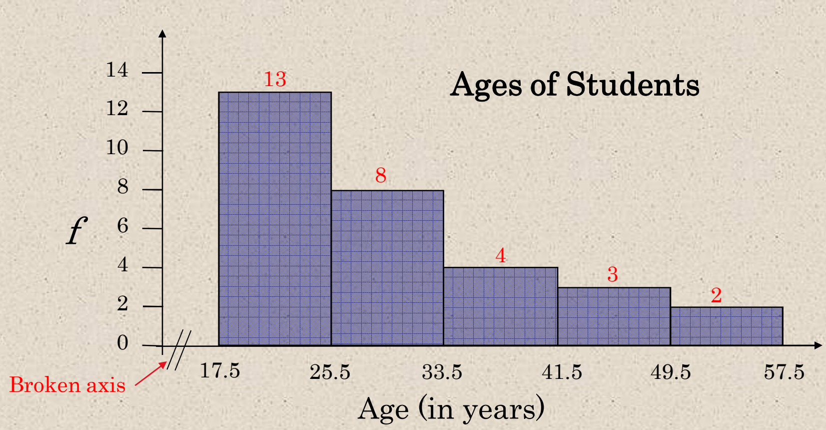
Chapter 2

Descriptive Statistics

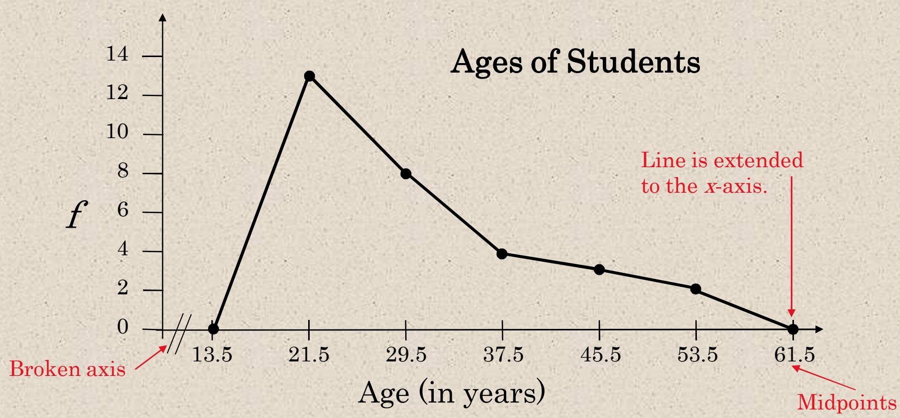
Frequency Histogram

Example:
Draw a frequency histogram for the "Ages of Students" frequency distribution. Use the class boundaries.



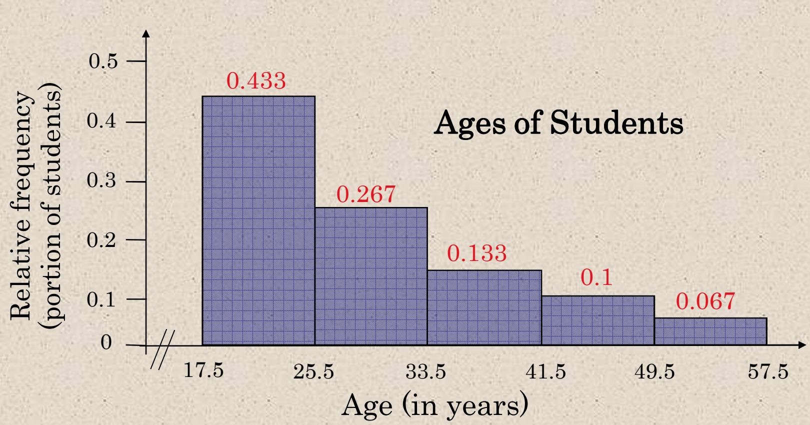
Frequency Polygon

A **frequency polygon** is a line graph that emphasizes the continuous change in frequencies.



