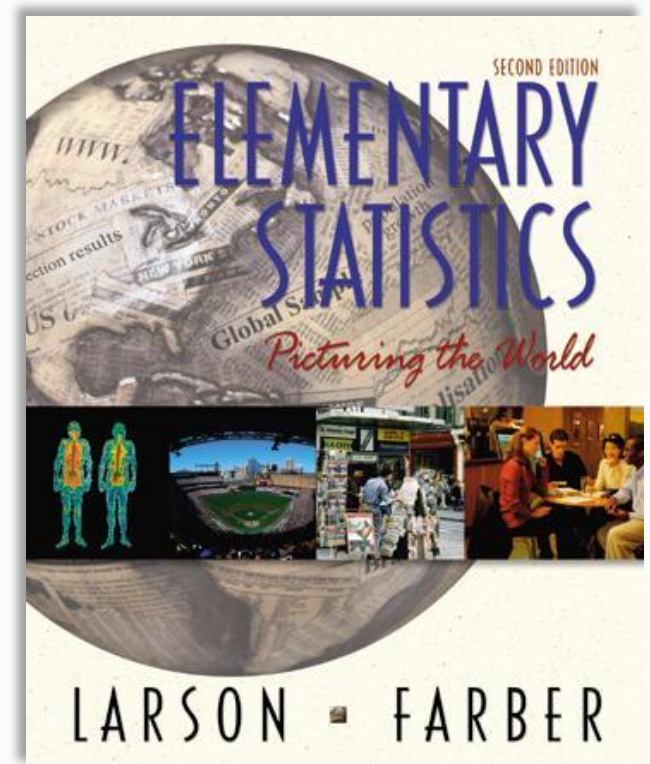


Elementary Statistics

Larson Farber

Section 2.1

Frequency Distributions and Their Graphs



Frequency Distributions

Minutes Spent on the Phone

102	124	108	86	103	82
71	104	112	118	87	95
103	116	85	122	87	100
105	97	107	67	78	125
109	99	105	99	101	92

Make a frequency distribution table with five classes.

Frequency Distributions

Classes - the intervals used in the distribution

Class width - the range divided by the number of classes,
round up to next number

greatest # - smallest # ALWAYS ROUND UP
of classes

Lower class limit - the smallest # that can be in the class

Upper class limit - the greatest # that can be in the class

Frequency - the number of items in the class

Frequency Distributions

Midpoint - the sum of the limits divided by 2

$$\frac{\text{lower class limit} + \text{upper class limit}}{2}$$

Relative frequency - the portion (%) of data in that class

$$\frac{\text{class frequency (f)}}{\text{sample size (n)}}$$

Cumulative frequency – the sum of the frequencies for that class and all previous classes

Construct a Frequency Distribution

Minimum = 67, Maximum = 125

Number of classes = 5

Class width = 12

	Class	Limits	Tally	\hat{f}
	67	78		
	79	90	— + + + + —	5
	91	102	— + + + + —	8
	103	114	— + + + + —	9
	115	126	— + + + + —	5
				<hr/> $\Sigma \hat{f} = 30$

Do all lower class limits first.

Other Information

Class	f	Midpoint	Relative Frequency	Cumulative Frequency
67 - 78	3	72.5	0.10	3
79 - 90	5	84.5	0.17	8
91 - 102	8	96.5	0.27	16
103 - 114	9	108.5	0.30	25
115 - 126	5	120.5	0.17	30

