



Chapter 2

Frequency Distributions and Graphs



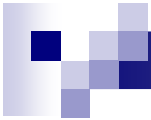
Chapter 2 Overview

- 2-1 Graphs, Pareto Diagrams, and Stem-and-Leaf Displays
- 2-2 Frequency Distributions and Histograms
- 2-3 Measures of Central Tendency
- 2-4 Measures of Dispersion
- 2-5 Measures of Position
- 2-6 Interpreting and Understanding Standard Deviation
- 2-7 The Art of Statistical Deception



2-1 Graphing Qualitative Data

- Pie charts (circle graphs), bar graphs, and pareto diagrams are used to summarize **qualitative**, or attribute, or categorical **data**.
- Pie charts (circle graphs) show the amount of data that belong to each category as a proportional part of a circle.
- Bar graphs show the amount of data that belong to each category as a proportionally sized rectangular area.
- Pareto diagrams are a bar graph with the bars arranged from the most numerous category to the least numerous category. They include a line graph displaying the cumulative percentages and counts for the bars



Example 1 – *Graphing Qualitative Data*

- Table 2.1 lists the number of cases of each type of operation performed at General Hospital last year.

Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23
<i>Total</i>	498

Operations Performed at General Hospital Last Year [TA02-01]

Table 2.1

Example 1 – *Graphing Qualitative Data*

- The data in Table 2.1 are displayed on a pie chart in Figure 2.1, with each type of operation represented by a proportion of a circle.

Relative proportion of circle
category/total #

Ex.)

proportion general operations

$$= \frac{\# \text{ general operations}}{\text{total \# operations}}$$

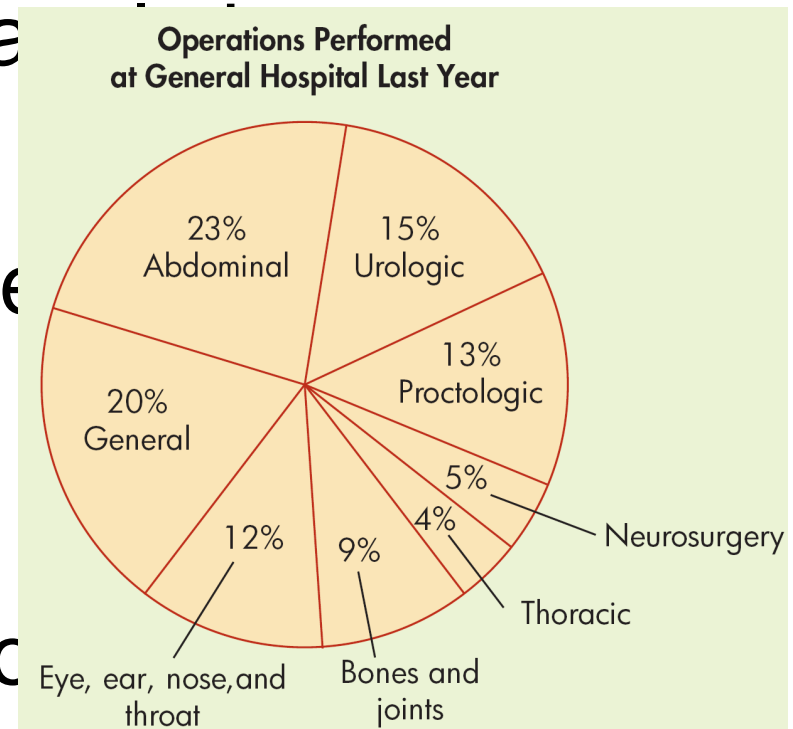
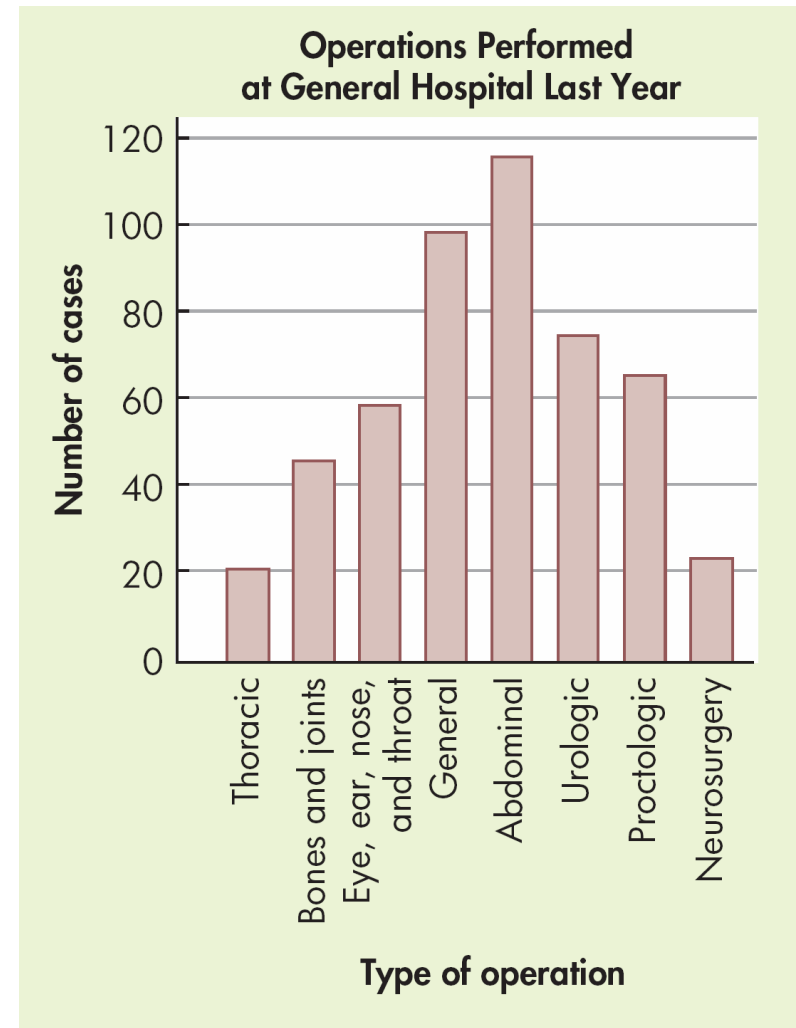


Figure 2.1

Example 1 – *Graphing Qualitative Data*

- Bar graphs of attribute data should be drawn with a space between bars of equal width.



Bar Graph

Figure 2.2



2-1 Graphing Qualitative Data

- Draw a pie graph & bar graph for the results from a Self.com survey “What is your top cold-weather beauty concern?”. The results were:
 - Dry skin – 57%
 - Chapped lips – 25%
 - Dull hair – 10%
 - Rough feet – 8%



Class Activity

- Draw a pie graph & bar graph for the results for the number of points scored for the top NBA teams in 2008-09:
 - Boston – 90
 - Chicago – 108
 - LA Lakers – 96

Example 2 – *Pareto Diagram of Hate Crimes*

■ The FBI reported the number of hate crimes by category for 2003 (<http://www.fbi.gov/>). The Pareto diagram in Figure 2.3 shows the 8715 categorized hate crimes, their percentages, and cumulative percentages.

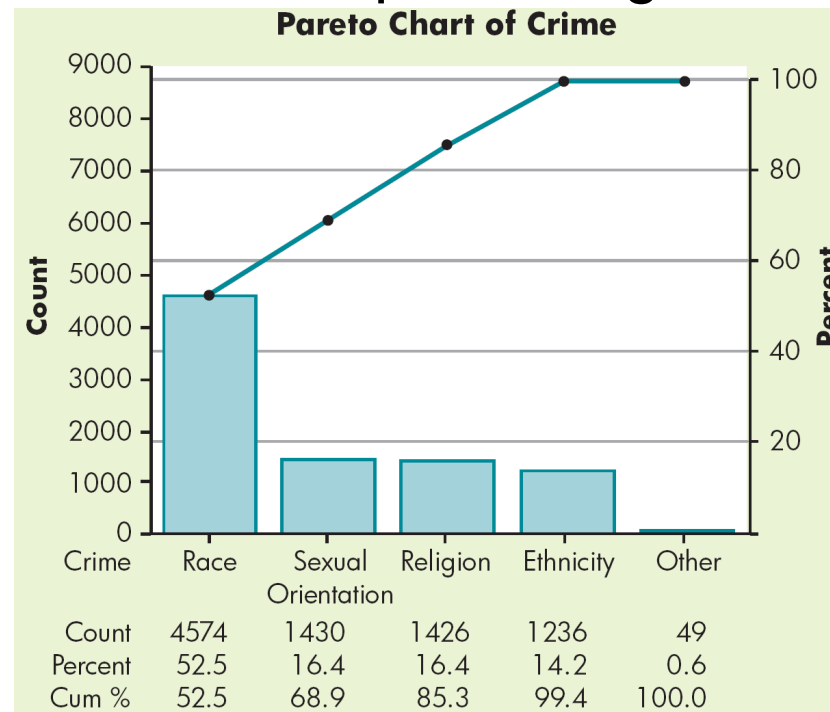


Figure 2.3

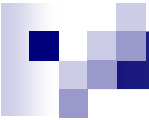


2-1 Graphing Qualitative Data

- The American Time Use Survey outlines the time use of an average weekday for full-time university and college students.

Construct a pareto diagram depicting the average time use for full-time college students.

Category	Hours
Sleeping	8.3
Leisure and sports	3.9
Educational activities	3.2
Working & related activities	3.0
Eating and drinking	1.0
Traveling	1.5
Grooming	0.8
Other	2.3
Total	24.0



2-1 Graphing Quantitative Data

- One major reason for constructing a graph of **quantitative data** is to display its *distribution*.
- **Distribution** The pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable.



2-1 Graphing Quantitative Data

- **Dotplot display** Displays the data of a sample by representing each data value with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale. We will skip this.

- **Stem-and-leaf display** Displays the data of a sample using the actual digits that make up the data values. Each numerical value is divided into two parts: The leading digit(s) becomes the stem, and the trailing digit(s) becomes the leaf. The stems are located along the main axis, and a leaf for each data value is located so as to display the distribution of the data.

- **Histogram** We will discuss in the next section.

- **Ogive** We will discuss in the next section.



Example 3 - Creating a Stem & Leaf Plot

The growth (cm) for a variety of plant after 20 days is shown. Construct a stem & leaf plot for the data.

20	12	39	38	41	43	51
52	59	55	53	59	50	58
35	38	23	32	43	53	



Example 3 - Creating a Stem & Leaf Plot (continued)

Step 1: Arrange the data in order

12, 20, 23, 32, 35, 38, 38, 39, 41, 43,
43, 50, 51, 52, 53, 53, 55, 58, 59, 59

Step 2: Separate the data according to
first digit

12

20, 23

32, 35, 38, 38, 39

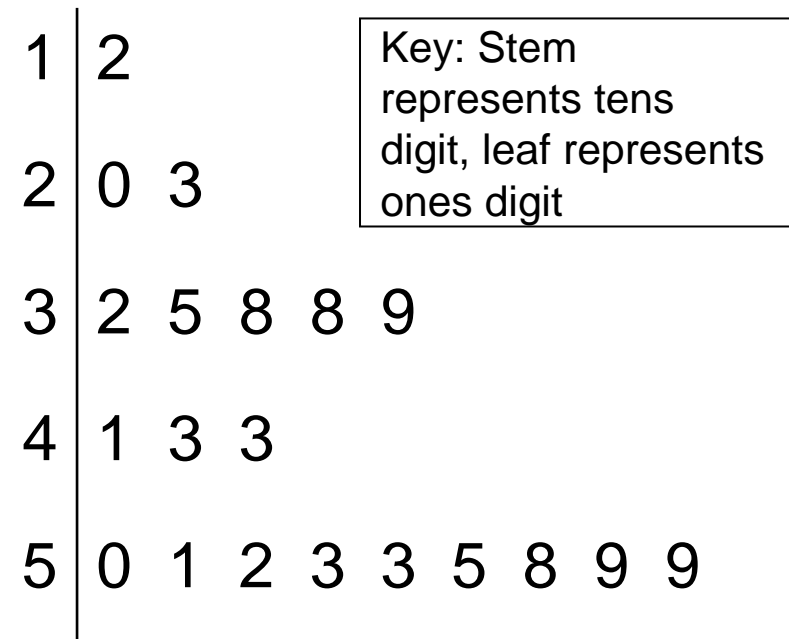
41, 43, 43

50, 51, 52, 53, 53, 55, 58, 59, 59

Example 3 - Creating a Stem & Leaf Plot (continued)

Step 3: Create a display using the leading digit as the *stem* and the trailing digit as the *leaf*.

