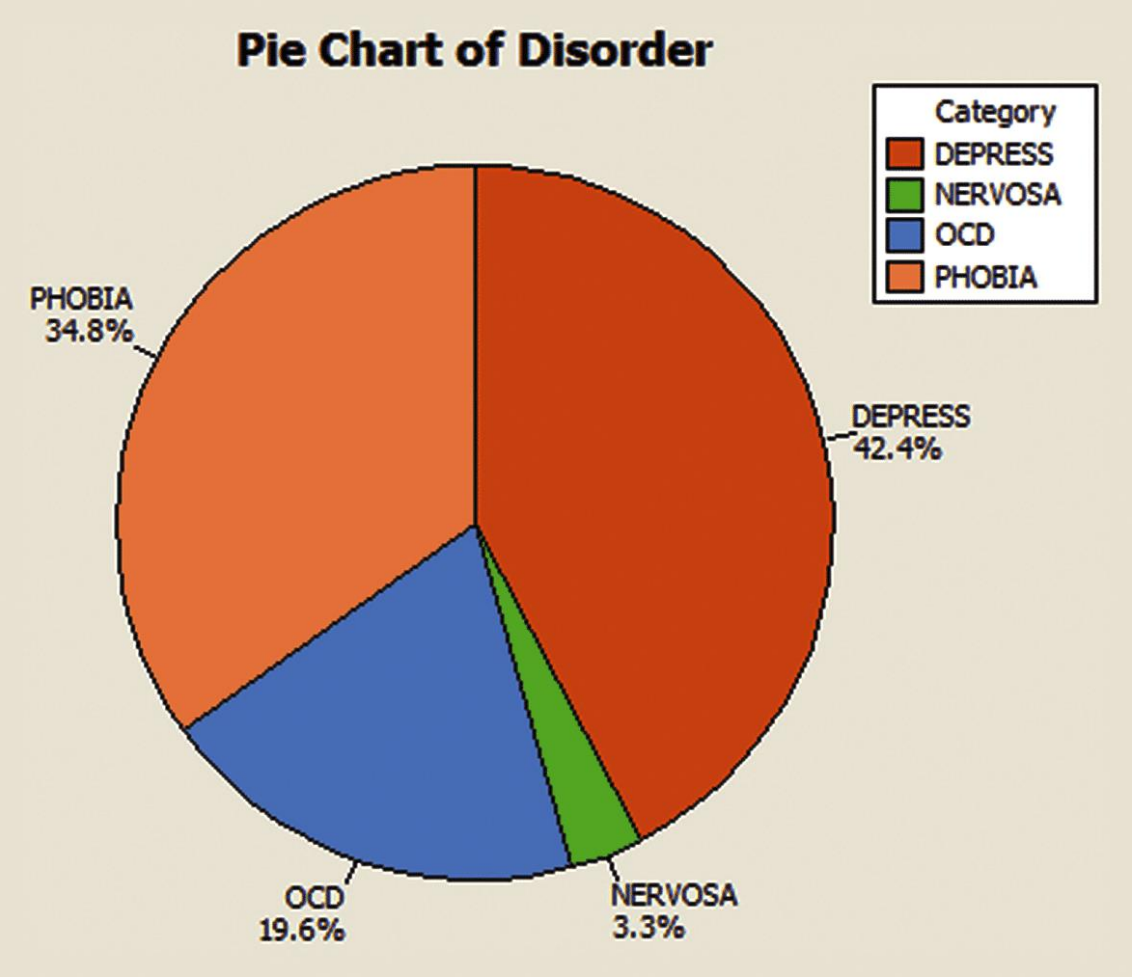


Figure SIA2.1c MINITAB pie charts for Gender, Disorder, and Dissatisfies with Looks

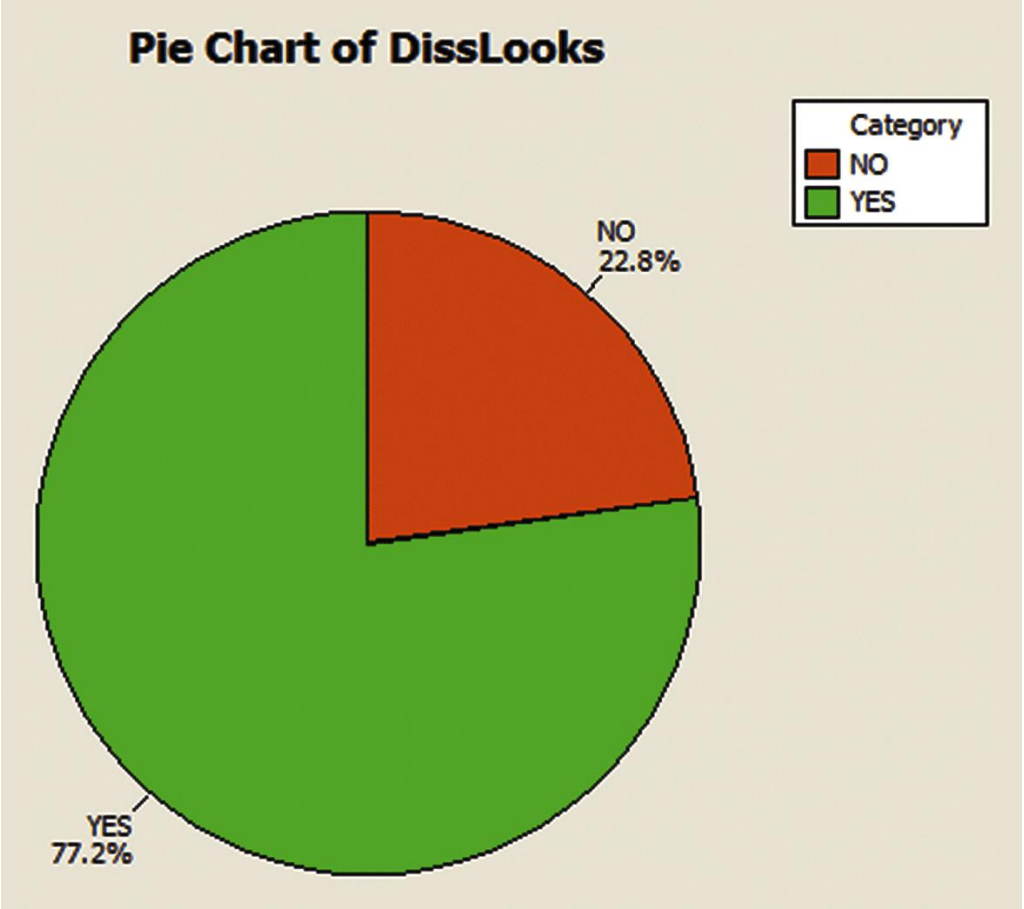


Figure SIA2.2 MINITAB side-by-side pie charts for Dissatisfied with Looks by Gender

- **Are BDD females tend to be more dissatisfied with their looks than BDD males?**
- We can gain insight into this question by forming side-by-side pie charts one for females and one for males.
- You can see that about 82% of the females are dissatisfied in some way with their body as compared to about 69% of the males.
- Thus, it **does appear** that BDD females tend to be more dissatisfied with their looks than males, at least for this sample of patients.