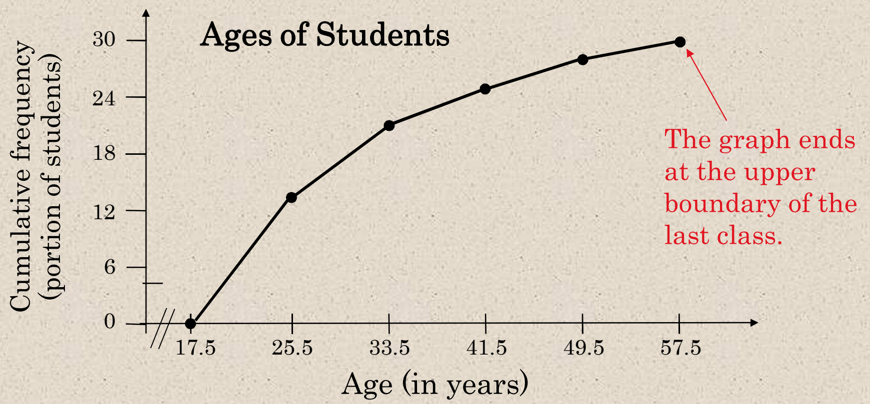
Relative Frequency Histogram

A **relative frequency histogram** has the same shape and the same horizontal scale as the corresponding frequency histogram.



Cumulative Frequency Graph

A **cumulative frequency graph** or **ogive**, is a line graph that displays the cumulative frequency of each class at its upper class boundary.



§ 2.2

More Graphs and Displays

Stem-and-Leaf Plot

In a **stem-and-leaf plot**, each number is separated into a stem (usually the entry's leftmost digits) and a leaf (usually the rightmost digit). This is an example of **exploratory data analysis**.

Example:
The following data represents the ages of 30 students in a statistics class. Display the data in a stem-and-leaf plot.

Ages of Students

18	20	21	27	29	20
19	30	32	19	34	19
24	29	18	37	38	22
30	39	32	44	33	46
54	49	18	51	21	21

Continued.

Stem-and-Leaf Plot

Ages of Students

Key: 1 | 8 = 18

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

Most of the values lie between 20 and 39.

This graph allows us to see the shape of the data as well as the actual values.

Stem-and-Leaf Plot

Example:
Construct a stem-and-leaf plot that has two lines for each stem.

Ages of Students

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

Key: 1 | 8 = 18

From this graph, we can conclude that more than 50% of the data lie between 20 and 34.

Pie Chart

A **pie chart** is a circle that is divided into sectors that represent categories. The area of each sector is proportional to the frequency of each category.

Accidental Deaths in the USA in 2002

Type	Frequency
Motor Vehicle	43,500
Falls	12,200
Poison	6,400
Drowning	4,600
Fire	4,200
Ingestion of Food/Object	2,900
Firearms	1,400

(Source: US Dept. of Transportation)

Continued.

Pie Chart

To create a pie chart for the data, find the relative frequency (percent) of each category.

Type	Frequency	Relative Frequency
Motor Vehicle	43,500	0.578
Falls	12,200	0.162
Poison	6,400	0.085
Drowning	4,600	0.061
Fire	4,200	0.056
Ingestion of Food/Object	2,900	0.039
Firearms	1,400	0.019

$n = 75{,}200$

Continued.

Pie Chart

Next, find the central angle. To find the central angle, multiply the relative frequency by 360°.

Type	Frequency	Relative Frequency	Angle
Motor Vehicle	43,500	0.578	208.2°
Falls	12,200	0.162	58.4°
Poison	6,400	0.085	30.6°
Drowning	4,600	0.061	22.0°
Fire	4,200	0.056	20.1°
Ingestion of Food/Object	2,900	0.039	13.9°
Firearms	1,400	0.019	6.7°

Continued.

Scatter Plot

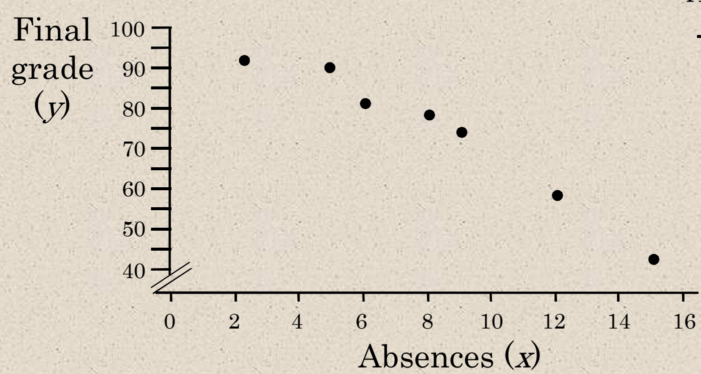
When each entry in one data set corresponds to an entry in another data set, the sets are called **paired data sets**.

In a **scatter plot**, the ordered pairs are graphed as points in a coordinate plane. The scatter plot is used to show the relationship between two quantitative variables.