12
20, 23
32, 35, 38, 38, 39
41, 43, 43
50, 51, 52, 53, 53, 55, 58, 59, 59





2-1 Graphing Quantitative Data

- Construct a stem-and-leaf display of the number of points scored during each basketball game last season:

□ 56 54 61 71 46 61 55 68
60 66 54 61 52 36 64 51



Class Activity

- The following lists the median house selling prices (in \$1000s) for the 20 suburbs of Rochester, New York, as listed in July 18, 2009, Democrat & Chronicle.

- 160 125 122 89 100 110 94 125
108 235 133 121 190 175 218 130
180 113 156 114

- Construct a stem & leaf display for the data.



2-2 Frequency Distributions and Histograms

- Data collected in original form is called **raw data**.
 - The table to the right is an example of raw data. It shows some common cereals & their sugar content.

Name	Sugars (grams)
Cap'n'Crunch	12
Cheerios	1
Cinnamon_Toast_Crunch	9
Cocoa_Puffs	13
Frosted_Flakes	11
Frosted_Mini-Wheats	7
Fruity_Pebbles	12
Lucky_Charms	12
Multi-Grain_Cheerios	6
Raisin_Bran	12
Rice_Krispies	3
Shredded_Wheat	0
Special_K	3



2-2 Frequency Distributions and Histograms

- A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.
 - In the previous example, we are trying to classify cereals & sugar content. It is easier to look at the data if we use a frequency distribution.



2-2 Frequency Distributions and Histograms

■ **Creating a Frequency Distribution**

- 1st: Determine Classes and Class Boundaries
- 2nd: Tally the number of entries that fall into each class
- 3rd: Add up the tally for each class to obtain the frequency



Ex. 4: Creating a Frequency Distribution

■ 1st: Determine Classes and Class Boundaries

We will choose sugar content as our class of interest & try to divide it up into class boundaries.

Notice that sugar content varies from 0 grams to 13 grams. We will choose our class boundaries to be sugar levels from 0-4 grams, 5-9 grams, and 10-14 grams (this can vary).

Name	Sugars (grams)
Cap'n'Crunch	12
Cheerios	1
Cinnamon_Toast_Crunch	9
Cocoa_Puffs	13
Frosted_Flakes	11
Frosted_Mini-Wheats	7
Fruity_Pebbles	12
Lucky_Charm	12
Multi-Grain_Cheerios	6
Raisin_Bran	12
Rice_Krispies	3
Shredded_Wheat	0
Special_K	3

Ex. 4: Creating a Frequency Distribution (continued)

■ **2nd: Tally the number of entries that fall into each class**

Class Boundaries	Tally
0-4 grams	IIII
5-9 grams	III
10-14 grams	IIIIII

Name	Sugars (grams)
Cap'n'Crunch	12
Cheerios	1
Cinnamon_Toast_Crunch	9
Cocoa_Puffs	13
Frosted_Flakes	11
Frosted_Mini-Wheats	7
Fruity_Pebbles	12
Lucky_Charm	12
Multi-Grain_Cheerios	6
Raisin_Bran	12
Rice_Krispies	3
Shredded_Wheat	0
Special_K	3

Ex. 4: Creating a Frequency Distribution (continued)

■ **3rd: Add up the tally for each class to obtain the frequency**

Class Boundaries	Tally	Frequency
0-4 grams	IIII	4
5-9 grams	III	3
10-14 grams	IIIII	6
	Total	13

Name	Sugars (grams)
Cap'n'Crunch	12
Cheerios	1
Cinnamon_Toast_Crunch	9
Cocoa_Puffs	13
Frosted_Flakes	11
Frosted_Mini-Wheats	7
Fruity_Pebbles	12
Lucky_Charm	12
Multi-Grain_Cheerios	6
Raisin_Bran	12
Rice_Krispies	3
Shredded_Wheat	0
Special_K	3



Grouped Frequency Distribution

- **Grouped frequency distributions** are used when the range of the data is large.
 - Similar to the cereal & sugar content example (usually on a larger scale)

Class Boundaries	Tally	Frequency
0-4 grams	IIII	4
5-9 grams	III	3
10-14 grams	IIIIII	6
	Total	13



Grouped Frequency Distribution

- The smallest and largest possible data values in a class are the ***lower*** and ***upper class boundaries***.
 - For the first class boundary:
 - Lower class boundary is 0
 - Upper class boundary is 4

Class Boundaries	Tally	Frequency
0-4 grams	IIII	4
5-9 grams	III	3
10-14 grams	IIIIII	6
	Total	13



Grouped Frequency Distribution

- The **class width** can be calculated by subtracting
 - successive lower class boundaries
 - First 2 lower class boundaries: class width is $5 - 0 = 5$
 - successive upper class boundaries
 - First 2 upper class boundaries: class width is $9 - 4 = 5$

Class Boundaries	Tally	Frequency
0-4 grams	IIII	4
5-9 grams	III	3
10-14 grams	IIIIII	6



Grouped Frequency Distribution

- The ***class midpoint X_m*** can be calculated by averaging
 - upper and lower class boundaries
 - Midpoint for the first class
 - $(0+4)/2 = 2$
 - Midpoint for the second class
 - $(5+9)/2 = 7$
 - Midpoint for the third class
 - $(10+14)/2 = 12$

Class Boundaries	Tally	Frequency
0-4 grams	IIII	4
5-9 grams	III	3
10-14 grams	IIIIII	6



Class Activity

- Answer the following questions about the frequency distribution representing 100 resort club managers and their annual salaries:

Ann. Sal. (\$1000)	15-24	25-34	35-44	45-54	55-64
No. Mgrs.	12	37	26	19	6

- What are the upper & lower class boundaries for the first class?
- What is the class width?
- What is the class midpoint for the third class?



Guidelines for Frequency Distributions


1. 5-15 Categories
2. Equal Intervals
3. Classes should not overlap
4. Avoid Open-Ended Classes
5. Endpoints and class widths should be “nice numbers”



Example: Constructing a Frequency Distribution

The number of passengers (in thousands) for the leading US passenger airlines in 2004 is indicated below. Use the data to construct a frequency distribution.

91	86	81	70	55	42	40	21	16	14	13	13	13	12
11	10	10	9	7	6	6	6	5	5	5			



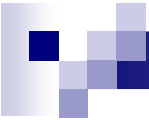
Example: Constructing a Frequency Distribution (continued)

STEP 1 Determine the classes.

Stem & Leaf Display:

91 86 81 70 55 42 40
21 16 14 13 13 13 12
11 10 10 9 7 6 6
6 5 5 5

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



Example: Constructing a Frequency Distribution (continued)

- We will use the following intervals:

Interval

90-99

80-89

70-79

60-69

50-59

40-49

30-39

20-29

10-19

0-9

Example: Constructing a Frequency Distribution (continued)

STEP 2 Tally the data.

(It helps if you order the data first)

91 86 81 70 55 42
40 21 16 14 13 13
13 12 11 10 10 9
7 6 6 6 5 5
5

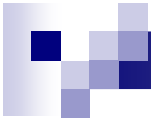
Interval	Tally
90-99	I
80-89	II
70-79	I
60-69	
50-59	I
40-49	II
30-39	
20-29	I
10-19	IIIIIIII
0-9	IIIIIIII 33



Example: Constructing a Frequency Distribution (continued)

STEP 3 Add up
the Tally category.

Interval	Tally	Frequency
90-99	I	1
80-89	II	2
70-79	I	1
60-69		0
50-59	I	1
40-49	II	2
30-39		0
20-29	I	1
10-19	IIIIIIII	9
0-9	IIIIIIII	8
		34



Example: Constructing a Frequency Distribution (continued)

STEP 4

Create Cumulative Frequency Column if needed.

Interval	Frequency	Cumulative Frequency
90-99	1	1
80-89	2	3
70-79	1	4
60-69	0	4
50-59	1	5
40-49	2	7
30-39	0	7
20-29	1	8
10-19	9	17
0-9	8	25
		35



2-1 Constructing a Frequency Distribution

- All of the third graders at Roth Elementary School were given a physical fitness strength test. The following data resulted:

□ 12 22 6 9 2 9 5 9
 3 5 16 1 22 18 6 12
 21 23 9 10 24 21 17 11

- Construct a grouped frequency distribution for the data.



Class Activity

- The hemoglobin A1c test, a blood test given to diabetic patients during their checkups, indicates the level of control of blood sugar during the past 2-3 months. The following data were obtained for diabetic patients at a clinic:

- 6.5 5.0 7.6 4.8 8.0 7.5 7.9 8.0
9.2 6.4 6.0 5.6 6.0 5.7 9.2 8.1
8.0 6.5 6.6 5.0 8.0 6.5 6.1 6.4

- Construct a grouped frequency distribution for the data.



2-2 Frequency Distributions and Histograms

The ***histogram*** is a graph that displays the data by using vertical bars of various heights to represent the frequencies of the classes.

The class boundaries or class midpoints are represented on the horizontal axis.



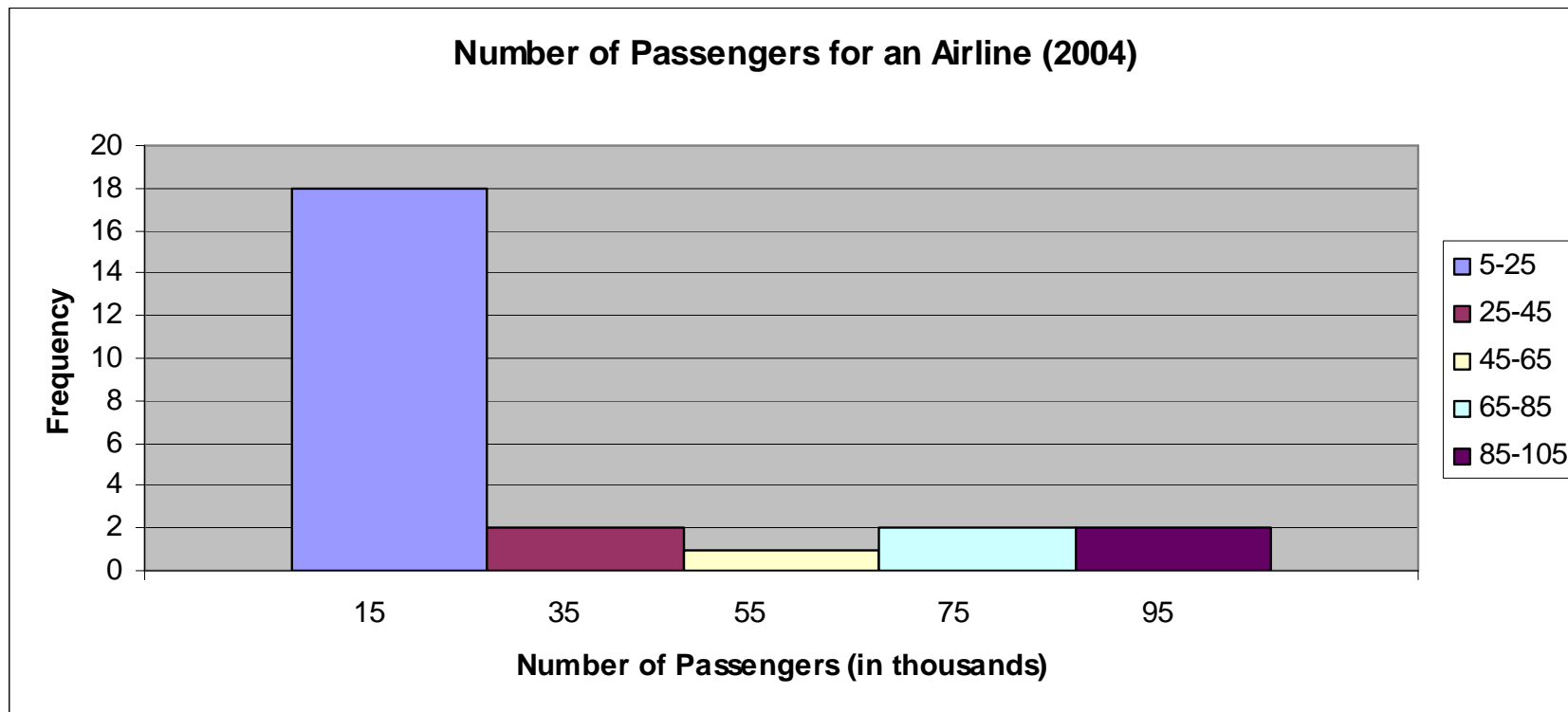
Ex. 6: Constructing Histograms

Construct a histogram for the following frequency distribution:

Histograms use class boundaries or midpoints and frequencies of the classes.