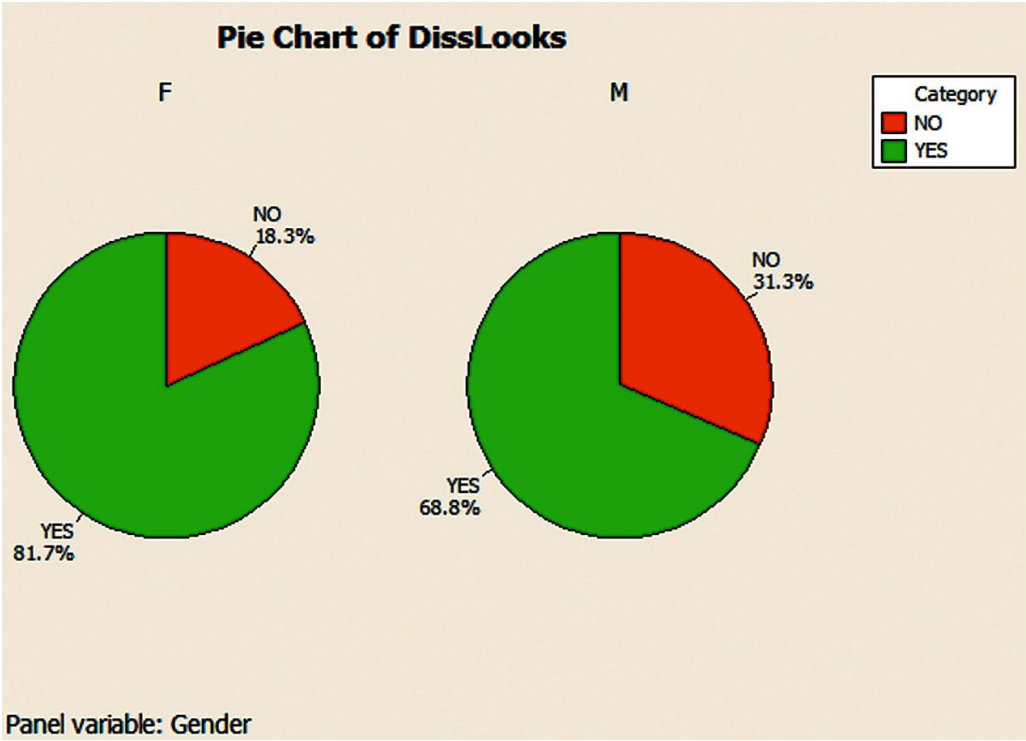
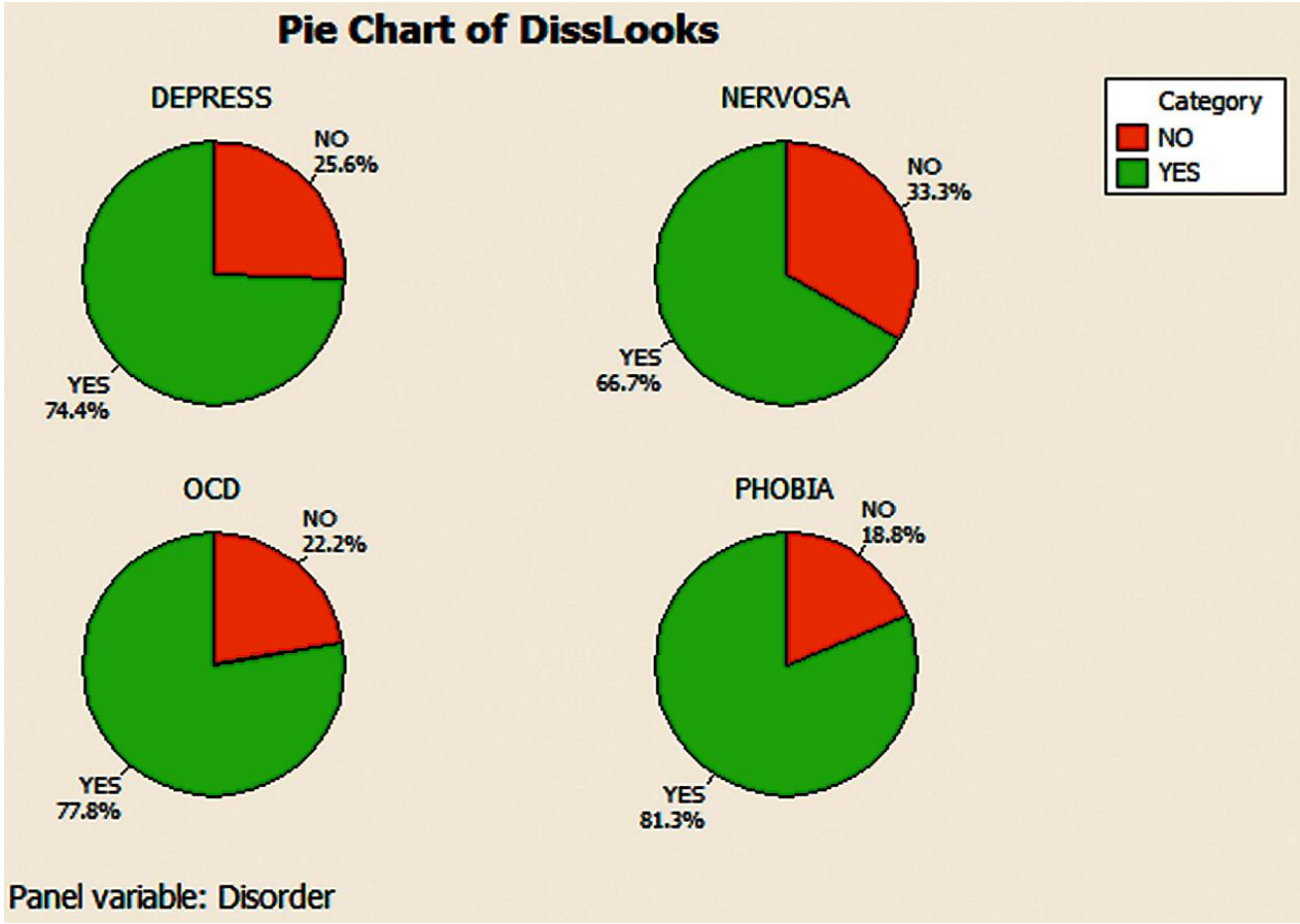


Figure SIA2.3 MINITAB side-by-side Pie charts for Dissatisfied with Looks by Comorbid Disorder

- **Are certain comorbid disorders lead to a higher level of dissatisfaction with body appearance?**
- The percentage of dissatisfied BDD patients range from about 67% for those with a nervosa disorder to about 81% for those diagnosed with a social phobia.



2.2

Graphical Methods for Describing Quantitative Data

Describing quantitative data

- Recall that **quantitative data sets** consist of data that are recorded on a **meaningful numerical scale**.
- To describe, summarize, and detect patterns in such data, we can use three graphical methods:
  - dot plots,
  - stem-and-leaf displays,
  - and histograms.

Table 2.2

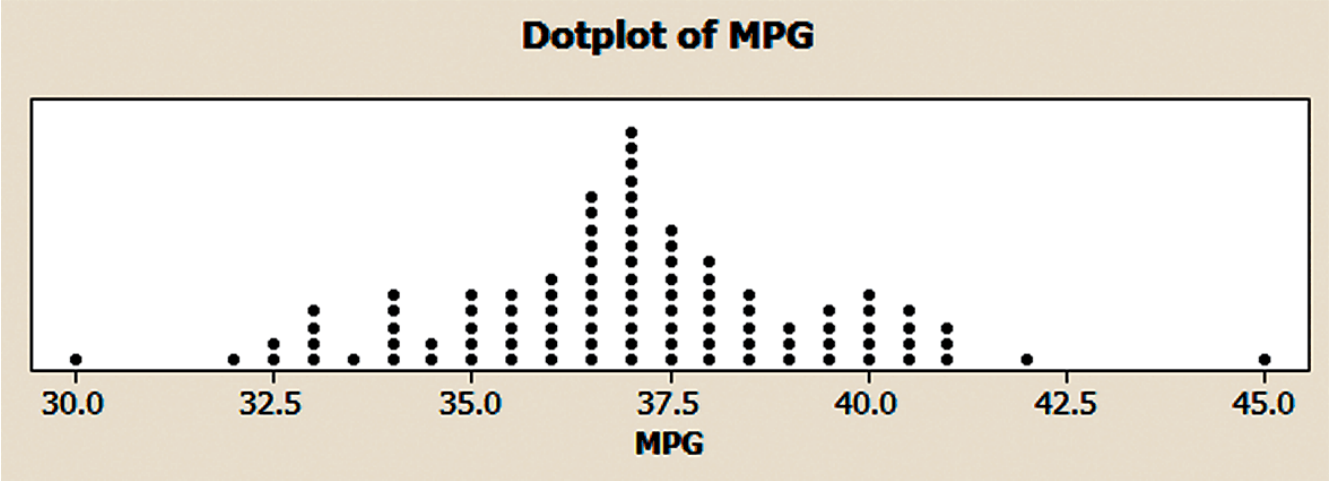
- The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. Suppose that the 100 measurements in Table 2.2 represent the results of such tests on a certain new car model.

- How can we summarize the information in this rather large sample?

Table 2.2 EPA Mileage Ratings on 100 Cars

36.3	41.0	36.9	37.1	44.9
32.7	37.3	41.2	36.6	32.9
40.5	36.5	37.6	33.9	40.2
36.2	37.9	36.0	37.9	35.9
38.5	39.0	35.5	34.8	38.6
36.3	36.8	32.5	36.4	40.5
41.0	31.8	37.3	33.1	37.0
37.0	37.2	40.7	37.4	37.1
37.1	40.3	36.7	37.0	33.9
39.9	36.9	32.9	33.8	39.8
36.8	30.0	37.2	42.1	36.7
36.5	33.2	37.4	37.5	33.6
36.4	37.7	37.7	40.0	34.2
38.2	38.3	35.7	35.6	35.1
39.4	35.3	34.4	38.8	39.7
36.6	36.1	38.2	38.4	39.3
37.6	37.0	38.7	39.0	35.8
37.8	35.9	35.6	36.7	34.5
40.1	38.0	35.2	34.8	39.5
34.0	36.8	35.0	38.1	36.9

Figure 2.8 MINITAB dot plot for 100 EPA mileage ratings



- The horizontal axis is a scale for the quantitative variable in miles per gallon.
- The rounded (to the nearest half gallon) numerical value of each measurement in the data set is located on the horizontal scale by a dot.
- When data values repeat, the dots are placed above one another, forming a pile at that particular numerical location.
- Dot plot verifies that almost all of the mileage ratings are in the 30s, with most falling between 35 and 40 miles per gallon.