Frequency Histogram

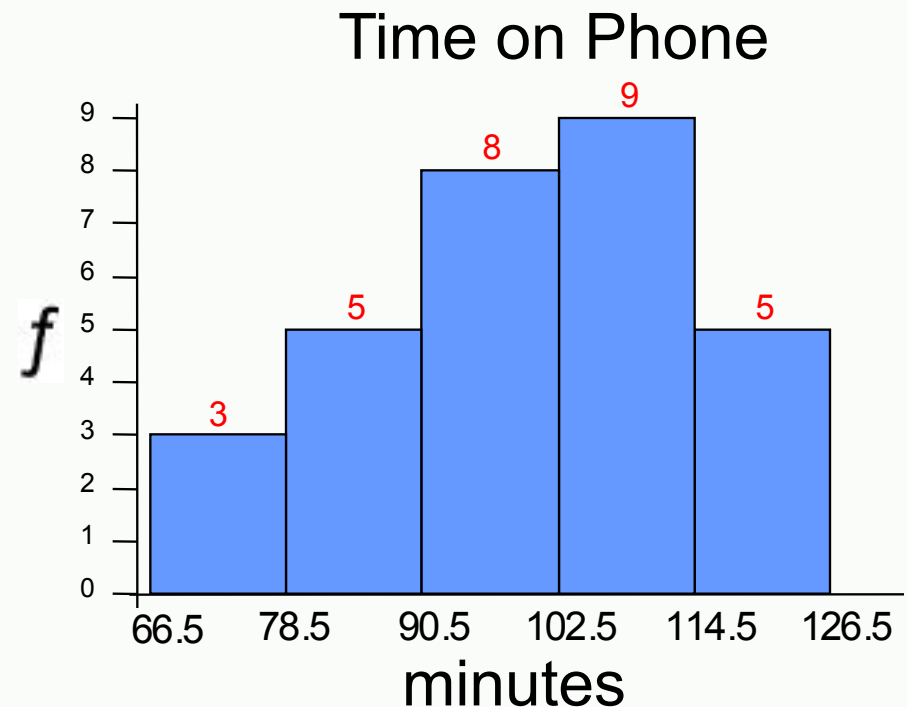
A bar graph that represents the frequency distribution of the data set

- 1. horizontal scale uses class boundaries or midpoints**
- 2. vertical scale measures frequencies**
- 3. consecutive bars must touch**

Class boundaries - numbers that separate classes without forming gaps between them

Frequency Histogram

Class	f	Boundaries
67 - 78		66.5 - 78.5
79 - 90	5	78.5 - 90.5
91 - 102	8	90.5 - 102.5
103 - 114	9	102.5 - 114.5
115 - 126	5	114.5 - 126.5



Relative Frequency Histogram

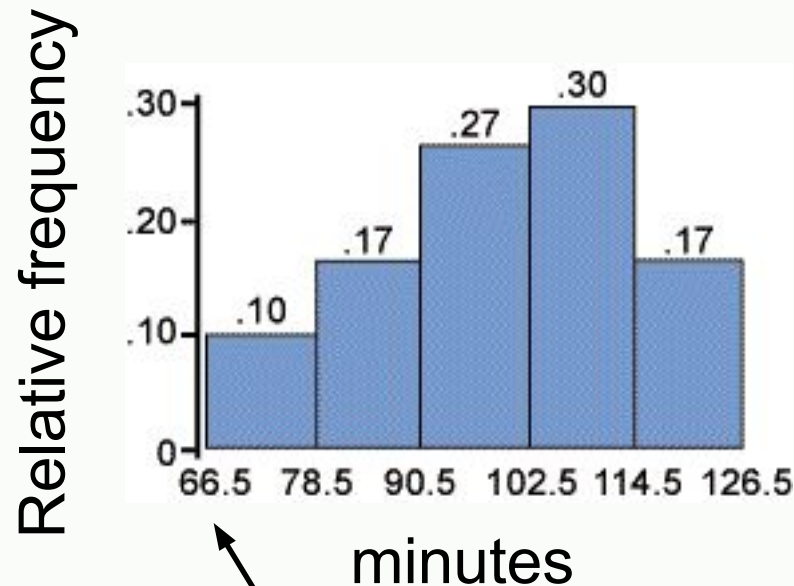
A bar graph that represents the relative frequency distribution of the data set

Same shape as frequency histogram

- 1. horizontal scale uses class boundaries or midpoints**
- 2. vertical scale measures relative frequencies**