Class Boundaries	Class Midpoints	Frequency
5 - 25	$(5+25)/2 = 15$	18
25 - 45	$(25+45)/2 = 35$	2
45 - 65	$(45+65)/2 = 55$	1
65 - 85	$(65+85)/2 = 75$	2
85 - 105	$(85+105)/2 = 95$	2

Ex. 6: Constructing Histograms (continued)





2-2 Histograms

- Construct a histogram for the following frequency distribution using class boundaries.

Ann. Sal. (\$1000)	15-24	25-34	35-44	45-54	55-64
No. Mgrs.	12	37	26	19	6



Class Activity

- During the Spring 2009 semester, 200 students took a statistics test from a particular instructor. The resulting grades are given in the following table:

Test Grades	Number
50-60	13
60-70	44
70-80	74
80-90	59
90-100	9
100-110	1
Total	200

- Draw a histogram of the statistics test grades.



2-2 Frequency Distributions and Histograms

- The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.
- The upper class boundaries are represented on the horizontal axis.



Ex. 7: Creating Ogives

Construct an ogive for the following frequency distribution.

Ogives use upper class boundaries and cumulative frequencies of the classes.

Class Boundaries	Frequency	Cumulative Frequency
5 - 25	18	18
25 - 45	2	20
45 - 65	1	21
65 - 85	2	23
85 - 105	2	25



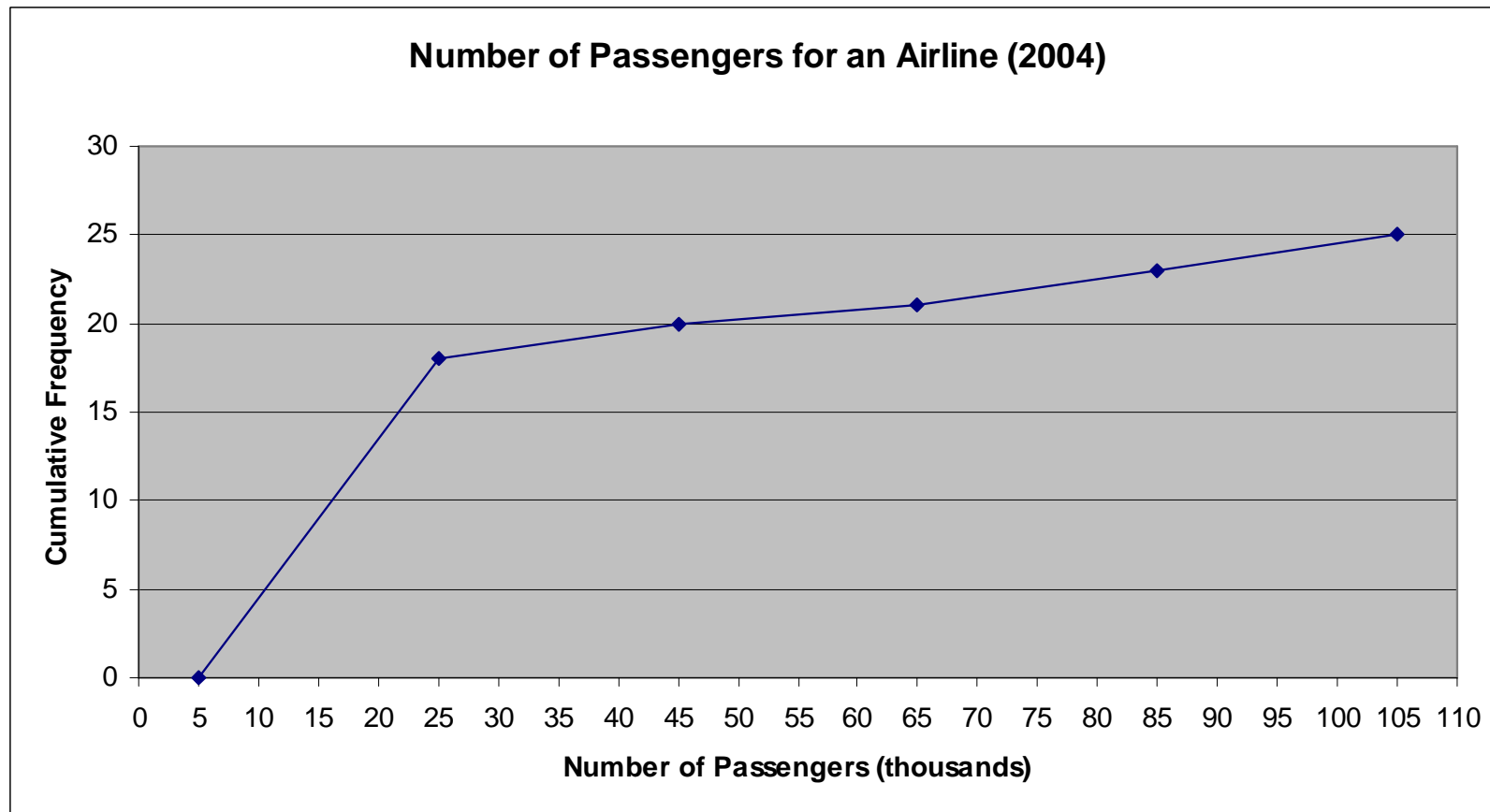
Ex. 7: Creating Ogives (continued)

Ogives use upper class boundaries and cumulative frequencies of the classes.

Class Boundaries	Cumulative Frequency
Less than 5	0
Less than 25	18
Less than 45	20
Less than 65	21
Less than 85	23
Less than 105	25

Ex. 7: Creating Ogives (continued)

Ogives use upper class boundaries and cumulative frequencies of the classes.





2-2 Ogives

- Create an ogive for the following distribution:

Ann. Sal. (\$1000)	15-24	25-34	35-44	45-54	55-64
No. Mgrs.	12	37	26	19	6



Procedure Table

Constructing Statistical Graphs


- 1: Draw and label the x and y axes.
- 2: Choose a suitable scale for the frequencies or cumulative frequencies, and label it on the y axis.
- 3: Represent the class boundaries for the histogram or ogive on the x axis.
- 4: Plot the points and then draw the bars or lines.



2-2 Frequency Distributions and Histograms

If proportions are used instead of frequencies, the graphs are called ***relative frequency graphs***.

Relative frequency graphs are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class.



Ex. 8: Creating Relative Frequency Graphs

Construct a histogram and ogive using relative frequencies for the distribution (shown here) of the potassium content (mg) of 75 breakfast cereals.

Class Boundaries	Frequencies
14.5-59.5	27
59.5-104.5	19
104.5-149.5	16
149.5-194.5	6
194.5-239.5	2
239.5-284.5	3
284.5-330.5	2
Total	75

Ex. 8: Creating Relative Frequency Graphs - Histograms

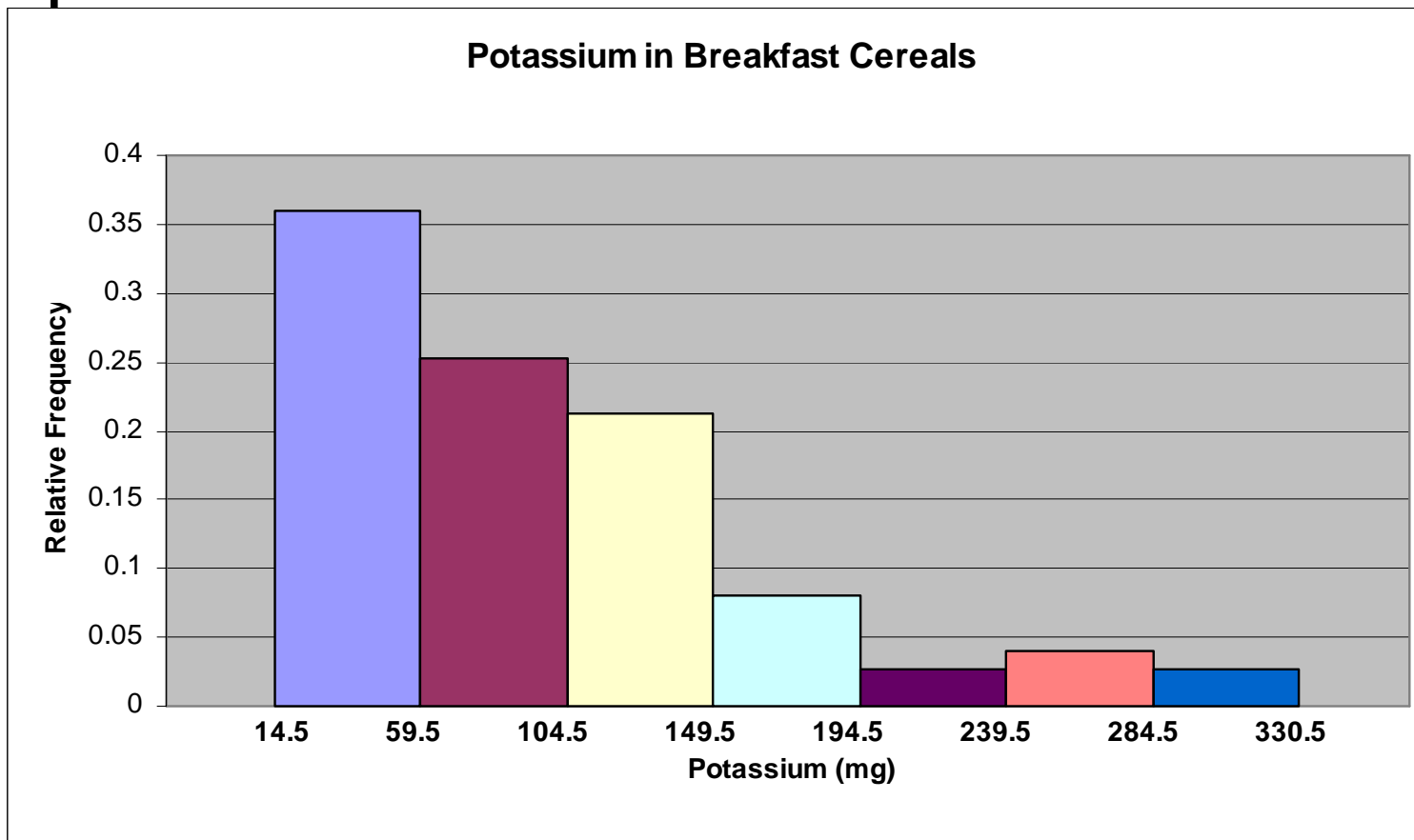
The following is a frequency distribution of the potassium content (mg) of 75 cereals.

Class Boundaries	Frequency	Relative Frequency
14.5-59.5	27	$27/75 = 0.36$
59.5-104.5	19	$19/75 = 0.25$
104.5-149.5	16	$16/75 = 0.21$
149.5-194.5	6	$6/75 = 0.08$
194.5-239.5	2	$2/75 = 0.03$
239.5-284.5	3	$3/75 = 0.04$
284.5-330.5	2	$2/75 = 0.03$
Total	75	1

Divide each frequency by the total frequency to get the relative frequency.

Ex. 8: Creating Relative Frequency Graphs - Histograms

Use the class boundaries and the relative frequencies of the classes.





Ex. 8: Creating Relative Frequency Graphs - Ogives

The following is a frequency distribution of potassium content (mg) in cereals.