Figure 2.9 MINITAB stem-and-leaf display for 100 mileage ratings

Stem: portion of the measurement (mpg) to the left of the decimal point.

Leaf: the remaining portion, to the right of the decimal point.

Stem-and-Leaf Display: MPG

```
Stem-and-leaf of MPG  N  = 100
Leaf Unit = 0.10

 1     30  0
 2     31  8
 6     32  5799
 12    33  126899
 18    34  024588
 29    35  01235667899
 49    36  012334455667778889 99
(21)   37  000011122334456677899
 30    38  0122345678
 20    39  00345789
 12    40  0123557
 5     41  002
 2     42  1
 1     43
 1     44  9
```

Figure 2.10 SPSS histogram for 100 EPA gas mileage ratings

- The horizontal axis of the figure, which gives the miles per gallon for a given automobile, is divided into **class intervals**, commencing with the interval from 30–31 and proceeding in intervals of equal size to 44–45 mpg.
- The vertical axis gives the number (or frequency) of the 100 readings that fall into each interval.
- Histograms can be used to display either the **frequency** or **relative frequency** of the measurements falling into the class intervals.

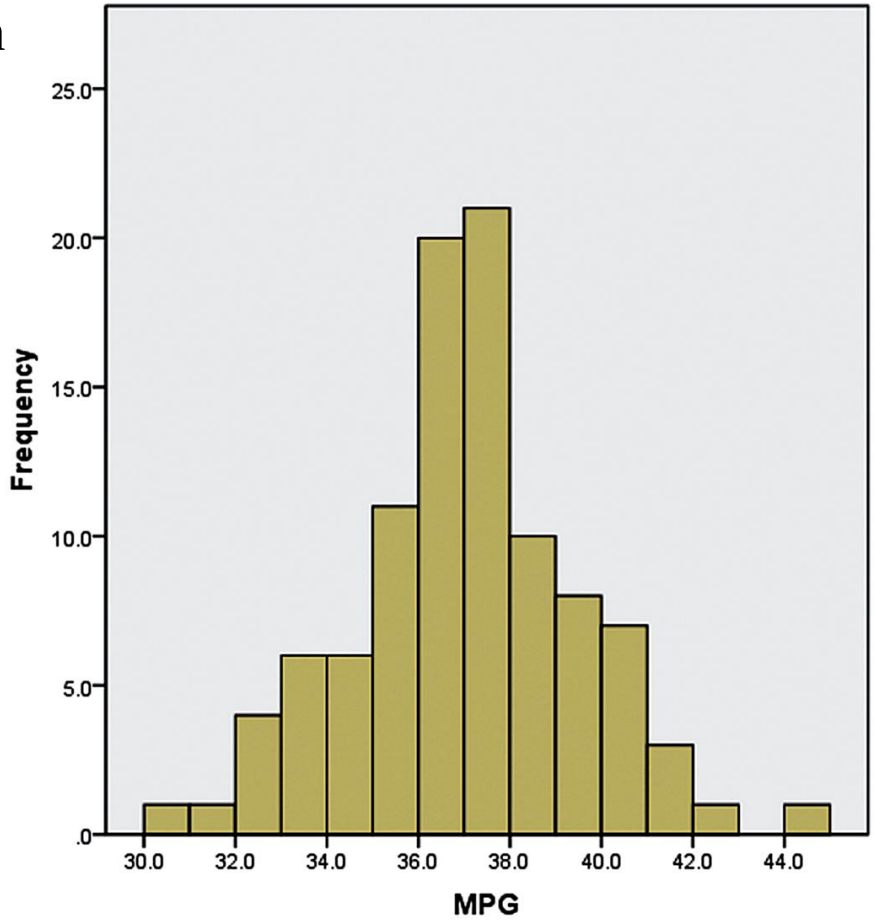


Table 2.3

Table 2.3 Class Intervals, Frequencies, and Relative Frequencies for the Gas Mileage Data

Class Interval	Frequency	Relative Frequency
30–31	1	0.01
31–32	1	0.01
32–33	4	0.04
33–34	6	0.06
34–35	6	0.06
35–36	11	0.11
36–37	20	0.20
37–38	21	0.21
38–39	10	0.10
39–40	8	0.08
40–41	7	0.07
41–42	3	0.03
42–43	1	0.01
43–44	0	0.00
44–45	1	0.01
Totals	100	1.00

Figure 2.10 SPSS histogram for 100 EPA gas mileage ratings

- In interpreting a histogram, consider two important facts.
- First, the proportion of the total area under the histogram that falls above a particular interval on the x-axis is equal to the relative frequency of measurements falling into that interval.
- For example, the relative frequency for the class interval 37–38 in Figure 2.10 is 21.
- Consequently, the rectangle above the interval contains 0.21 of the total area under the histogram.

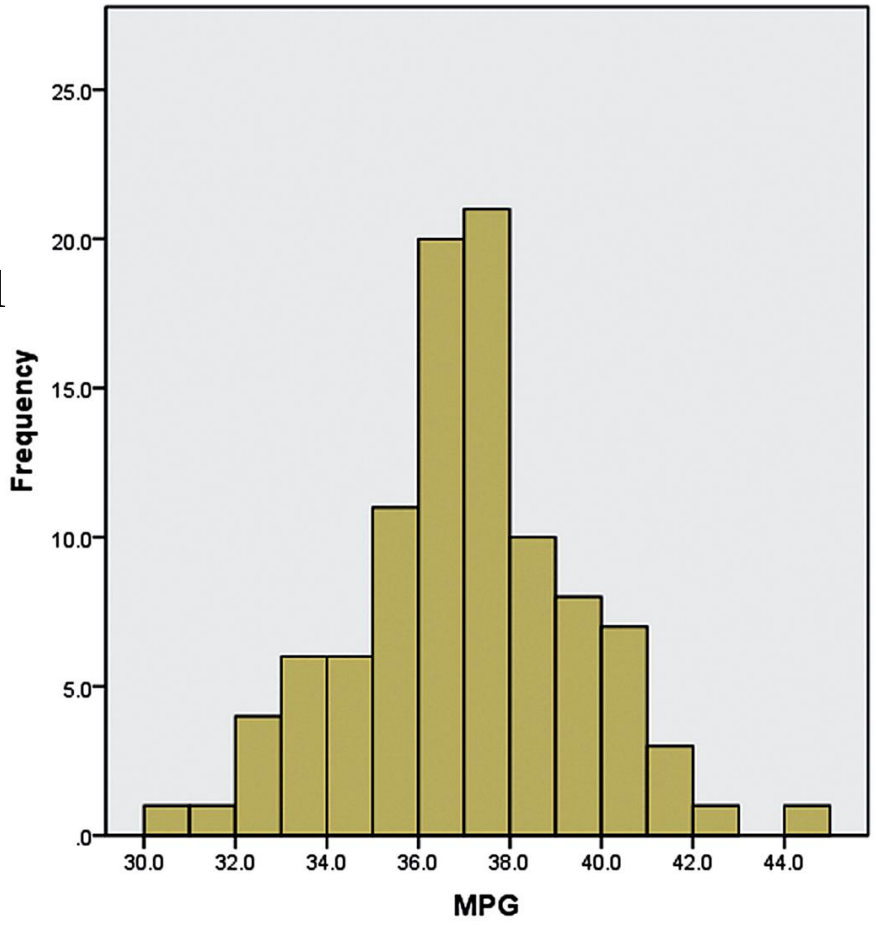
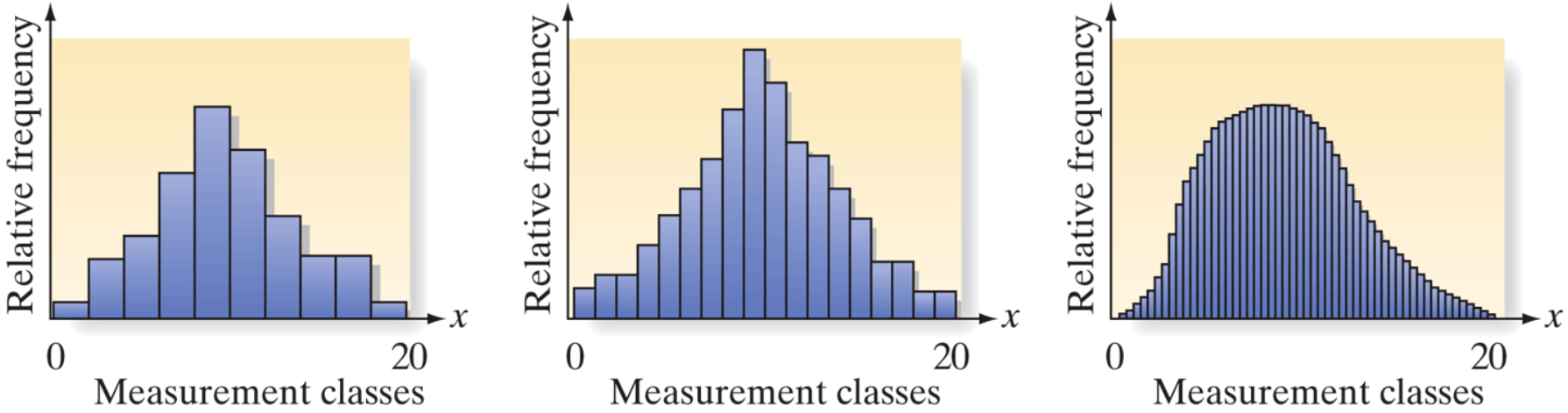