The following scatter plot represents the relationship between the number of absences from a class during the semester and the final grade.

Continued.

Scatter Plot



Absences x	Grade y
8	78
2	92
5	90
12	58
15	43
9	74
6	81

From the scatter plot, you can see that as the number of absences increases, the final grade tends to decrease.

Times Series Chart

A data set that is composed of quantitative data entries taken at regular intervals over a period of time is a **time series**. A **time series chart** is used to graph a time series.

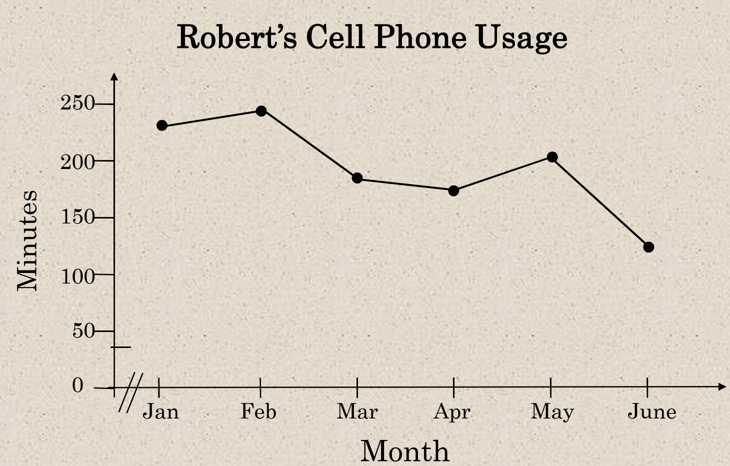
Example:
The following table lists the number of minutes Robert used on his cell phone for the last six months.

Construct a time series chart for the number of minutes used.

Month	Minutes
January	236
February	242
March	188
April	175
May	199
June	135

Continued.

Times Series Chart



§ 2.3

Measures of Central Tendency

Mean

A measure of central tendency is a value that represents a typical, or central, entry of a data set. The three most commonly used measures of central tendency are the mean, the median, and the mode.

The **mean** of a data set is the sum of the data entries divided by the number of entries.

Population mean: $\mu = \frac{\sum x}{N}$ Sample mean: $\bar{x} = \frac{\sum x}{n}$

"mu" "*x*-bar"

Mean

Example:
The following are the ages of all seven employees of a small company:

53 32 61 57 39 44 57

Calculate the population mean.

$$\mu = \frac{\sum x}{N} = \frac{343}{7}$$

Add the ages and divide by 7.

$$= 49 \text{ years}$$

The mean age of the employees is 49 years.

Median

The **median** of a data set is the value that lies in the middle of the data when the data set is ordered. If the data set has an odd number of entries, the median is the middle data entry. If the data set has an even number of entries, the median is the mean of the two middle data entries.

Example:
Calculate the median age of the seven employees.

53 32 61 57 39 44 57

To find the median, sort the data.

32 39 44 **53** 57 57 61

The median age of the employees is 53 years.

Mode

The **mode** of a data set is the data entry that occurs with the greatest frequency. If no entry is repeated, the data set has no mode. If two entries occur with the same greatest frequency, each entry is a mode and the data set is called **bimodal**.

Example:
Find the mode of the ages of the seven employees.

53 32 61 57 39 44 57

The mode is 57 because it occurs the most times.

An **outlier** is a data entry that is far removed from the other entries in the data set.

Shapes of Distributions

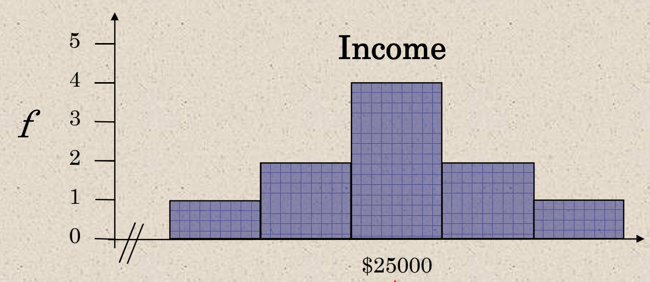
A frequency distribution is **symmetric** when a vertical line can be drawn through the middle of a graph of the distribution and the resulting halves are approximately the mirror images.

A frequency distribution is **uniform** (or **rectangular**) when all entries, or classes, in the distribution have equal frequencies. A uniform distribution is also symmetric.