Class Boundaries	Frequency	Cumulative Frequency	Cum. Rel. Frequency
14.5-59.5	27	27	0.36
59.5-104.5	19	46	0.61
104.5-149.5	16	62	0.83
149.5-194.5	6	68	0.91
194.5-239.5	2	70	0.93
239.5-284.5	3	73	0.97
284.5-330.5	2	75	1



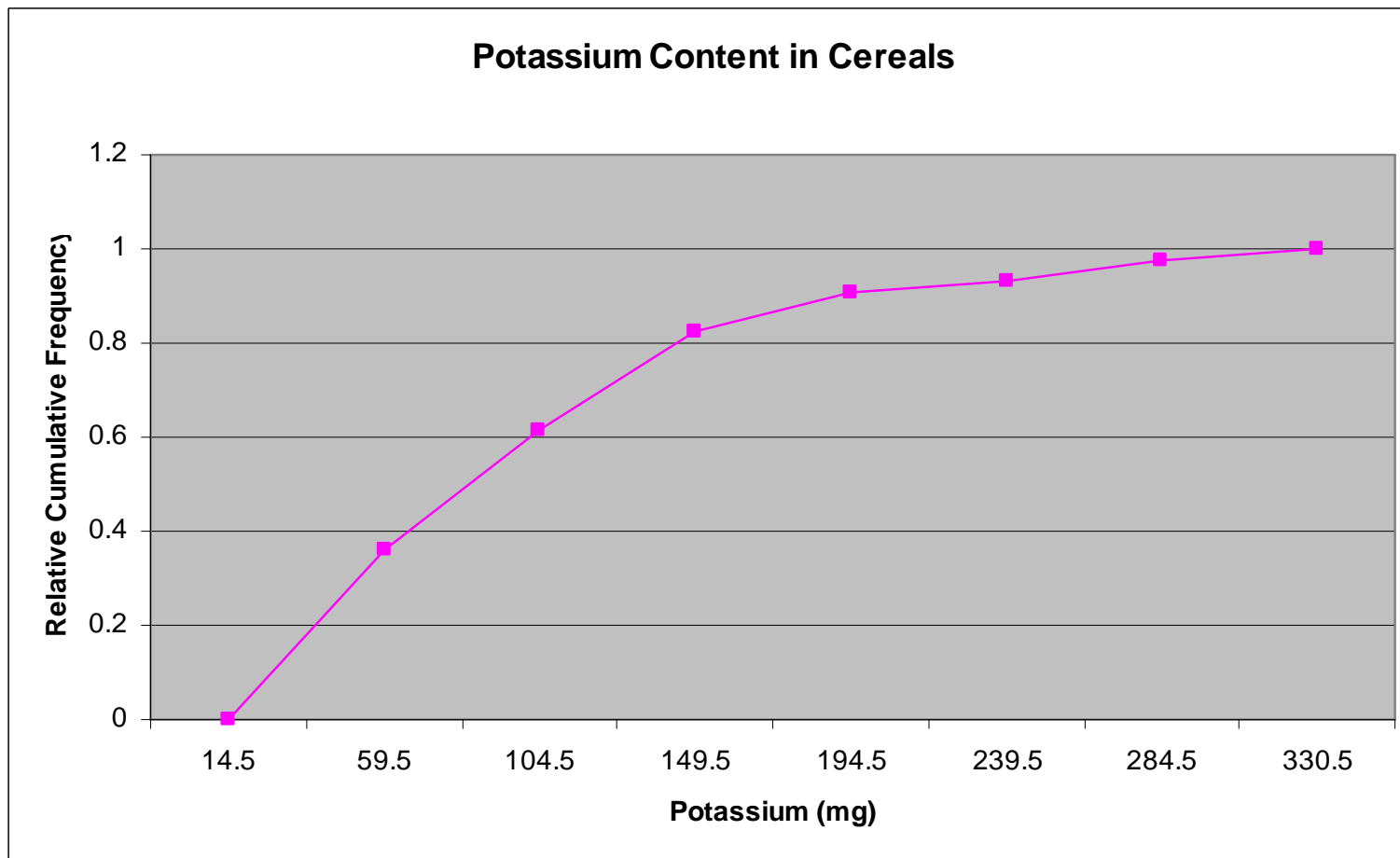
Ex. 8: Creating Relative Frequency Graphs - Ogives

Ogives use upper class boundaries and cumulative frequencies of the classes.

Class Boundaries	Cum. Rel. Frequency
Less than 59.5	0.36
Less than 104.5	0.61
Less than 149.5	0.83
Less than 194.5	0.91
Less than 239.5	0.93
Less than 285.5	0.97
Less than 330.5	1

Ex. 8: Creating Relative Frequency Graphs - Ogives

Use the upper class boundaries and the cumulative relative frequencies.





2-2 Creating Relative Frequency Graphs

- Create a relative frequency histogram and ogive for the following table:

Ann. Sal. (\$1000)	15-24	25-34	35-44	45-54	55-64
No. Mgrs.	12	37	26	19	6

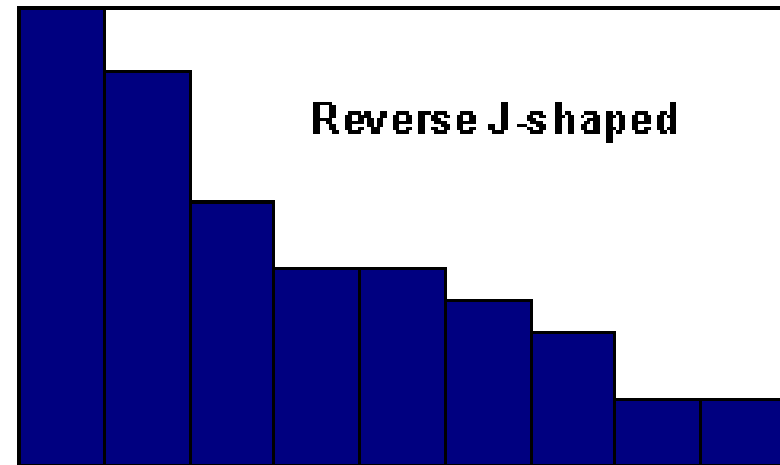
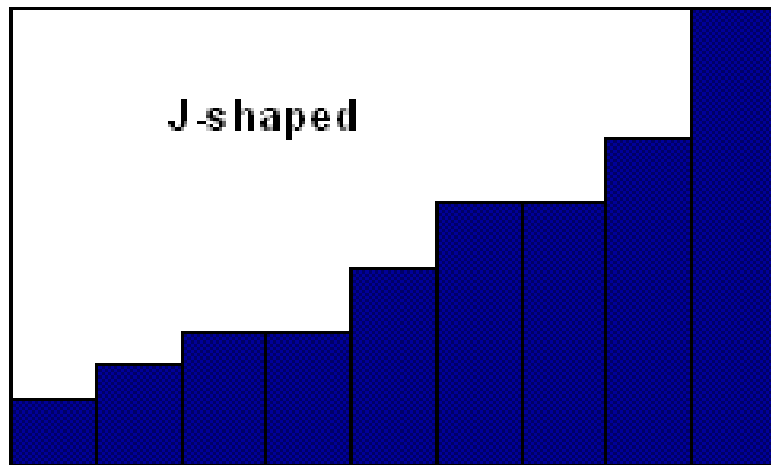
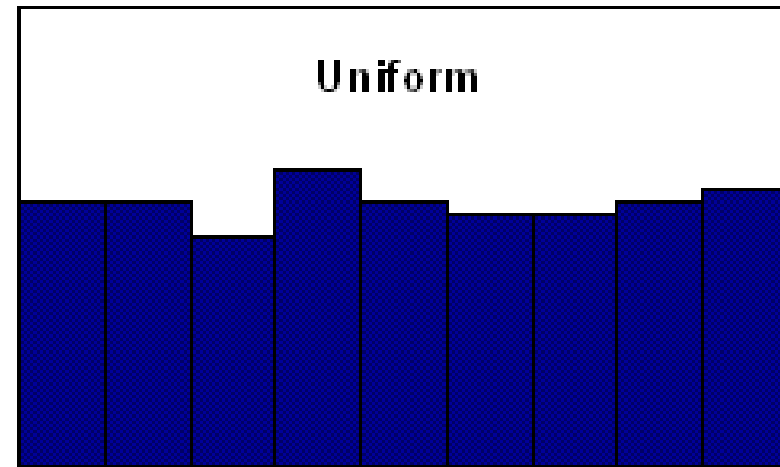
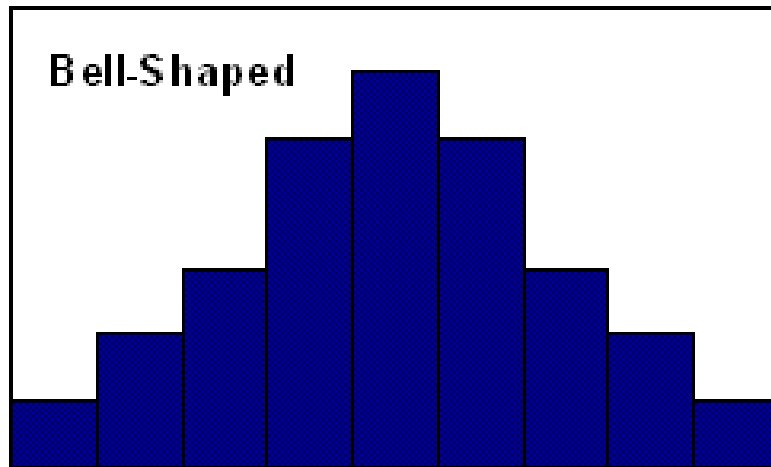


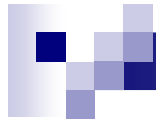
Class Activity

- Create a relative frequency histogram for the following table:

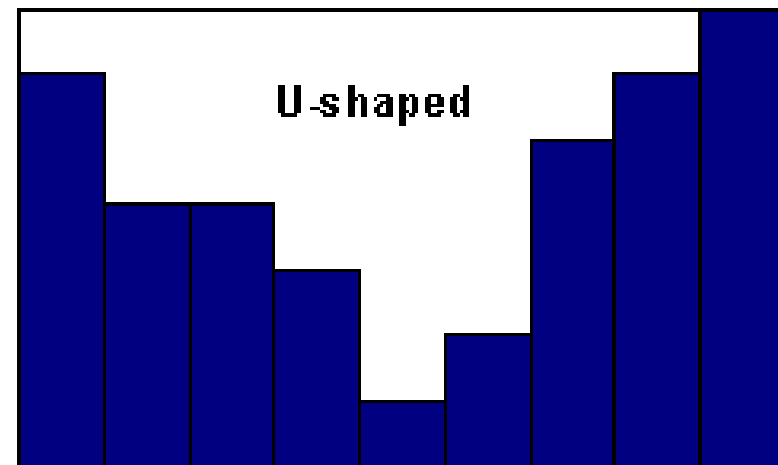
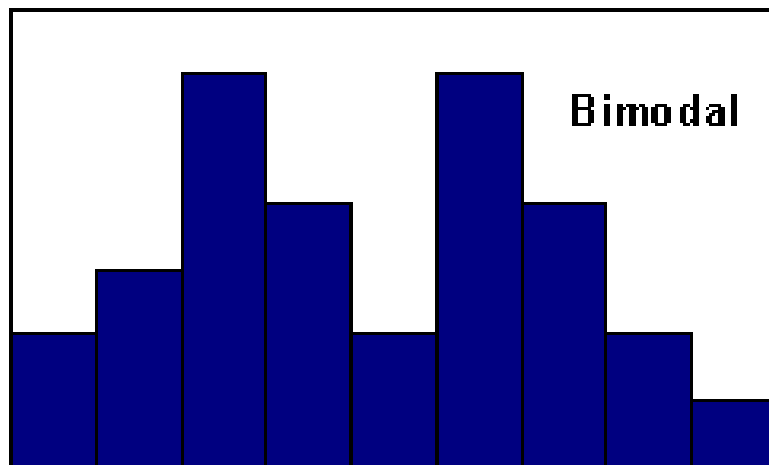
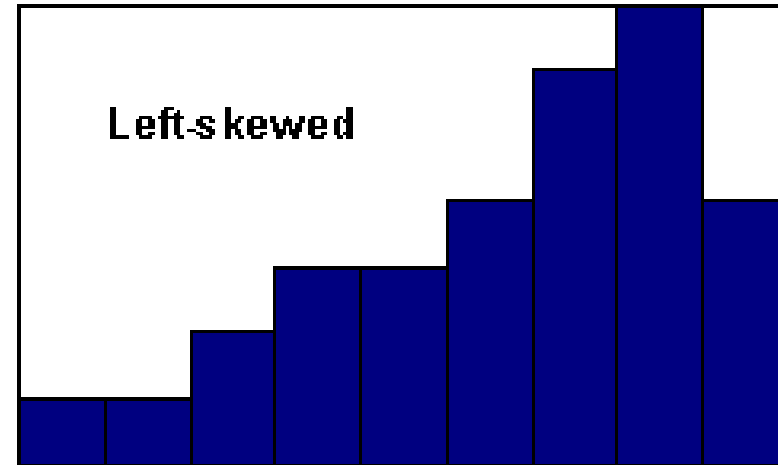
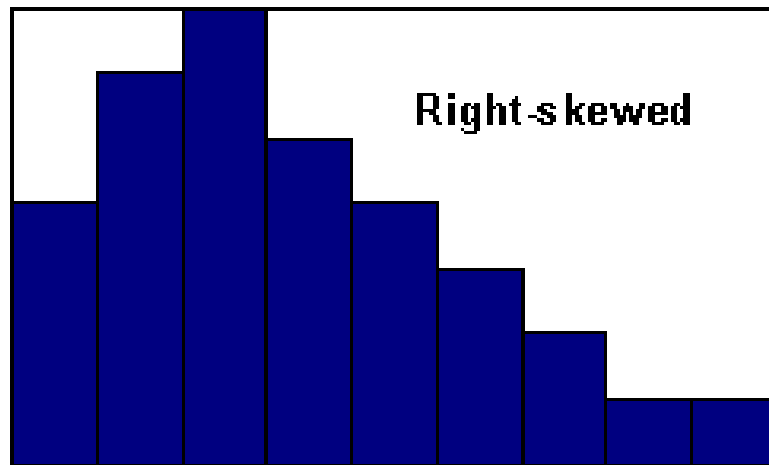
Test Grades	Number
50-60	13
60-70	44
70-80	74
80-90	59
90-100	9
100-110	1
Total	200