Figure 2.11 The effect of the size of a data set on the outline of a histogram

- Second, imagine the appearance of the relative frequency histogram for a very large set of data (representing, say, a population). As the number of measurements in a data set is increased, you can obtain a better description of the data by decreasing the width of the class intervals.
- When the class intervals become small enough, a relative frequency histogram will (for all practical purposes) appear as a smooth curve.



Procedure

Determining the Number of Classes in a Histogram

Number of Observations in Data Set	Number of Classes
Fewer than 25	5–6
25–50	7–14
More than 50	15–20

- While **histograms provide good visual descriptions** of data sets—particularly very large ones—they **do not let us identify individual measurements**.
- In contrast, each of the **original measurements is visible** to some extent in a **dot plot** and is **clearly visible in a stem-and-leaf display**.
- However, **stem-and-leaf displays can become unwieldy for very large data sets**.

Example

- **Problem** Over 60 years ago, famous child psychologist Jean Piaget devised a test of basic perceptual and conceptual skills dubbed the "water-level task." Subjects were shown a drawing of a glass being held at a 45° angle and asked to draw a line representing the true surface of the water. Today, research psychologists continue to use the task to test the perception of both adults and children. In one study, the water-level task was given to several groups that included 20 male bartenders and 20 female waitresses. For each participant, the researchers measured the deviation (in angle degrees) of the judged line from the true line. These deviations are shown in Table 2.4. [Note: Deviations can be negative if the judged angle is smaller than the angle of the true line.]

Table 2.4

Table 2.4	Water-Level Task Deviations (angle degrees)																			
Bartenders:	−9	6	10	6	10	−3	7	8	6	14	7	8	−5	2	−1	0	2	3	0	2
Waitresses:	7	10	25	8	10	8	12	9	35	10	12	11	7	10	21	−1	4	0	16	−1

- **a.** Create a frequency histogram for the combined data in Table 2.4. Then, shade the area under the histogram that corresponds to deviations recorded for waitresses.
  Interpret the result.

- **b.** Create a stem-and-leaf display for these combined data. Again, shade each leaf of the display that corresponds to a deviation recorded for a waitress.
  Interpret the result.

Figure 2.12 SPSS histogram for task deviations

- Note that SPSS formed 20 classes, with class intervals -20 to -15, -15 to -10, . . . , 30 to 35, and 35 to 40.
- This histogram clearly shows the clustering of the deviation angles between 0° and 15°, with a few deviations in the upper end of the distribution (greater than 20°).
- The graph clearly shows that waitresses tend to have greater (positive) deviations than do bartenders and fewer deviations near 0° relative to bartenders.

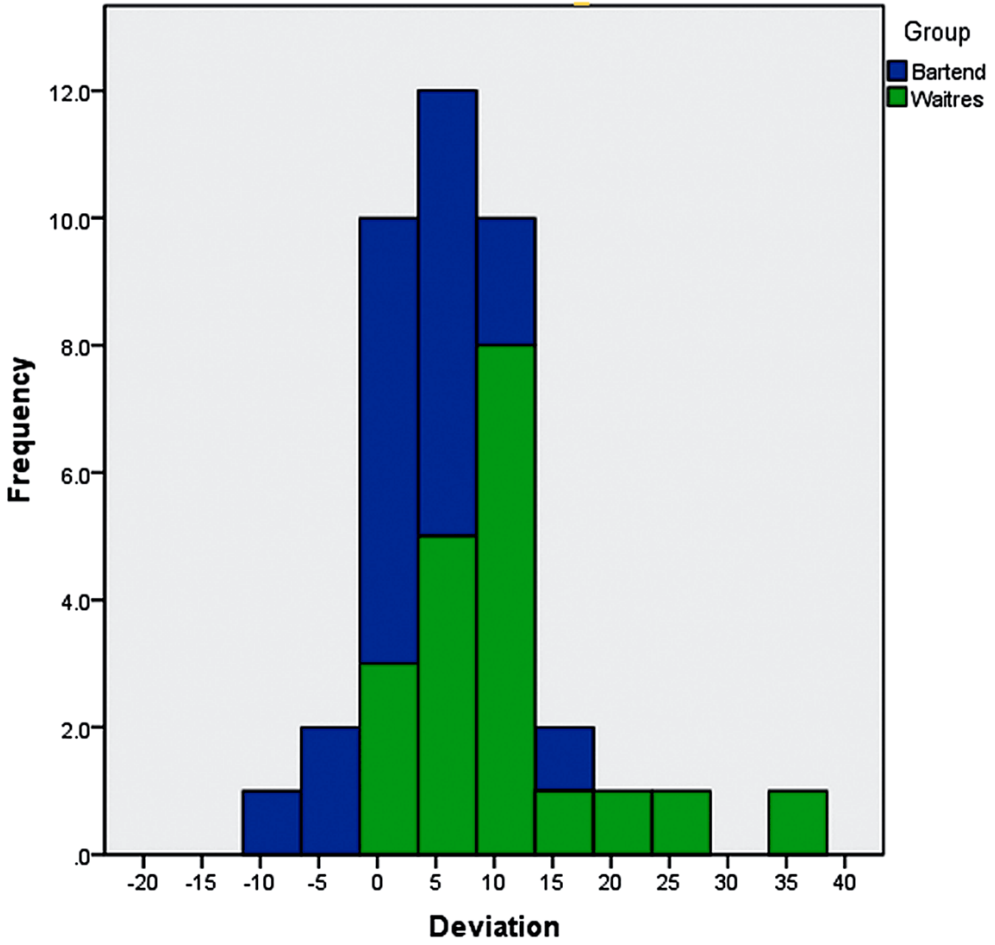


Figure 2.13 MINITAB stem-and-leaf display for task deviations

- Note that the stem (the second column on the printout) represents the first digit (including 0) in the deviation angle measurement while the leaf (the third column on the printout) represents the second digit.
- Thus, the leaf 5 in the stem 2 row represents the deviation angle of 25°.
- The shaded leaves represent deviations recorded for waitresses.
- As with the histogram, the stem-and-leaf display shows that deviations for waitresses tend to appear in the upper tail of the distribution.

Stem-and-Leaf Display: Deviation

```
Stem-and-leaf of Deviation   N   = 40
Leaf Unit = 1.0

 2     -0   95
 7     -0   31110
 14    0    0022234
(12)   0    666777788889
 14    1    0000001224
 4     1    6
 3     2    1
 2     2    5
 1     3
 1     3    5
```

Together, the graphs **imply** that waitresses tend to overestimate the angle of the true line relative to bartenders.

Definition

Summary of Graphical Descriptive Methods for Quantitative Data

Dot Plot: The numerical value of each quantitative measurement in the data set is represented by a dot on a horizontal scale. When data values repeat, the dots are placed above one another vertically.

Stem-and-Leaf Display: The numerical value of the quantitative variable is partitioned into a "stem" and a "leaf." The possible stems are listed in order in a column. The leaf for each quantitative measurement in the data set is placed in the corresponding stem row. Leaves for observations with the same stem value are listed in increasing order horizontally.