A frequency distribution is skewed if the "tail" of the graph elongates more to one side than to the other. A distribution is **skewed left** (**negatively skewed**) if its tail extends to the left. A distribution is **skewed right** (**positively skewed**) if its tail extends to the right.

Symmetric Distribution

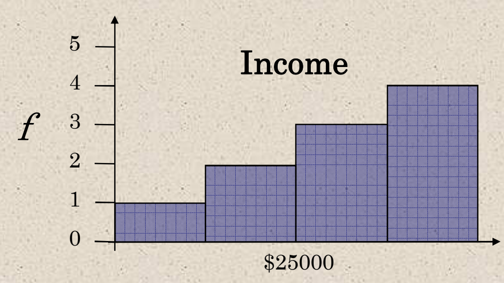
10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000



mean = median = mode
= $25,000

Skewed Left Distribution

10 Annual Incomes
0
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
35,000

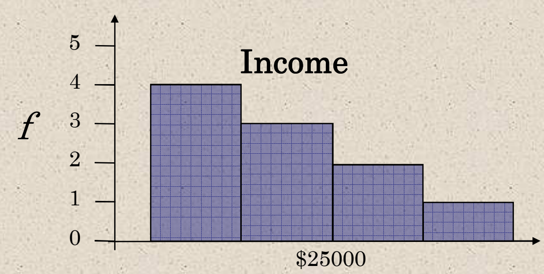


mean = $23,500
median = mode = $25,000

Mean < Median

Skewed Right Distribution

10 Annual Incomes
15,000
20,000
22,000
24,000
25,000
25,000
26,000
28,000
30,000
1,000,000

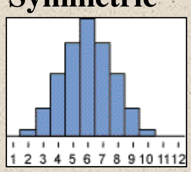


mean = $121,500
median = mode = $25,000

Mean > Median

Summary of Shapes of Distributions

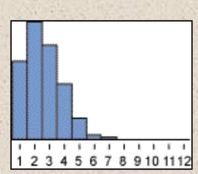
Symmetric



Uniform

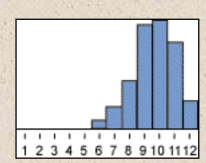
Mean = Median

Skewed right



Mean > Median

Skewed left



Mean < Median

§ 2.4

Measures of Variation

Range

The **range** of a data set is the difference between the maximum and minimum date entries in the set.
Range = (Maximum data entry) – (Minimum data entry)

Example:
The following data are the closing prices for a certain stock on ten successive Fridays. Find the range.

Stock	56	56	57	58	61	63	63	67	67	67

The range is 67 – 56 = 11.

Deviation

The **deviation** of an entry x in a population data set is the difference between the entry and the mean μ of the data set.

Deviation of $x = x - \mu$

Example:
The following data are the closing prices for a certain stock on five successive Fridays. Find the deviation of each price.

The mean stock price is
$\mu = 305/5 = 61$.

Stock x	Deviation $x - \mu$
56	$56 - 61 = -5$
58	$58 - 61 = -3$
61	$61 - 61 = 0$
63	$63 - 61 = 2$
67	$67 - 61 = 6$
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$

Variance and Standard Deviation

The **population variance** of a population data set of N entries is

$$\text{Population variance} = \sigma^2 = \frac{\Sigma(x-\mu)^2}{N}.$$

"sigma squared"

The **population standard deviation** of a population data set of N entries is the square root of the population variance.

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}.$$

"sigma"

Finding the Population Standard Deviation

Guidelines

In Words	*In Symbols*
1. Find the mean of the population data set.	$\mu = \frac{\Sigma x}{N}$
2. Find the deviation of each entry.	$x - \mu$
3. Square each deviation.	$(x - \mu)^2$
4. Add to get the **sum of squares**.	$SS_x = \Sigma(x - \mu)^2$
5. Divide by N to get the **population variance**.	$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$
6. Find the square root of the variance to get the **population standard deviation**.	$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$

Finding the Sample Standard Deviation

Guidelines

In Words	*In Symbols*
1. Find the mean of the sample data set.	$\bar{x} = \frac{\Sigma x}{n}$
2. Find the deviation of each entry.	$x - \bar{x}$
3. Square each deviation.	$(x - \bar{x})^2$
4. Add to get the **sum of squares**.	$SS_x = \Sigma(x - \bar{x})^2$
5. Divide by $n - 1$ to get the **sample variance**.	$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$
6. Find the square root of the variance to get the **sample standard deviation**.	$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$

Finding the Population Standard Deviation

Example:
The following data are the closing prices for a certain stock on five successive Fridays. The population mean is 61. Find the population standard deviation.