Shapes of Distributions



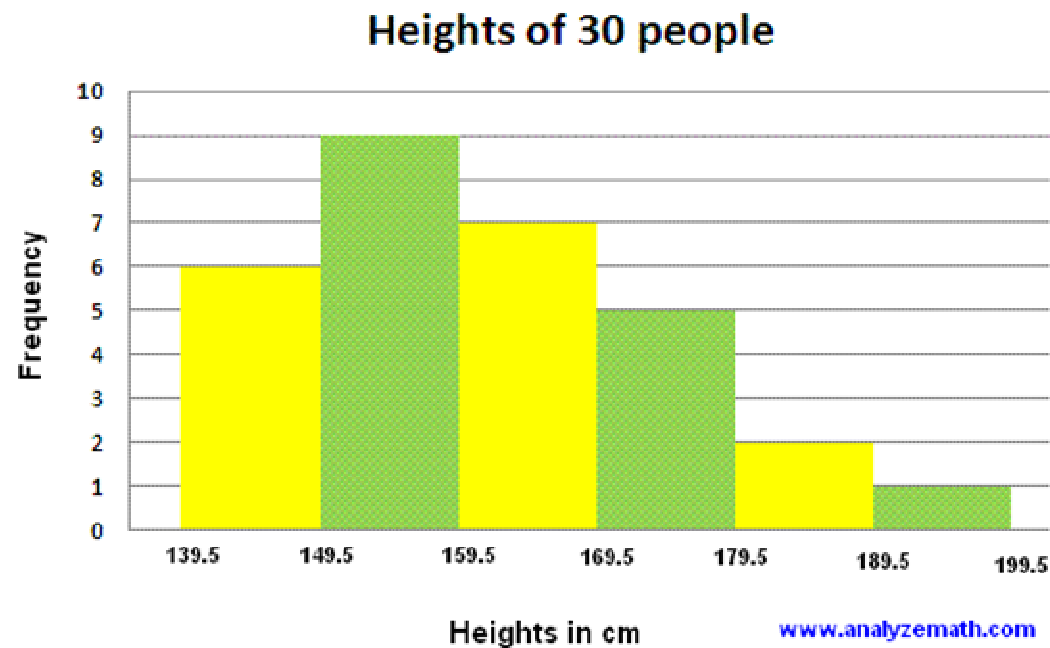


Shapes of Distributions



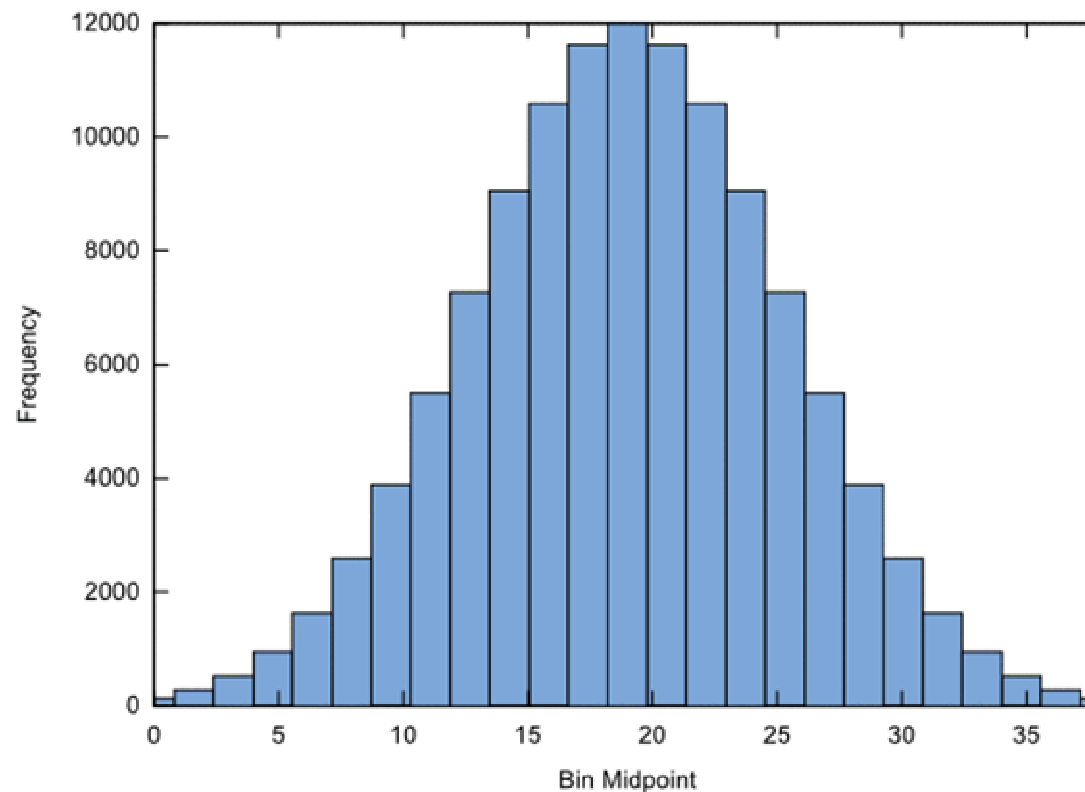
Class Activity

- What is the distribution of the following histograms:



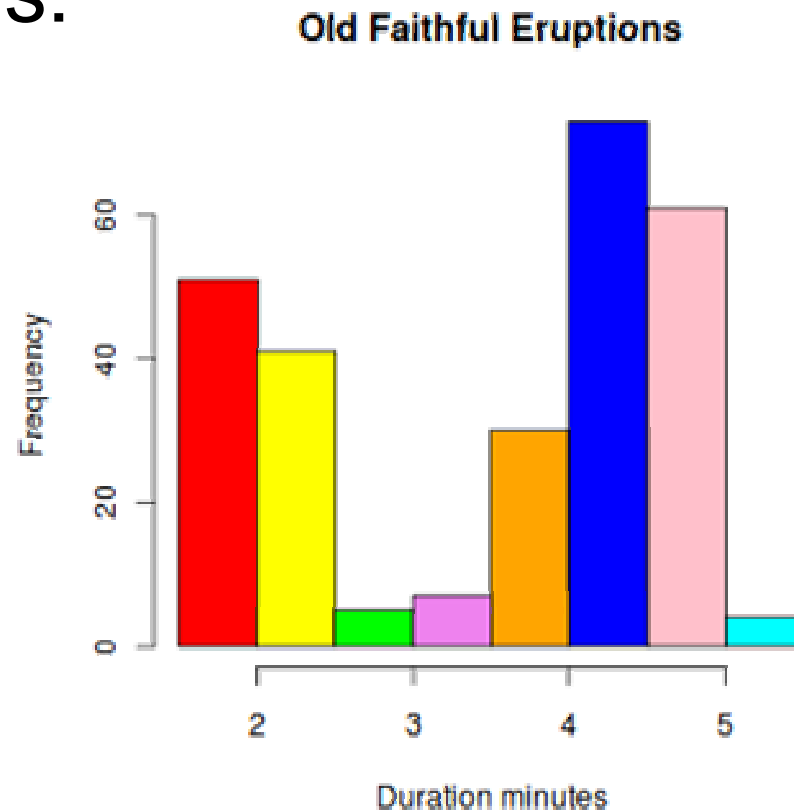
Class Activity

- What is the distribution of the following histograms:



Class Activity

- What is the distribution of the following histograms:





2-3 Measures of Central Tendency

General Rounding Rule

The basic rounding rule is that rounding should not be done until the final answer is calculated. Use of parentheses on calculators or use of spreadsheets help to avoid early rounding error.



2-3 Measures of Central Tendency

What Do We Mean By **Average**?

- ☐ Mean
- ☐ Median
- ☐ Mode
- ☐ Midrange
- ☐ Weighted Mean



2-3 Measures of Central Tendency: Mean

- The **mean** is the quotient of the sum of the values and the total number of values.
- The symbol \bar{X} is used for sample mean.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$



Ex.) 9 - Calculating the Mean

The data represent the Richter magnitudes for 12 major earthquakes in a region. Find the mean.

7.0, 7.2, 6.2, 5.4, 7.7, 6.4, 8.0, 6.5, 6.4, 7.2,
6.2, 5.4

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

$$\begin{aligned}\bar{X} &= \frac{7.0 + 7.2 + 6.2 + 5.4 + 7.7 + 6.4 + 8.0 + 6.5 + 6.4 + 7.2 + 6.2 + 5.4}{12} \\ &= \frac{79.6}{12} = 6.63\end{aligned}$$

The mean Richter magnitude is 6.63.



2-3 Mean

- Find the mean of the following:

☐ 3 5 3 8 6



Rounding Rule: Mean


The mean should be rounded to one more decimal place than occurs in the raw data.

The mean, in most cases, is not an actual data value.



2-3 Measures of Central Tendency: Median

- The **median** is the midpoint of the data array. The symbol for the median is MD.
- The median will be one of the data values if there is an odd number of values.
- The median will be the average of two data values if there is an even number of values.



Ex.) 10 – Finding the Median (odd number of data values)

An athlete ran the following times in seconds for the 200 meter dash in a given session. Find the median time.

26.1, 25.6, 25.7, 25.2, 25.0

Step 1: Sort in ascending order.


25.0 25.2 25.6 25.7 26.1



Step 2: Select the middle value.

MD = 25.6

The median is 25.6 seconds.



Ex.) 11 – Finding the Median (even number of data values)

The data represent the Richter magnitudes for 12 major earthquakes in a region. Find the median.

7.0, 7.2, 6.2, 5.4, 7.7, 6.4, 8.0, 6.5, 6.4, 7.2,
6.2, 5.4

Step 1: Sort in ascending order.

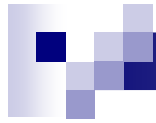
5.4 5.4 6.2 6.2 6.4 6.4 6.5 7.0 7.2 7.2 7.7 8.0



Step 2: Average the 2 middle values

$$\text{MD} = \frac{6.4 + 6.5}{2} = \frac{12.9}{2} = 6.45$$

The median is 6.45.



2-3 Median

- Find the median of the following:

☐ 3 5 3 8 6



Measures of Central Tendency:

Mode

- The **mode** is the value that occurs most often in a data set.
- It is sometimes said to be the most typical case.
- There may be no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal).



Ex.) 12 – Finding the Mode

Consider the DDT levels (in ppm) in a sample of eggs taken from a wintering ground in Mexico:

21, 18, 35, 21, 21, 21, 70, 25, 21, 25, 25, 21, 25,
18, 25, 21, 35, 50, 21, 25

Step 1: Sort the data (unnecessary but helpful)

18, 18, 21, 21, 21, 21, 21, 21, 21, 21, 25, 25, 25, 25, 25, 25, 35, 35, 50, 70
 ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

Step 2: Select the value that occurs the most.

The mode is 21 ppm.



Ex.) 13 – Finding the Mode

Find the mode for weight (in lbs) of 6 randomly selected students.

120, 130, 135, 140, 132, 141

Step 1: Sort the data

120, 130, 132, 135, 140, 141

Step 2: Select the value that occurs the most.

No value occurs more than once.

There is no mode.



Ex.) 14 – Finding the Mode

Find the mode for weight (in lbs) of 8 randomly selected students.

120, 130, 135, 130, 140, 132, 135, 141

Step 1: Sort the data

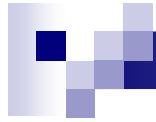
120, 130, 130, 132, 135, 135, 140, 141

Step 2: Select the value that occurs the most.

130 and 135 both occur the most.

The data set is said to be bimodal.

The modes are 130 and 135.



2-3 Mode

- Find the mode for the following data:

☐ 3 5 3 8 6

Ex.) 15 – Finding the Modal Class

Find the modal class for the frequency distribution of the hourly compensation (in U.S. \$) of production workers.

Class	Frequency
2.48 - 7.48	6
7.49 - 12.49	3
12.50 - 17.50	1
17.51 - 22.51	7
22.52 - 27.52	5
27.53 - 32.53	5
2.48 - 7.48	4

The modal class is
17.51 – 22.51.



Measures of Central Tendency: Midrange

- The **midrange** is the average of the lowest and highest values in a data set.

$$MR = \frac{Lowest + Highest}{2}$$



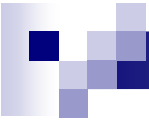
Ex.) 16 – Finding the Midrange

An athlete ran the following times in seconds for the 200 meter dash in a given session. Find the midrange.