Histogram: The possible numerical values of the quantitative variable are partitioned into class intervals, each of which has the same width. These intervals form the scale of the horizontal axis. The frequency or relative frequency of observations in each class interval is determined. A vertical bar is placed over each class interval, with the height of the bar equal to either the class frequency or class relative frequency.

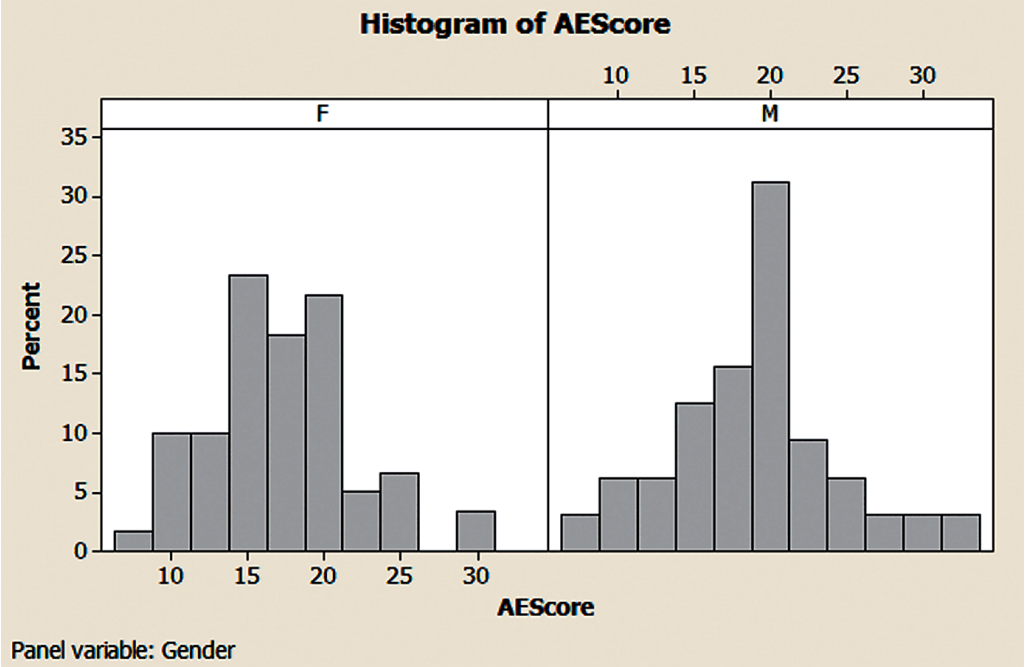
Statistics in Action

Interpreting Histograms for the Body image Data

- In the *Body Image: An International Journal of Research* (Jan. 2010) study of 92 BDD patients, the researchers asked each patient to respond to a series of questions on body image (e.g., “How satisfied are you with your physical attractiveness and looks?”). Recall that the scores were summed to yield an Appearance Evaluation score that ranged from 7 to 35 points. This score represents a quantitative variable. Consequently, to graphically investigate whether BDD females tend to be more dissatisfied with their looks than BDD males, we can form side-by-side histograms for the total score, one histogram for females and one for males.

Figure SIA2.4 MINITAB side-by-side histograms for Appearance Evaluation by Gender

- For females, the histogram for appearance evaluation score is centered at about 17 points, while for males the histogram is centered higher, at about 20 points.
- Also, from the histograms you can see that about 55% of the female patients had a score of less than 20, compared to only about 45% of the males.



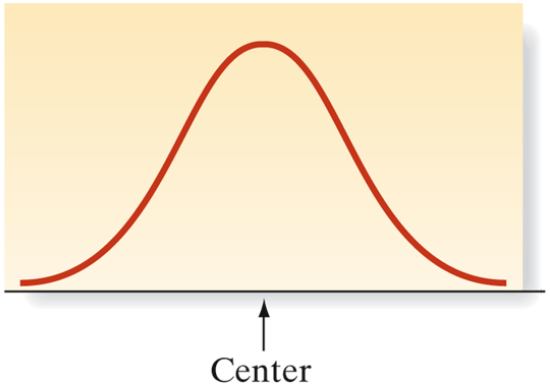
- Again, the histograms seem to indicate that BDD females tend to be more dissatisfied with their looks than males.
- We'll learn how to attach a measure of reliability to such an inference.

2.3

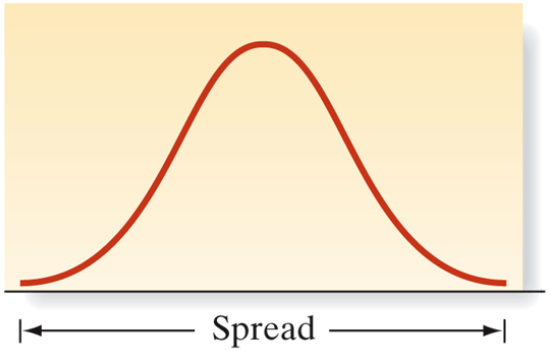
Numerical Measures of Central Tendency

Figure 2.14 Numerical descriptive measures

- Dataset → refers to either
  - sample
  - or, population
- If **statistical inference** is our goal
  - need: sample **numerical descriptive measures**
- Numerical methods to describe quantitative data measure
  - 1. The **central tendency** of the set of measurements—that is, the tendency of the data to cluster, or center, about certain numerical values.
  - 2. The **variability** of the set of measurements—that is, the spread of the data.



a.



b.

Definition

The most popular and best understood measure of central tendency for a quantitative data set is the ***arithmetic mean*** (or simply the **mean**) of the data set.

> The **mean** of a set of quantitative data is the sum of the measurements, divided by the number of measurements contained in the data set.

Formula

In everyday terms, the mean is the average value of the data set and is often used to represent a "typical" value. We denote the **mean** of a sample of measurements by $\overline{x}$ (read "x-bar") and represent the formula for its calculation as shown in the following box:

Formula for a Sample Mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

[*Note:* $\sum_{i=1}^{n} x_i = (x_1 + x_2 + \cdots + x_n)$. For more details and examples on this **summation notation**, see Appendix A.]

Example

- **Problem** Calculate the mean of the following five sample measurements: 5, 3, 8, 5, 6.
- **Solution** Using the definition of sample mean and the summation notation, we find that

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{5 + 3 + 8 + 5 + 6}{5} = \frac{27}{5} = 5.4$$

Thus, the mean of this sample is 5.4.

Figure 2.15 SAS numerical descriptive measures for 100 EPA gas mileages

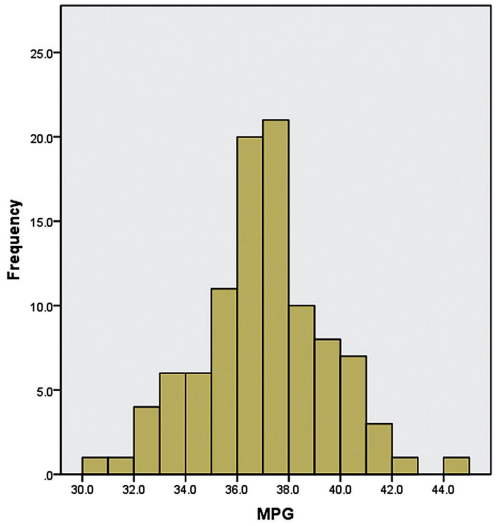
- **Problem** Calculate the sample mean for the 100 EPA mileages given in Table 2.2.

The MEANS Procedure

Analysis Variable : MPG

Mean	Std Dev	Variance	N	Minimum	Maximum	Median
36.9940000	2.4178971	5.8462263	100	30.0000000	44.9000000	37.0000000

Given this information, you can visualize a distribution of gas mileage readings centered in the vicinity of $\overline{x} \approx 37$. An examination of the relative frequency histogram confirms that $\overline{x}$ does in fact fall near the center of the distribution.



Definition

- The sample mean $\bar{x}$ will play an important role in accomplishing our objective of making inferences about populations on the basis of information about the sample.
- For this reason, we need to use a different symbol for the *mean of a population*—the mean of the set of measurements on every unit in the population.
- We use the Greek letter μ (mu) for the population mean.

Symbols for the Sample Mean and the Population Mean

In this text, we adopt a general policy of using Greek letters to represent numerical descriptive measures of the population and Roman letters to represent corresponding descriptive measures of the sample. The symbols for the mean are

$\bar{x}$ = Sample mean μ = Population mean

How much can we trust $\overline{x}$

- We'll often use the sample mean $\overline{x}$ to estimate (make an inference about) the population mean μ.
- For example, the EPA mileages for the population consisting of all cars has a mean equal to some value μ. Our sample of 100 cars yielded mileages with a mean of $\overline{x}$ = 36.9940. If, as is usually the case, we don't have access to the measurements for the entire population, we could use $\overline{x}$ as an estimator or approximator for μ.
- Then we'd need to know something about the reliability of our inference. That is, we'd need to know how accurately we might expect $\overline{x}$ to estimate μ. We'll later find that this accuracy depends on two factors:
  - 1. The size of the sample. The larger the sample, the more accurate the estimate will tend to be.
  - 2. The variability, or spread, of the data. All other factors remaining constant, the more variable the data, the less accurate is the estimate.