Relative Frequency Histogram

Time on Phone



Relative frequency on vertical scale

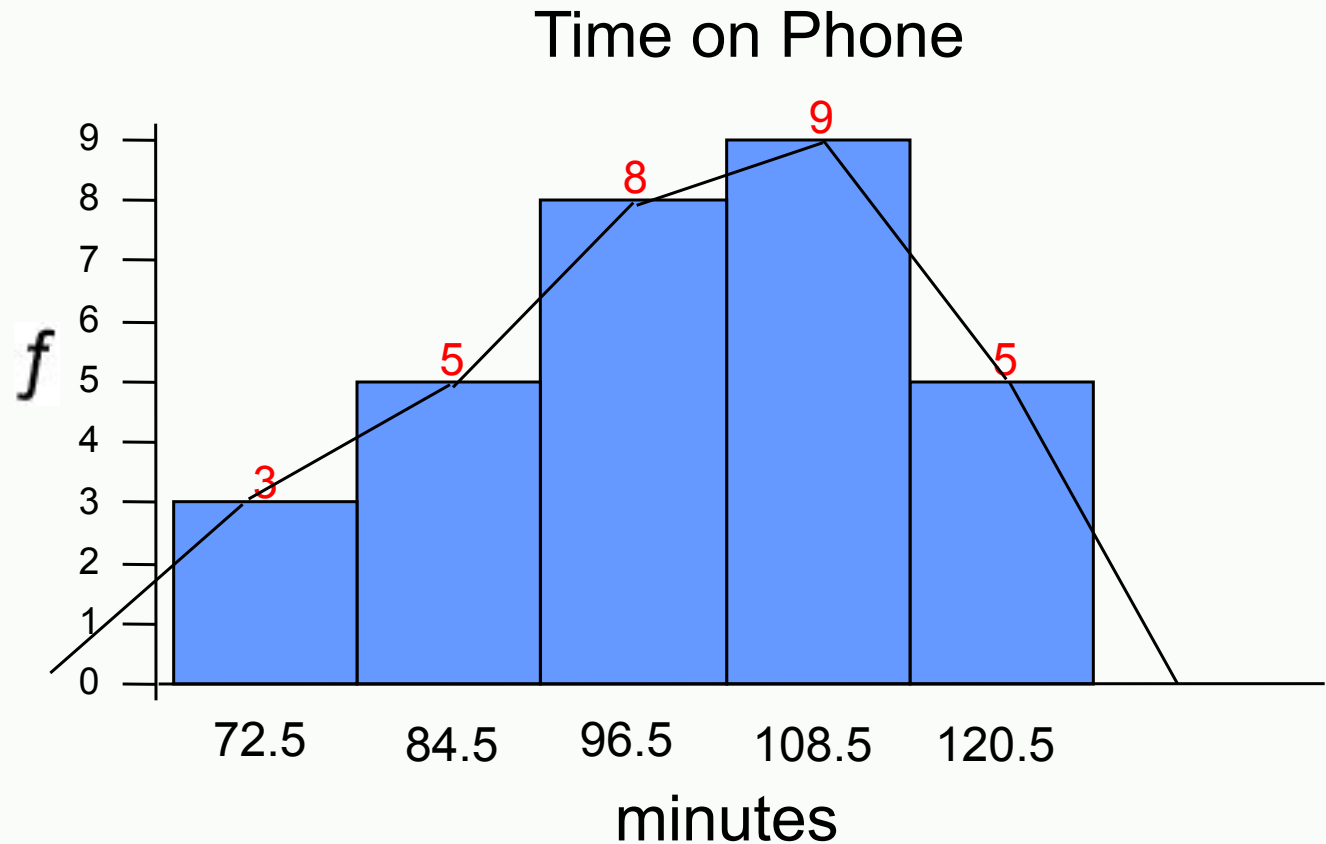
Frequency Polygon

A line graph that emphasizes the continuous change in frequencies

- 1. horizontal scale uses class midpoints**
- 2. vertical scale measures frequencies**

Frequency Polygon

Class	f
67 - 78	
79 - 90	5
91 - 102	8
103 - 114	9
115 - 126	5



Mark the midpoint at the top of each bar. Connect consecutive midpoints. Extend the frequency polygon to the axis.