Always positive!

Stock x	Deviation $x - \mu$	Squared $(x - \mu)^2$
56	– 5	25
58	– 3	9
61	0	0
63	2	4
67	6	36
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$	$\Sigma(x - \mu)^2 = 74$

$SS_2 = \Sigma(x - \mu)^2 = 74$

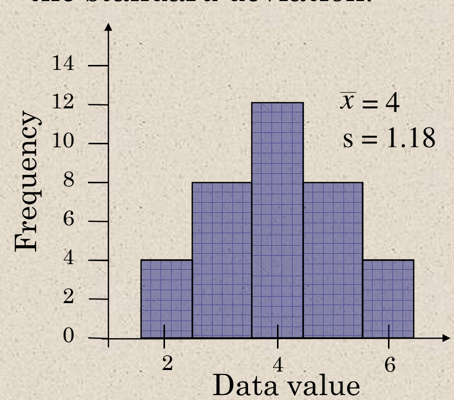
$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{74}{5} = 14.8$$

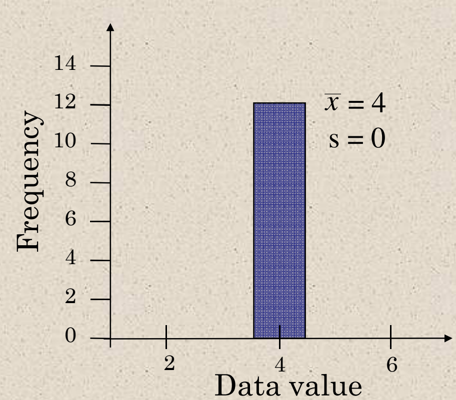
$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{14.8} \approx 3.85$$

$\sigma \approx \$3.85$

Interpreting Standard Deviation

When interpreting standard deviation, remember that is a measure of the typical amount an entry deviates from the mean. The more the entries are spread out, the greater the standard deviation.



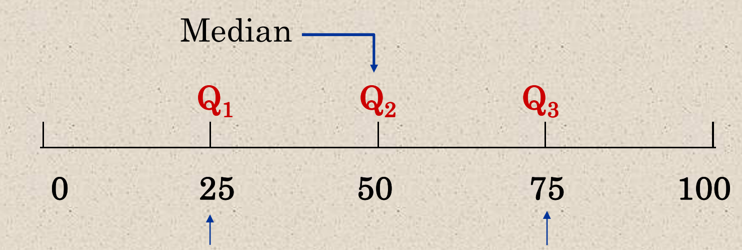


§ 2.5

Measures of Position

Quartiles

The three **quartiles**, Q_1, Q_2, and Q_3, approximately divide an ordered data set into four equal parts.



Q_1 is the median of the data below Q_2.

Q_3 is the median of the data above Q_2.

Finding Quartiles

Example:

The quiz scores for 15 students is listed below. Find the first, second and third quartiles of the scores.

28 43 48 51 43 30 55 44 48 33 45 37 37 42 38

Order the data.

Lower half: 28 30 33 37 37 38 42

43

Upper half: 43 44 45 48 48 51 55

Q_1 = 37, Q_2 = 43, Q_3 = 48

About one fourth of the students scores 37 or less; about one half score 43 or less; and about three fourths score 48 or less.

Interquartile Range

The **interquartile range (IQR)** of a data set is the difference between the third and first quartiles.

Interquartile range (IQR) = $Q_3 - Q_1$.

Example:

The quartiles for 15 quiz scores are listed below. Find the interquartile range.

$Q_1 = 37$ $Q_2 = 43$ $Q_3 = 48$

(IQR) = $Q_3 - Q_1$
= 48 – 37
= 11

The quiz scores in the middle portion of the data set vary by at most 11 points.

Box and Whisker Plot

A **box-and-whisker plot** is an exploratory data analysis tool that highlights the important features of a data set.

The **five-number summary** is used to draw the graph.

- The minimum entry
- Q_1
- Q_2 (median)
- Q_3
- The maximum entry

Example:

Use the data from the 15 quiz scores to draw a box-and-whisker plot.

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

Continued.

Box and Whisker Plot

Five-number summary

- The minimum entry 28
- Q_1 37
- Q_2 (median) 43
- Q_3 48
- The maximum entry 55