26.1, 25.6, 25.7, 25.2, 25.0

$$\text{MR} = \frac{25.0 + 26.1}{2} = \frac{51.1}{2} = 25.55$$

The midrange is 25.55.




Properties of the Mean

- Uses all data values.
- Varies less than the median or mode
- Used in computing other statistics, such as the variance
- Unique, usually not one of the data values
- Cannot be used with open-ended classes
- Affected by extremely high or low values, called outliers



Properties of the Median

- Gives the midpoint
- Used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- Can be used for an open-ended distribution.
- Affected less than the mean by extremely high or extremely low values.



Properties of the Mode

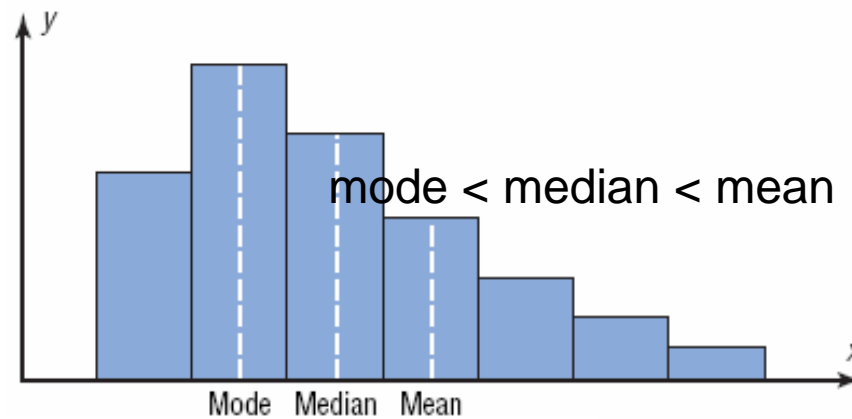
- Used when the most typical case is desired
- Easiest average to compute
- Can be used with nominal data
- Not always unique or may not exist



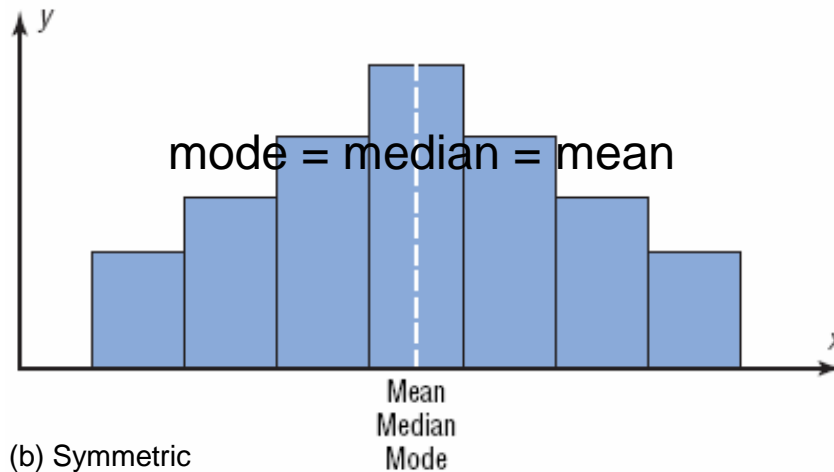
Properties of the Midrange

- Easy to compute.
- Gives the midpoint.
- Affected by extremely high or low values in a data set

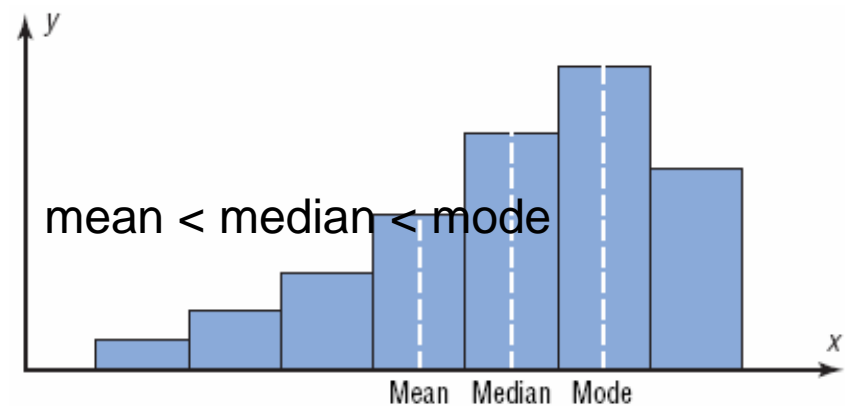
Distributions



(a) Positively skewed or right-skewed



(b) Symmetric



(c) Negatively skewed or left-skewed



Class Activity – Measures of Central Tendency

Find the mean, median, and mode for the following data sets and determine what kind of distribution they follow:

Set 1: 2.42, 3.90, 3.65, 3.30, 2.42, 0.98, 0.28

Set 2: 69.99, 69.99, 52.64, 15.62, 7.52, 10.91, 96.26



Class Activity – Measures of Central Tendency

- What would be the best measure of central tendency to use for the following?
 - ☐ eggs, eggs, eggs, eggs, cheese, chicken, fish
 - ☐ 2.43, 2.42, 2.46, 2.50, 2.38, 2.62, 10.11
 - ☐ 2, 4, 9, 11, 1, 5, 7, 3, 1, 6, 8



2-4 Measures of Dispersion

How Can We Measure Variability?

- ☐ Range
- ☐ Variance
- ☐ Standard Deviation



Measures of Dispersion: Range

- The **range** is the difference between the highest and lowest values in a data set.

$$R = \text{Highest} - \text{Lowest}$$



Ex.) 17 – Finding the Range

Two different brands of light bulbs are tested to see how long each will last before burning out. Six light bulbs of each brand constitute are sampled. The results (in hours) are shown. Find the mean and range of each group.

Brand A	Brand B
1000	1200
1200	900
1150	1350
1075	775
1000	1250
1050	1000

Ex.) 17 – Finding the Range

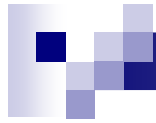
Brand A	Brand B
1000	1200
1200	900
1150	1350
1075	775
1000	1250
1050	1000

$$\text{Brand A: } \bar{x} = \frac{\sum X}{n} = \frac{6475}{6} = 1079.2$$
$$R = 1200 - 1000 = 200$$

$$\text{Brand B: } \bar{x} = \frac{\sum X}{n} = \frac{6475}{6} = 1079.2$$
$$R = 1350 - 775 = 575$$

The average for both brands is the same, but the range for Brand B is much greater than the range for Brand A.

Which brand would you buy?



2-4 Range

- Find the range of the following data:

□ 3 5 3 8 6



Measures of Dispersion: Variance & Standard Deviation

- The **variance** is the average of the squares of the distance each value is from the mean.
- The **standard deviation** is the square root of the variance.
- The standard deviation is a measure of how spread out your data are.



Uses of the Variance and Standard Deviation

- To determine the spread of the data.
- To determine the consistency of a variable.
- To determine the number of data values that fall within a specified interval in a distribution (Chebyshev's Theorem).
- Used in inferential statistics.



Measures of Dispersion: Variance & Standard Deviation

- The **sample variance** is

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

- The **sample standard deviation** is

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

- The numerator for the sample variance, $\sum (x - \bar{x})^2$, is often called the *sum of squares for x* and symbolized by $SS(x)$.



Measures of Dispersion: Variance & Standard Deviation

- If we rewrite $SS(x)$ as:

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

- Then we can write a mathematically equivalent formula that:
 - Saves time when calculating by hand
 - Does not use the mean
 - Is more accurate when the mean has been rounded.



Measures of Dispersion: Variance & Standard Deviation

- The **sample variance** is

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

- The **sample standard deviation** is

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



Ex.) 18 – Finding Sample Variance & Standard Deviation

The number of incidents in which police were needed for a sample of 10 schools is:

7, 37, 3, 8, 48, 11, 6, 0, 10, 3

Find the variance and standard deviation for the data.

Solution:

Use the second formula for finding variance.

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

Ex.) 18 – Finding Sample Variance & Standard Deviation (continued)

Set up a table using the data and use the formula:

X	X^2
7	49
37	1369
3	9
8	64
48	2304
11	121
6	36
0	0
10	200
3	9
$(\sum x) = 133$	$\sum x^2 = 4061$

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

$$\begin{aligned}
 s^2 &= \frac{4061 - \frac{(133)^2}{10}}{9} \\
 &= \frac{4061 - 1768.9}{9} \\
 &= \frac{2292.1}{9}
 \end{aligned}$$

$$\begin{aligned}
 s^2 &= 254.68 \\
 s &= 15.96
 \end{aligned}$$



2-4 Variance & Standard Deviation

- Find the variance and standard deviation for the following data set:

□ 3 5 3 8 6



Class Activity

- A random sample of 10 of the 2007 NASCAR drivers produced the following ages:

□	36	26	48	28	45	21	21	38
	27	32						
- Find the variance and standard deviation.



2-5 Measures of Position

- Z-score
- Percentile
- Quartile
- Outlier



Measures of Position: Z-score

- A **z-score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation.

$$z = \frac{X - \bar{X}}{s}$$

- A z-score represents the number of standard deviations a value is above or below the mean.



Ex.) 19 – Calculating z-score

If the average number of vacation days for a selection of various countries has a mean of 29.4 days and a standard deviation of 8.6, find the z-score for Canada which has 26 vacation days and for Italy which has 42 vacation days.

$$z = \frac{X - \bar{X}}{s} = \frac{26 - 29.4}{8.6} = -0.40 \quad \text{Canada}$$

$$z = \frac{X - \bar{X}}{s} = \frac{42 - 29.4}{8.6} = 1.47 \quad \text{Italy}$$



2-5 Z-score

Indicate which student has the higher z score.

Art Major	$X = 46$	$\bar{X} = 50.5$	$s = 5.2$
Theater Major	$X = 70$	$\bar{X} = 75.1$	$s = 7.3$



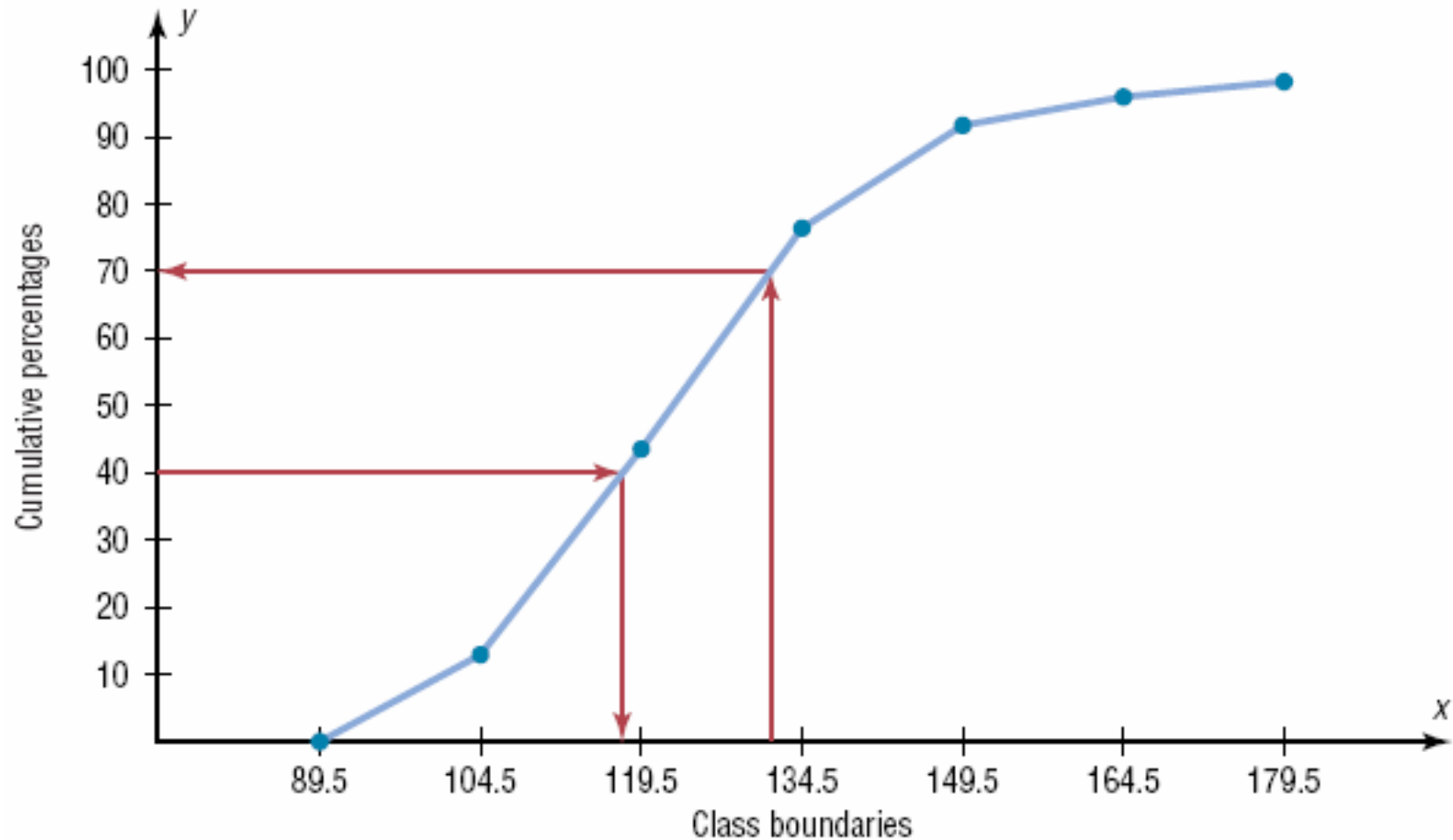
Measures of Position: Percentiles

- **Percentiles** separate the data set into 100 equal groups.
- A percentile rank for a datum represents the percentage of data values below the datum.