Ogive

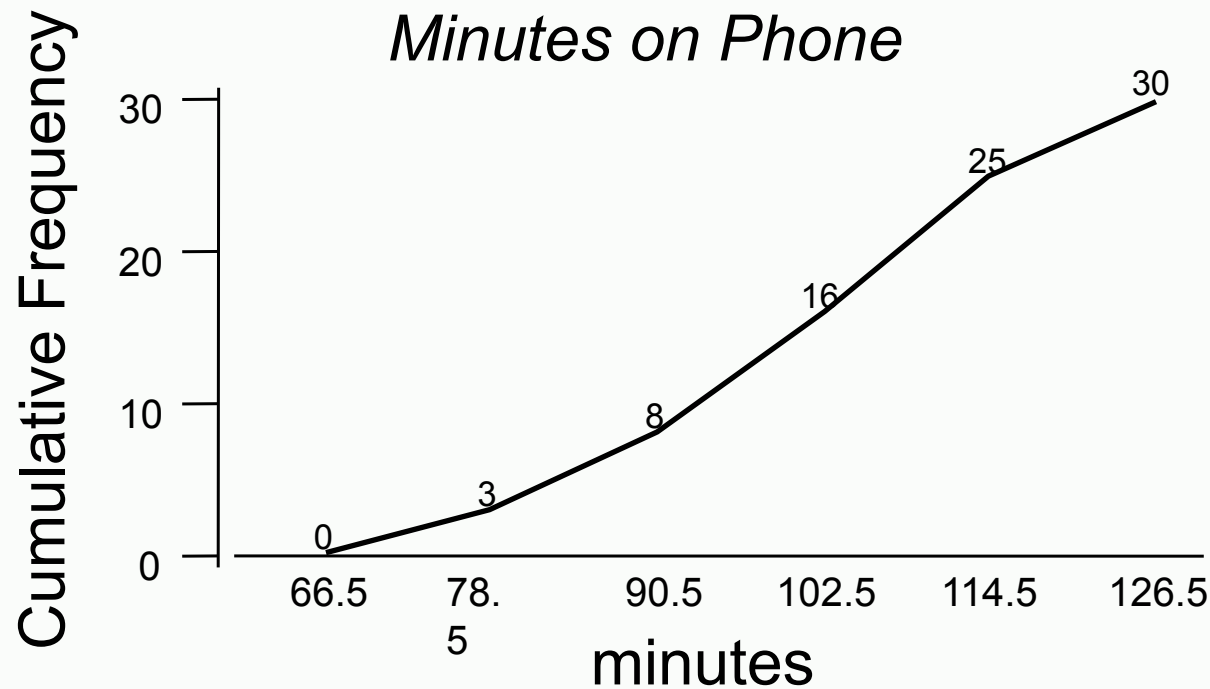
Also called a cumulative frequency graph

A line graph that displays the cumulative frequency of each class

- 1. horizontal scale uses upper boundaries**
- 2. vertical scale measures cumulative frequencies**