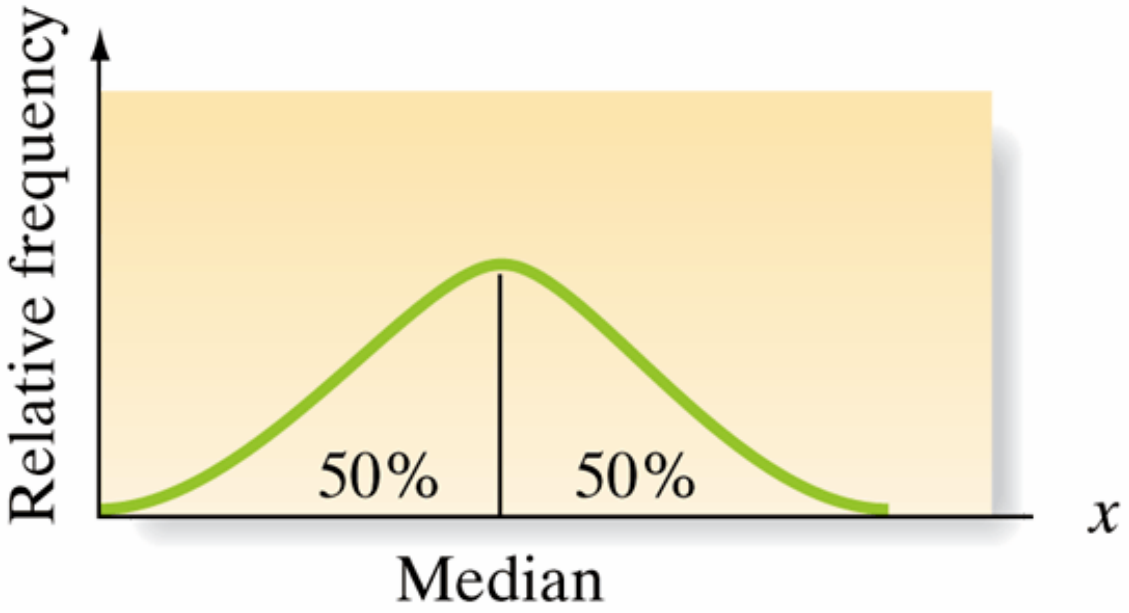
Definition

Another important measure of central tendency is the **median**.

> The **median** of a quantitative data set is the middle number when the measurements are arranged in ascending (or descending) order.

Figure 2.16 Location of the Median

- The median is of **most value** in describing large data sets.
- If a data set is characterized by a relative frequency histogram, the median is the point on the *x*-axis such that half the area under the histogram lies above the median and half lies below.



Procedure and Symbols

We denote the *median* of a *sample* by M. Like with the population mean, we use a Greek letter (η) to represent the population median.

Calculating a Sample Median M

Arrange the n measurements from the smallest to the largest.

1. If n is odd, M is the middle number.
2. If n is even, M is the mean of the middle two numbers.

Symbols for the Sample and Population Median

M = Sample median

η = Population median

Example

Problem Consider the following sample of $n = 7$ measurements: 5, 7, 4, 5, 20, 6, 2.

- a. Calculate the median M of this sample.
- b. Eliminate the last measurement (the 2), and calculate the median of the remaining n = 6 measurements.

Solution

- a. The seven measurements in the sample are ranked in ascending order: 2, 4, 5, 5, 6, 7, 20. Because the number of measurements is odd, the median is the middle measurement. Thus, the median of this sample is $\boldsymbol{M = 5}$.
- b. After removing the 2 from the set of measurements, we rank the sample measurements in ascending order as follows: 4, 5, 5, 6, 7, 20. Now the number of measurements is even, so we average the middle two measurements. The median is $M = (5 + 6) / 2 =$ **5.5**.

Median may be a better measure of central tendency than the Mean

- Why?
- Household incomes of a community being studied by a sociologist.
- The presence of just a few households with very high incomes will affect the mean more than the median.
- Thus, the median will provide a more accurate picture of the typical income for the community. The mean could exceed the vast majority of the sample measurements (household incomes), making it a misleading measure of central tendency.

Example

Problem Calculate the median for the 100 EPA mileages given in Table 2.2. Compare the median with the mean computed in Example 2.4.

Solution

- M is 37.0.
- Thus, half of the 100 mileages in the data set fall below 37.0 and half lie above 37.0.
- Note that the median, 37.0, and the mean, 36.9940, are almost equal, a relationship that indicates a **lack of skewness** in the data.
- In other words, the data exhibit a tendency to have as many measurements in the left tail of the distribution as in the right tail.

Definition

A data set is said to be **skewed** if one tail of the distribution has more extreme observations than the other tail.

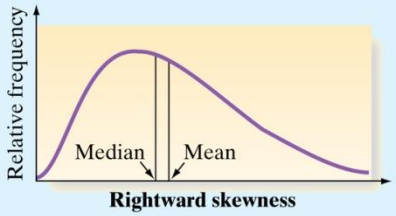
- A comparison of the mean and the median gives us a general method for detecting skewness in data sets.

Procedure

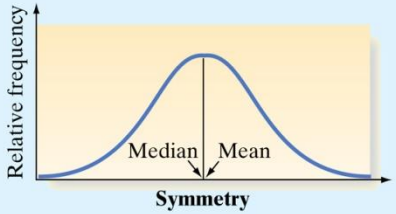
- With ***rightward skewed*** data, the right tail (high end) of the distribution has more extreme observations.
- Conversely, with ***leftward skewed*** data, the left tail (low end) of the distribution has more extreme observations.

Detecting Skewness by Comparing the Mean and the Median

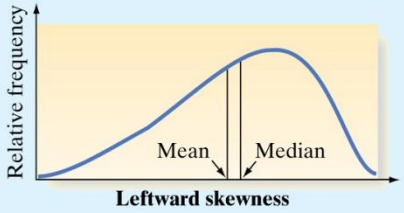
If the data set is skewed to the right, then the median is less than the mean.



If the data set is symmetric, then the mean equals the median.



If the data set is skewed to the left, then the mean is less than the median.



Definition

A third measure of central tendency is the **mode** of a set of measurements

> The **mode** is the measurement that occurs most frequently in the data set.

Therefore, the mode shows where the data **tend to concentrate**.

Example

- **Problem** Each of 10 taste testers rated a new brand of barbecue sauce on a 10-point scale, where 1 = awful and 10 = excellent. Find the mode for the following 10 ratings:

  8 7 9 6 8 10 9 9 5 7

- **Solution** Since 9 occurs most often (three times), the mode of the ten taste ratings is 9.

- Note that the data are actually qualitative in nature (e.g., “awful,” “excellent”). The mode is particularly useful for describing qualitative data. The modal category is simply the category (or class) that occurs most often.

Strengths and weaknesses of the Mode

- may be **useful** with quantitative data sets
  - modal neck size and sleeve length of potential customers
- may be **unuseful** with quantitative data sets
  - mode of EPA mileage is 37.0 (not make sense)
- **more meaningful** measure can be obtained from a relative frequency histogram
  - largest relative frequency is called the modal class.
- The mean and median provide more descriptive information than the mode for quantitative data.

Example

Problem Seismologists use the term “aftershock” to describe the smaller earthquakes that follow a main earthquake. Following the Northridge earthquake, the Los Angeles area experienced a record 2,929 aftershocks in a three-week period. The magnitudes (measured on the Richter scale) of these aftershocks as well as their interarrival times (in minutes) were recorded by the U.S. Geological Survey. Today seismologists continue to use these data to model future earthquake characteristics. Find and interpret the **mean**, **median**, and **mode** for both of these variables. Which measure of central tendency is better for describing the magnitude distribution? The distribution of interarrival times?

Figure 2.17 MINITAB descriptive statistics for earthquake data

Solution Measures of central tendency for the two variables, magnitude and interarrival time, were produced using MINITAB. The means medians, and modes are displayed in Figure 2.17.

Descriptive Statistics: MAGNITUDE, INT-TIME

Variable	N	Mean	Median	Mode	N for Mode
MAGNITUDE	2929	2.1197	2.0000	1.8	298
INT-TIME	2928	9.771	6.000	2	354

Comment on this.