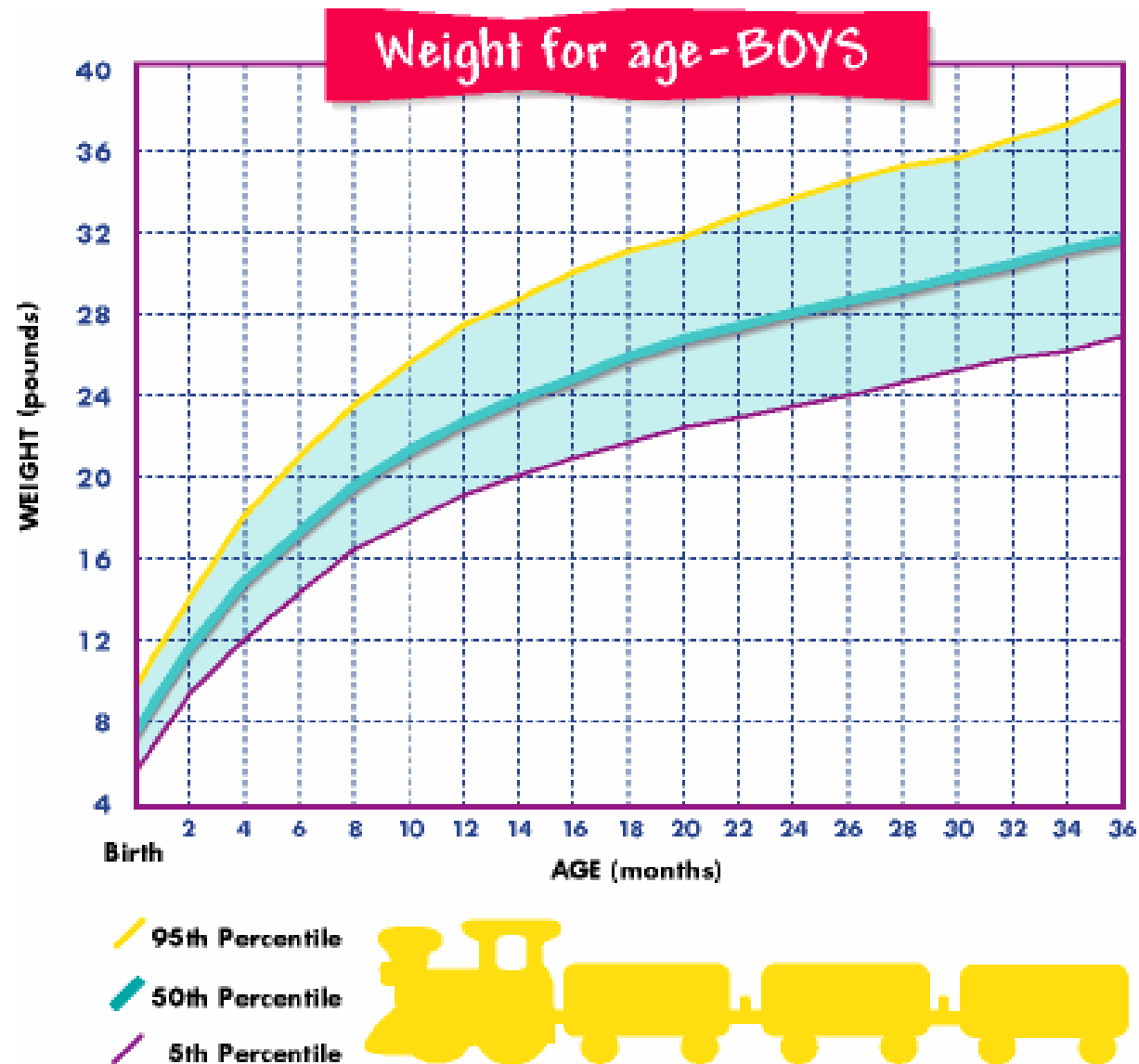
$$\textit{Percentile} = \frac{(\# \text{ of values below } X) + 0.5}{\text{total \# of values}} \cdot 100\%$$

$$c = \frac{nk}{100}$$

Measures of Position: Example of a Percentile Graph



Example - Percentiles



Ex.) 20 - Percentiles

The average weekly earnings in dollars for various industries are listed below. Find the percentile rank of \$524 in weekly earnings.

804, 736, 659, 489, 777, 623, 597, 524, 228

Sort in ascending order & then use the formula.

228, 489, 524, 597, 623, 659, 736, 777, 804

2 values ↑

$$\text{Percentile} = \frac{(\# \text{ of values below } X) + 0.5}{\text{total \# of values}} \cdot 100\%$$

$$\begin{aligned} &= \frac{2 + 0.5}{9} \cdot 100\% \\ &= 27.8\% \end{aligned}$$

Someone earning \$524 a week makes a salary that is 27.8% better than those in other industries.



2-5 Percentiles

- Below are the ACT scores attained by the 25 members of a local high school graduating class.

□ 23 26 25 19 33 21 21 22
21 27 19 25 18 23 22 30
27 27 23 16 21 19 20 30
22

- What is the percentile rank of a score of 26?



Ex.) 21 - Percentiles

The average weekly earnings in dollars for various industries are listed below. Find the value corresponding to the 75th percentile.

804, 736, 659, 489, 777, 623, 597, 524, 228

Use the formula:

$$c = \frac{nk}{100} = \frac{9 \cdot 75}{100} = 6.75 \approx 7$$

Arrange the data from lowest to highest & select entry number 7:

228, 489, 524, 597, 623, 659, 736, 777, 804



The value 736 corresponds to the 75th percentile.



2-5 Percentiles

- Below are the ACT scores attained by the 25 members of a local high school graduating class.


□ 23 26 25 19 33 21 21 22
21 27 19 25 18 23 22 30
27 27 23 16 21 19 20 30
22

- What ACT score represents the 80th percentile?



Measures of Position: Quartiles and Deciles

- **Deciles** separate the data set into 10 equal groups. $D_1=P_{10}$, $D_4=P_{40}$
- **Quartiles** separate the data set into 4 equal groups. $Q_1=P_{25}$, $Q_2=MD$, $Q_3=P_{75}$
- Q_2 = median of all the data
 Q_1 = median of the data up to Q_2
 Q_3 = median of the data from Q_2 to the highest value
- The **Interquartile Range**, $IQR = Q_3 - Q_1$.



Ex.) 22 – Quartiles

Find Q_1 , Q_2 , and Q_3 for the data set.

5, 12, 16, 25, 32, 38

Step 1: Sort in ascending order (already done).

5, 12, 16, 25, 32, 38

Step 2: Find the median. This is Q_2 .

$$Q_2 = \text{MD} = \frac{16 + 25}{2} = \boxed{20.5}$$



Ex.) 22 – Quartiles (continued)

Find Q_1 , Q_2 , and Q_3 for the data set.

5, 12, 16, 25, 32, 38

Step 3: Find the median of the values less than Q_2 .
This is Q_1 .

5, 12, 16

$$Q_1 = \text{MD} = \boxed{12}$$

Step 4: Find the median of the values greater than Q_2 . This is Q_3 .

25, 32, 38

$$Q_3 = \text{MD} = \boxed{32}$$



2-5 Quartiles

- Consider the following sample:
 - 26, 49, 9, 42, 60, 11, 43, 26, 30, and 14.
- Find Q_1 , Q_2 , and Q_3 for the data set.



Measures of Position:

Outliers

- An **outlier** is an extremely high or low data value when compared with the rest of the data values.
- A data value less than $Q_1 - 1.5(IQR)$ or greater than $Q_1 + 1.5(IQR)$ can be considered an outlier.



2-5 Outliers

- Consider the following sample:

- 26, 49, 9, 42, 60, 11, 43, 26, 30, and 14.

- Is the value 60 an outlier?



Measures of Position: Five Number Summary & Boxplot

- The **Five-Number Summary** is composed of the following numbers:
Low, Q_1 , MD, Q_3 , High
- The Five-Number Summary can be graphically represented using a **Boxplot (Box & Whiskers Display)**.



Constructing Boxplots

Procedure Table

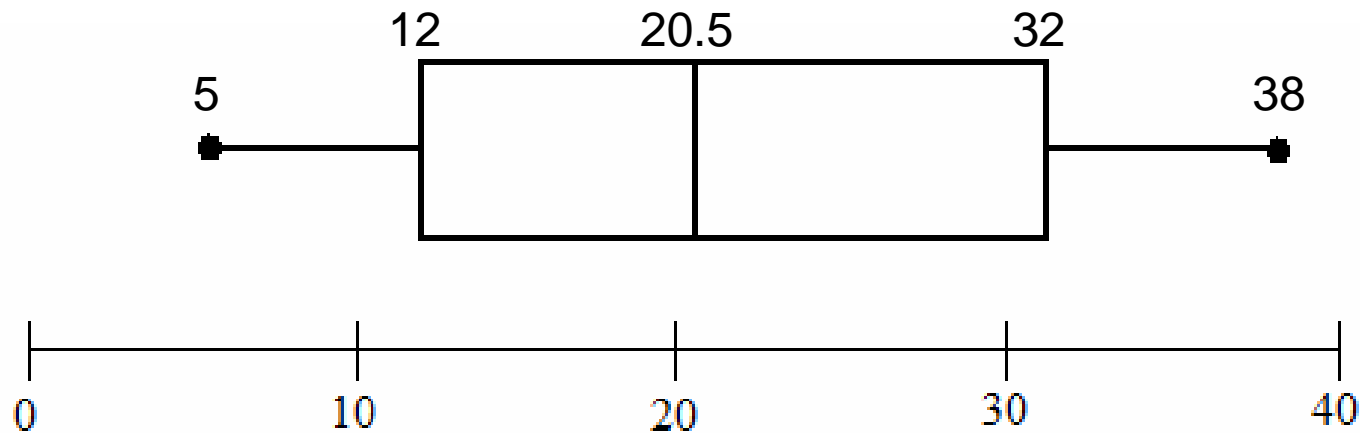
1. Find the five-number summary.
2. Draw a horizontal axis with a scale that includes the maximum and minimum data values.
3. Draw a box with vertical sides through Q_1 and Q_3 , and draw a vertical line through the median.
4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.


Ex.) 23 – Constructing a Boxplot

Using the data from example 22, construct a boxplot.

5, 12, 16, 25, 32, 38
↑ ↑ ↑ ↑ ↑
Low Q_1 MD Q_3 High

Five-Number Summary: 5-12-20.5-32-38






Class Activity – Exploratory Data Analysis

Remember the following data sets we looked at in a previous class activity:

Set 1: 2.42, 3.90, 3.65, 3.30, 2.42, 0.98, 0.28

Set 2: 69.99, 69.99, 52.64, 15.62, 7.52, 10.91, 96.26

For each data set, draw a boxplot and determine if it is consistent with your earlier conclusion about the distribution of the data.