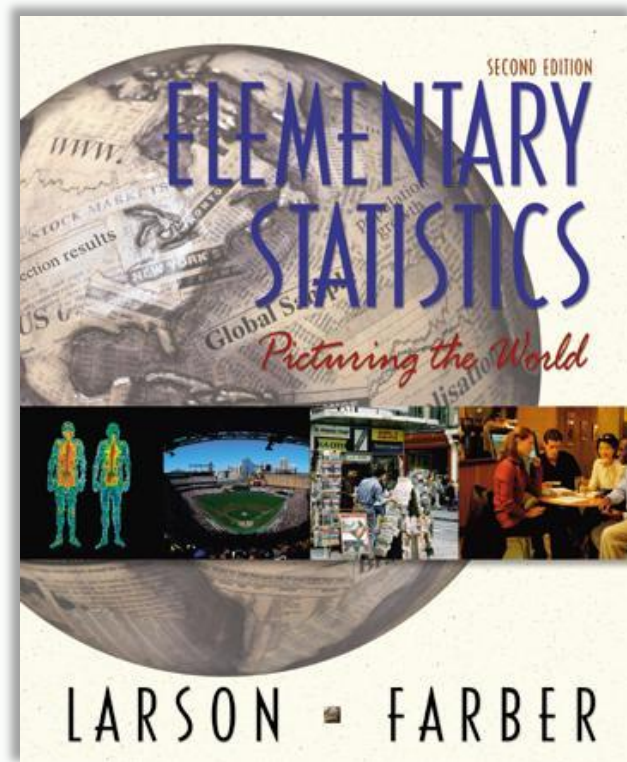
Ogive

An ogive reports the number of values in the data set that are less than or equal to the given value, x .



Section 2.2

More Graphs and Displays



Stem-and-Leaf Plot

- contains all original data**
- easy way to sort data & identify outliers**

Minutes Spent on the Phone

102	124	108	86	103	82
71	104	112	118	87	95
103	116	85	122	87	100
105	97	107	67	78	125
109	99	105	99	101	92

Key values:

Minimum value = 67

Maximum value = 125

Stem-and-Leaf Plot

Lowest value is 67 and highest value is 125, so list stems from 6 to 12.

Never skip stems. You can have a stem with NO leaves.

<u>Stem</u>	<u>Leaf</u>	<u>Stem</u>	<u>Leaf</u>
6		12	
7		11	
8		10	
9		9	
10		8	
11		7	
12		6	