Figure 2.18a MINITAB Histogram for Magnitudes of Aftershocks

- The average magnitude is **2.12**; half the magnitudes fall below **2.0**; and the most commonly occurring magnitude is **1.8**.
- a slight **rightward skewness** in the data
- any of the three measures would be **adequate** for describing the "center"

Descriptive Statistics: MAGNITUDE, INT-TIME

Variable	N	Mean	Median	Mode	N for Mode
MAGNITUDE	2929	2.1197	2.0000	1.8	298
INT-TIME	2928	9.771	6.000	2	354

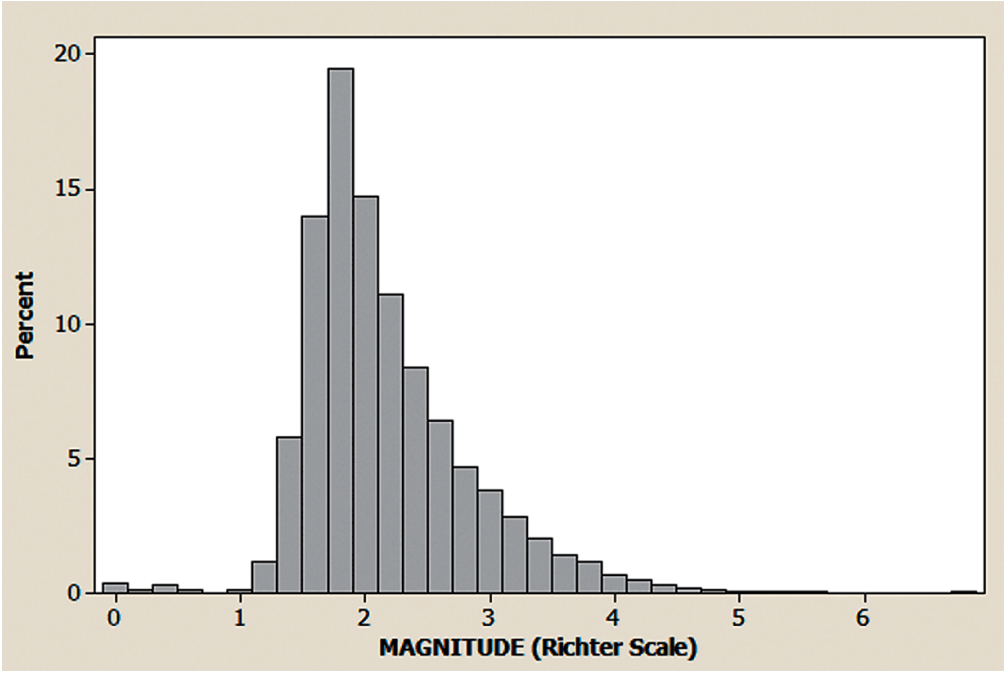
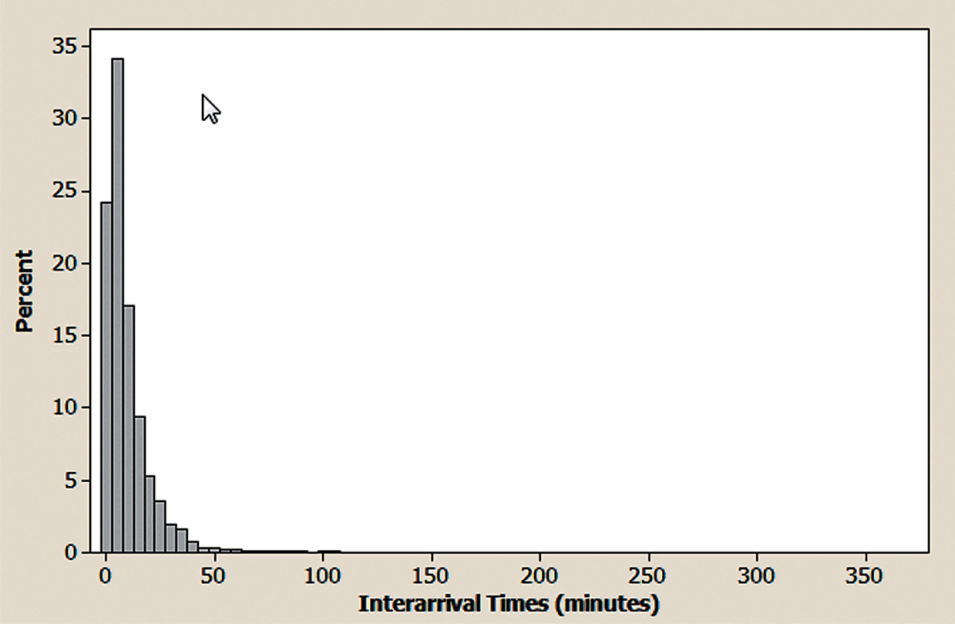


Figure 2.18b MINITAB Histogram for Inter-Arrival Times of Aftershocks

- On average, the aftershocks arrive **9.77** minutes apart; half the aftershocks have interarrival times below **6.0** minutes; and the most commonly occurring interarrival time is **2.0** minutes.
- **highly skewed to the right**
- probably want to use the median of 6.0 minutes as the "typical" interarrival time for the aftershocks



2.4

Numerical Measures of Variability

Figure 2.14 Numerical descriptive measures

- Dataset → refers to either
  - sample
  - or, population
- If **statistical inference** is our goal
  - need: sample **numerical descriptive measures**
- Numerical methods to describe quantitative data measure
  - 1. The **central tendency** of the set of measurements—that is, the tendency of the data to cluster, or center, about certain numerical values.
  - 2. The **variability** of the set of measurements—that is, the spread of the data.

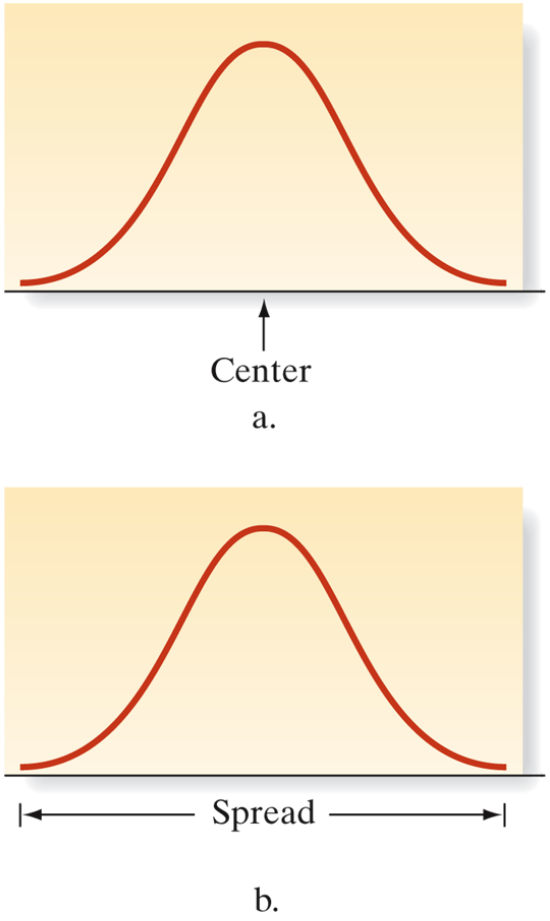
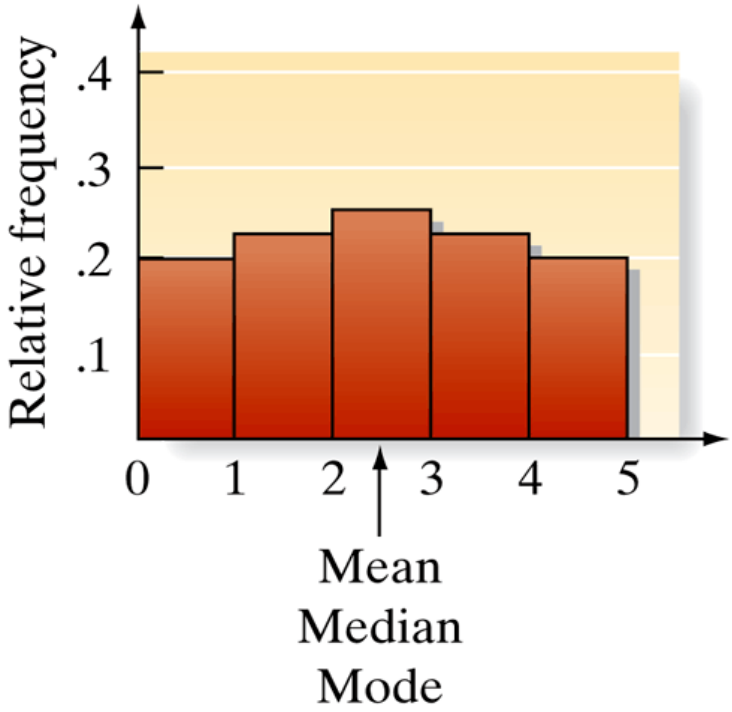
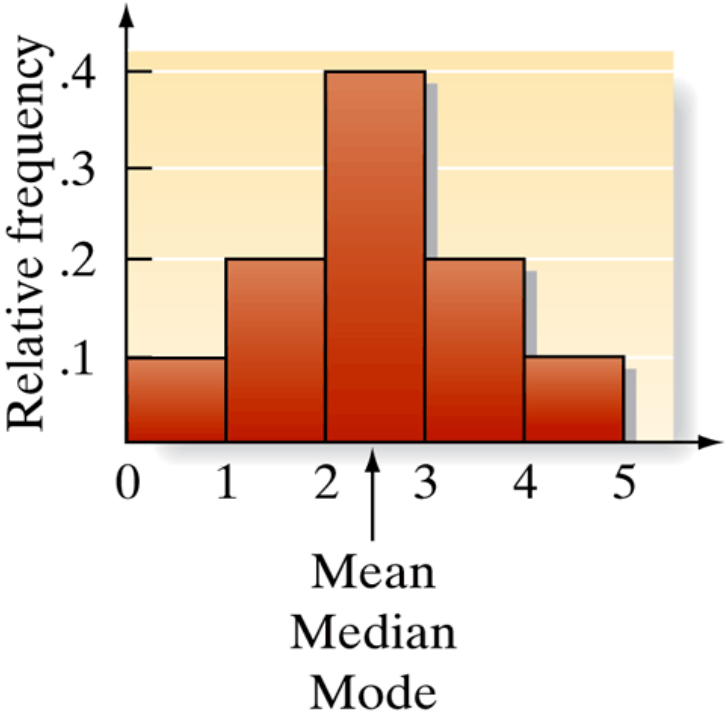