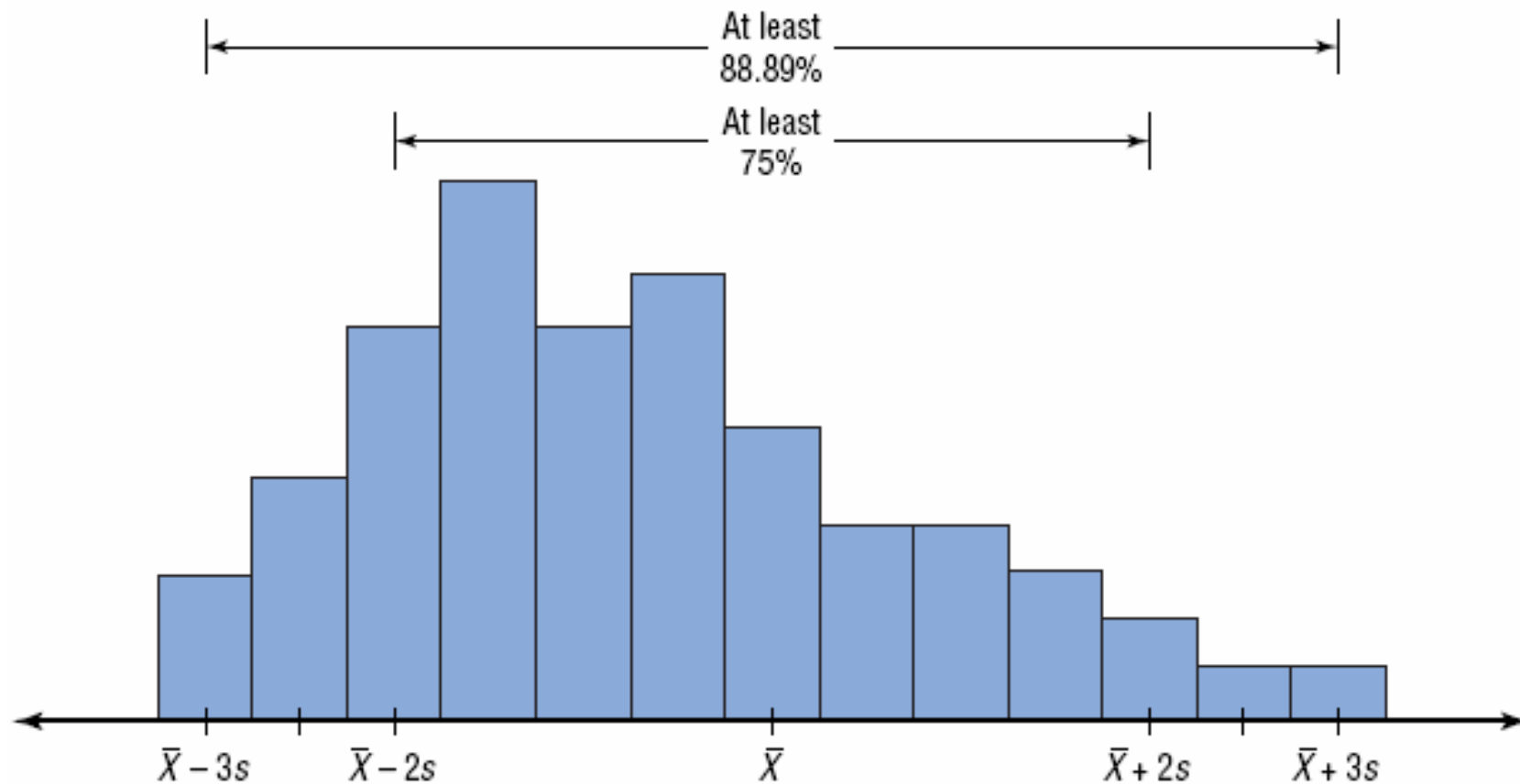


2-6 Interpreting and Understanding Standard Deviation: Chebyshev's Theorem

The proportion of values from any data set that fall within k standard deviations of the mean will be at least $1 - 1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

# of standard deviations, k	Minimum Proportion within k standard deviations	Minimum Percentage within k standard deviations
2	$1 - 1/4 = 3/4$	75%
3	$1 - 1/9 = 8/9$	88.89%
4	$1 - 1/16 = 15/16$	93.75%

Interpreting and Understanding Standard Deviation: Chebyshev's Theorem





Ex.) 24 – Chebyshev's Theorem

The average number of calories in a medium sized bagel is 240. If the standard deviation is 38 calories, find the range in which at least 75% of the data will lie.

Chebyshev's Theorem states that at least 75% of a data set will fall within 2 standard deviations of the mean.

$$\bar{x} - 2s = 240 - 2(38) = 164$$

$$\bar{x} + 2s = 240 + 2(38) = 316$$

At least 75% of all medium sized bagels will be between 164 and 316 calories.



Ex.) 25 – Chebyshev's Theorem

The average number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Using Chebyshev's Theorem, find the minimum percentage of data values that will fall in the range of 4 to 20 trials.


$$k = (X_U - \bar{X}) / s = (20 - 12) / 3 = 2.67$$

or

$$k = (\bar{X} - X_L) / s = (12 - 4) / 3 = 2.67$$

$$P = (1 - 1/k^2) \cdot 100\% = (1 - 1/2.67^2) \cdot 100\% = 85.9\%$$

At least 85.9% of the data values will fall between 4 and 20 trials.

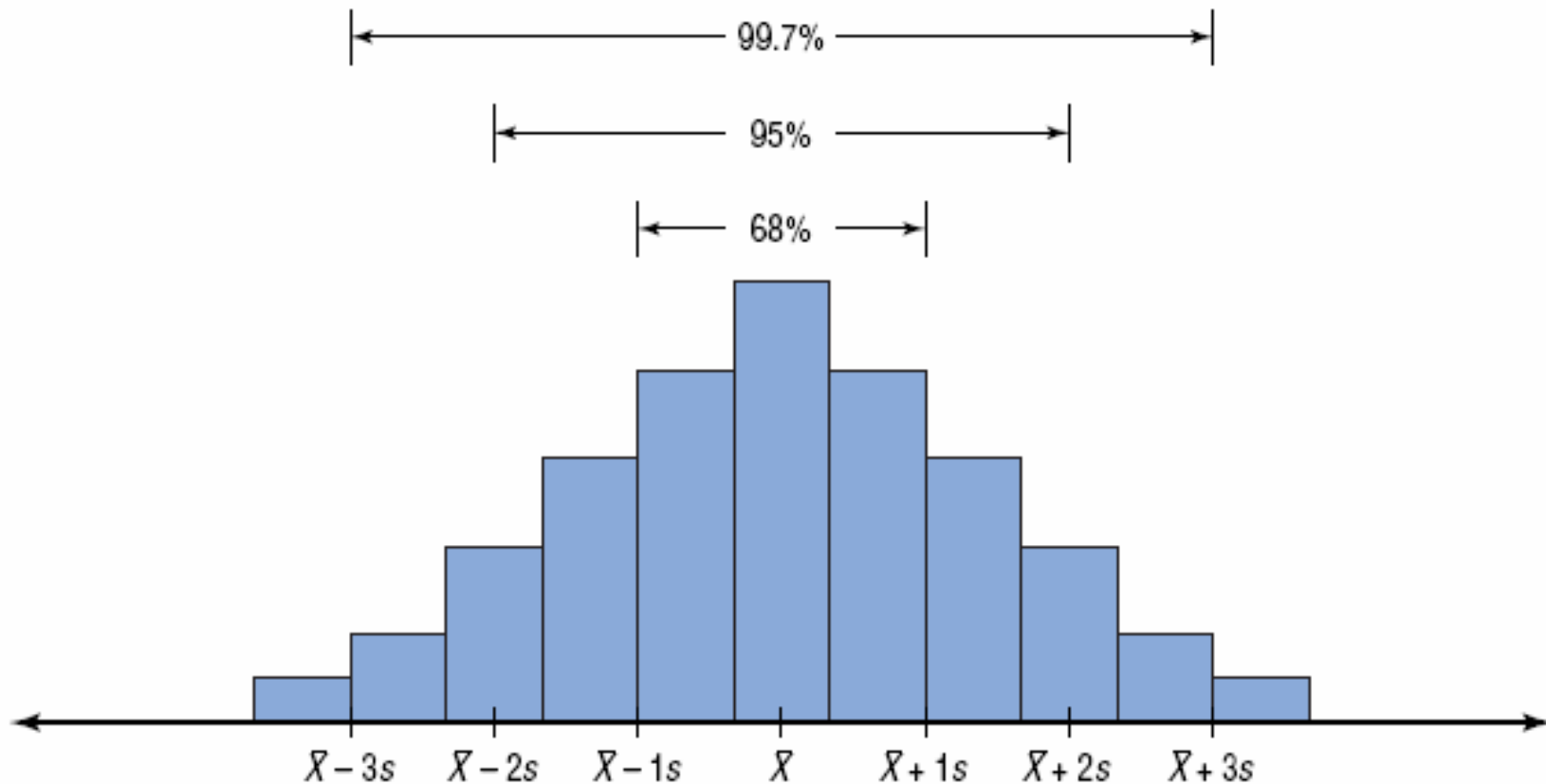


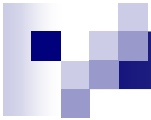
Interpreting and Understanding Standard Deviation: Empirical Rule

The percentage of values from a data set that fall within k standard deviations of the mean in a normal (bell-shaped) distribution is listed below.

# of standard deviations, k	Proportion within k standard deviations
1	68%
2	95%
3	99.7%

Interpreting and Understanding Standard Deviation: Empirical Rule (Normal)





Class Activity – Measures of Variation

Remember the 2 data sets we looked at in the last class activity:

Set 1: 2.42, 3.90, 3.65, 3.30, 2.42, 0.98, 0.28

Set 2: 69.99, 69.99, 52.64, 15.62, 7.52, 10.91, 96.26

For each data set, determine if it would be better to use Chebyshev's Theorem or the Empirical Rule to find percentile ranges. Then determine the range for which at least 95% of the data lie.



2-7 The Art of Statistical Deception

■ **Suspect Samples**

- ☐ Is the sample large enough?
- ☐ How was the sample selected?
- ☐ Is the sample representative of the population?

■ **Ambiguous Averages**

- ☐ What particular measure of average was used and why?



2-7 The Art of Statistical Deception

■ Changing the Subject

- Are different values used to represent the same data?

■ Detached Statistics

- One third fewer calories.....than what?

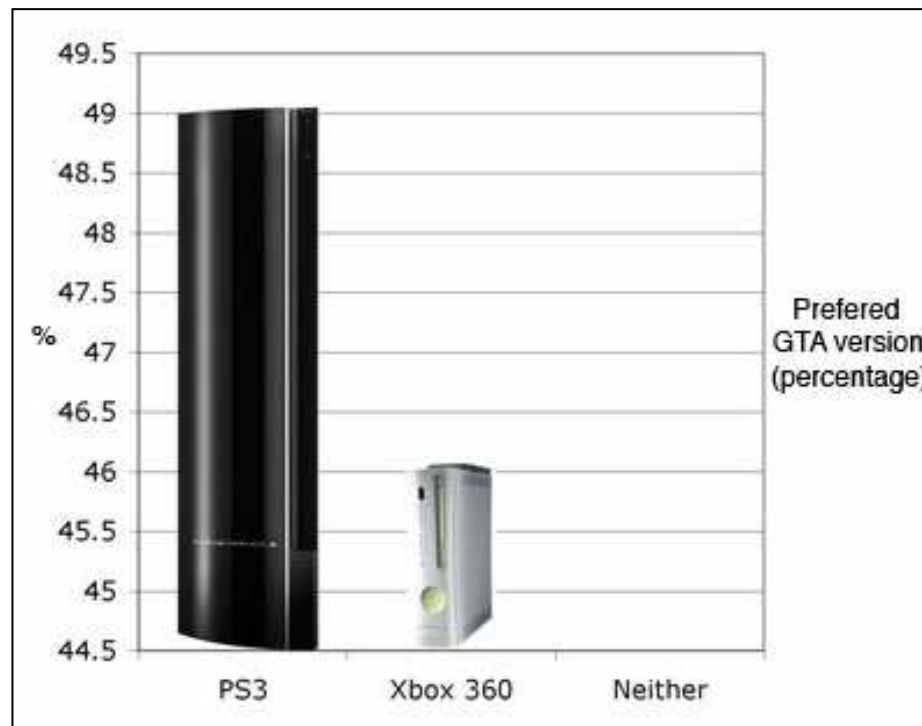
■ Implied Connections

- Studies *suggest* that *some people may* understand what this statement means.

2-7 The Art of Statistical Deception

■ Misleading Graphs

- Are the scales for the x-axis and y-axis appropriate for the data?





2-7 The Art of Statistical Deception

■ **Faulty Survey Questions**

☐ **Do you feel that statistics teachers should be paid higher salaries?**

vs.

☐ **Do you favor increasing tuition so that colleges can pay statistics teachers higher salaries?**



2-7 The Art of Statistical Deception

- **Who's doing the research for the study?**

- ☐ **Is the study showing that more people prefer PS3 to XBOX 360 funded by Playstation?**
- ☐ **Is the lead researcher known to favor one outcome over the other?**

Class Activity: Misleading Statistical Graphs

- How is the following graph misleading?