Figure 2.19 Response time histograms for two drugs

Same **mean**, **mode**, and **median**. Symmetric data.
For which drug, the response time is ***less variable***?



a. Drug A



b. Drug B

Definition

Perhaps the simplest measure of the variability of a quantitative data set is its ***range***.

The **range** of a quantitative data set is equal to the largest measurement minus the smallest measurement.

- The range is easy to compute and easy to understand, but it is a rather insensitive measure of data variation when the data sets are large.
- This is because two data sets can have the same range and be vastly different with respect to data variation.

Let's see if we can find a measure of data variation that is more sensitive than the range.

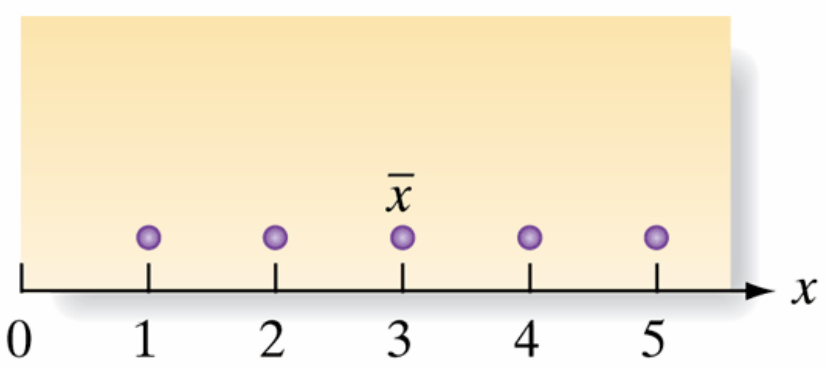
Table 2.5 Two Hypothetical Data Sets

	Sample 1	Sample 2
Measurements	$1, 2, 3, 4, 5$	$2, 3, 3, 3, 4$
Mean	$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$	$\bar{x} = \frac{2+3+3+3+4}{5} = \frac{15}{5} = 3$
Deviations of measurement values from $\bar{x}$	$(1-3), (2-3), (3-3), (4-3), (5-3)$ or $-2, -1, 0, 1, 2$	$(2-3), (3-3), (3-3), (3-3), (4-3)$ or $-1, 0, 0, 0, 1$

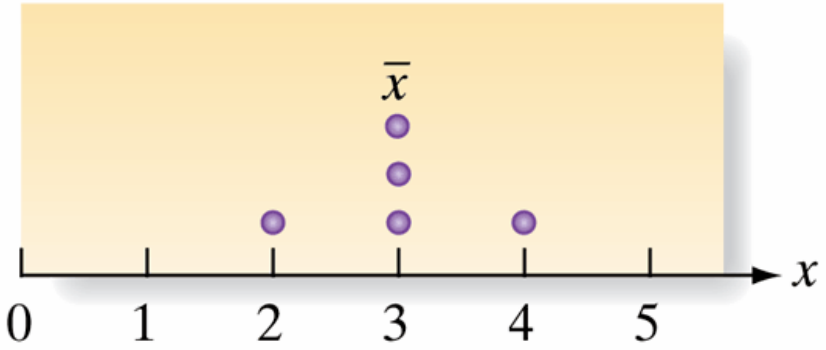
- If they tend to be large in magnitude, as in sample 1, the data are spread out, or highly variable.
- If the deviations are mostly small, as in sample 2, the data are clustered around the mean, and therefore do not exhibit much variability

Figure 2.20 Dot plots for deviations in Table 2.5

- You can see that these deviations provide information about the **variability** of the sample measurements.
- The next step is to condense the information in these distances into a single numerical measure of variability.
- Averaging the deviations from $\overline{x}$ won't help because the negative and positive deviations cancel.



a. Sample 1



b. Sample 2

Definition

To use the squared deviations calculated from a data set, we first calculate the ***sample variance***.

The **sample variance** for a sample of n measurements is equal to the sum of the squared deviations from the mean, divided by $(n - 1)$. The symbol s^2 is used to represent the sample variance.

Formula

Formula for the Sample Variance:

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

Note: A shortcut formula for calculating s^2 is

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

Example

Table 2.5 Two Hypothetical Data Sets

	Sample 1	Sample 2
Measurements	$1, 2, 3, 4, 5$	$2, 3, 3, 3, 4$
Mean	$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$	$\bar{x} = \frac{2+3+3+3+4}{5} = \frac{15}{5} = 3$
Deviations of measurement values from $\bar{x}$	$(1-3), (2-3), (3-3), (4-3), (5-3)$ or $-2, -1, 0, 1, 2$	$(2-3), (3-3), (3-3), (3-3), (4-3)$ or $-1, 0, 0, 0, 1$

Calculate the variance for sample 1.

$$s^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5-1}$$

$$= \frac{4+1+0+1+4}{4} = 2.5$$

Definition

The second step in finding a meaningful measure of data variability is to calculate the ***standard deviation*** of the data set.

The **sample standard deviation**, s, is defined as the positive square root of the sample variance, s^2, or, mathematically,

$$s = \sqrt{s^2}$$

Definition

The population variance, denoted by the symbol σ^2 (sigma squared), is the average of the squared deviations from the mean, μ, of the measurements on all units in the population, and σ (sigma) is the square root of this quantity.

Symbols for Variance and Standard Deviation

s^2 = Sample variance

s = Sample standard deviation

σ^2 = Population variance

σ = Population standard deviation

Notice that, unlike the variance, the standard deviation is expressed in the original units of measurement.

Example

Problem Calculate the variance and standard deviation of the following sample: 2, 3, 3, 3, 4.

Solution

Table 2.6 Calculating s^2		
x	$(x - \bar{x})$	$(x - \bar{x})^2$
2	−1	1
3	0	0
3	0	0
3	0	0
4	1	1
$\Sigma x = 15$		$\Sigma (x - \bar{x})^2 = 2$

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} = \frac{2}{5 - 1} = \frac{2}{4} = .5$$

$$s = \sqrt{.5} = .71$$

Figure 2.21 SAS numerical descriptive measures for 100 EPA mileages

The MEANS Procedure

Analysis Variable : MPG

Mean	Std Dev	Variance	N	Minimum	Maximum	Median
36.9940000	2.4178971	5.8462263	100	30.0000000	44.9000000	37.0000000

- You now know that the standard deviation measures the variability of a set of data, and you know how to calculate the standard deviation.
- The larger the standard deviation, the more variable the data are.
- The smaller the standard deviation, the less variation there is in the data.
- But how can we practically interpret the standard deviation and use it to make inferences? This is the topic of next week.