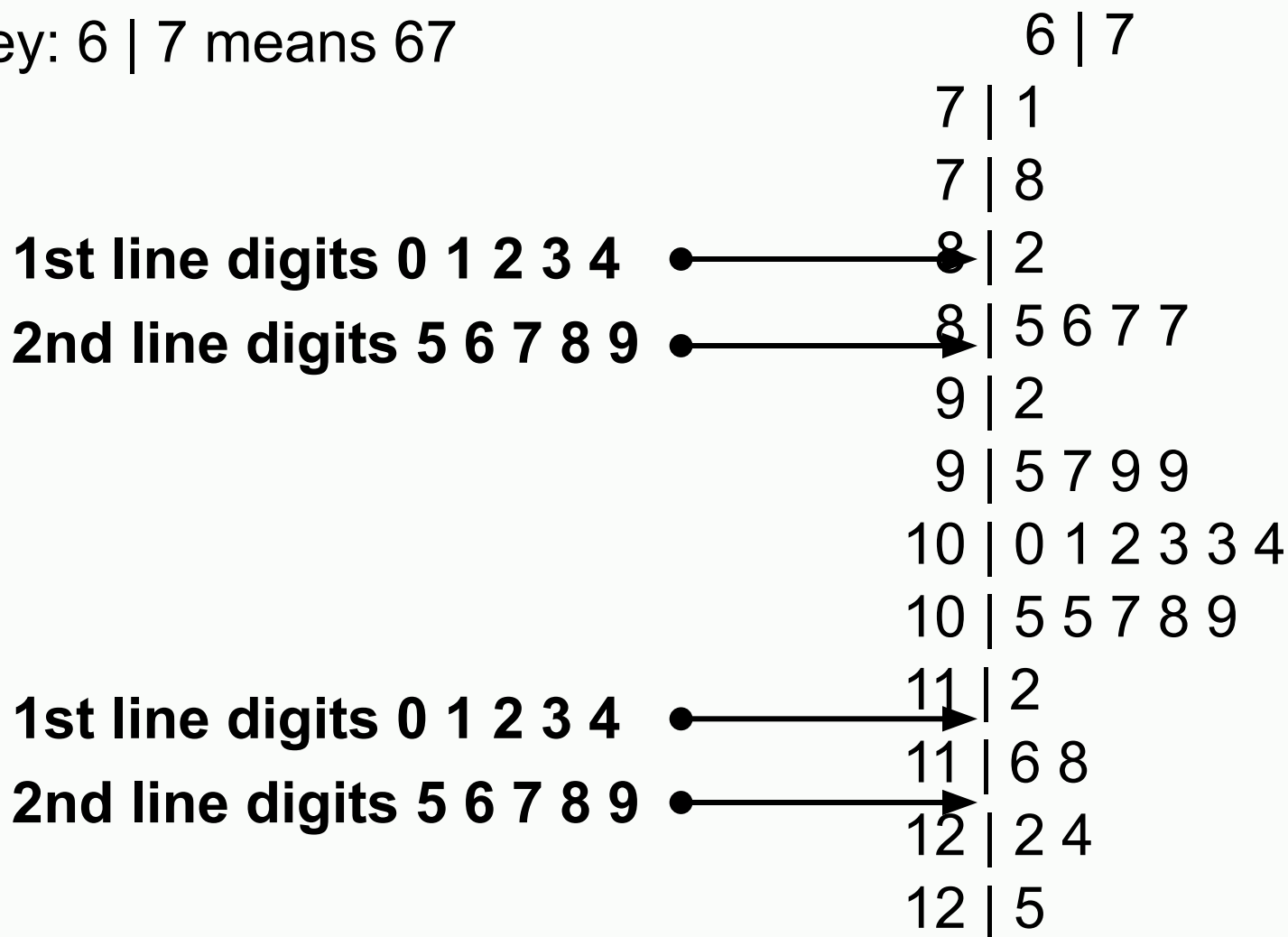
Stem-and-Leaf Plot

6 | 7
7 | 1 8
8 | 2 5 6 7 7
9 | 2 5 7 9 9
10 | 0 1 2 3 3 4 5 5 7 8 9
11 | 2 6 8
12 | 2 4 5

Key: 6 | 7 means 67

Stem-and-Leaf with two lines per stem

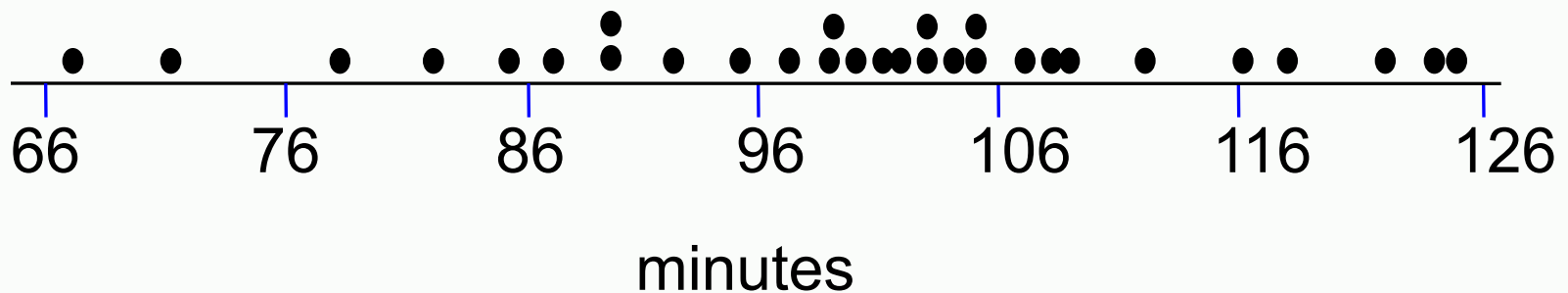
Key: 6 | 7 means 67



Dot Plot

- contains all original data
- easy way to sort data & identify outliers

Minutes Spent on the Phone



Pie Chart / Circle Graph

- Used to describe parts of a whole
- Central Angle for each segment

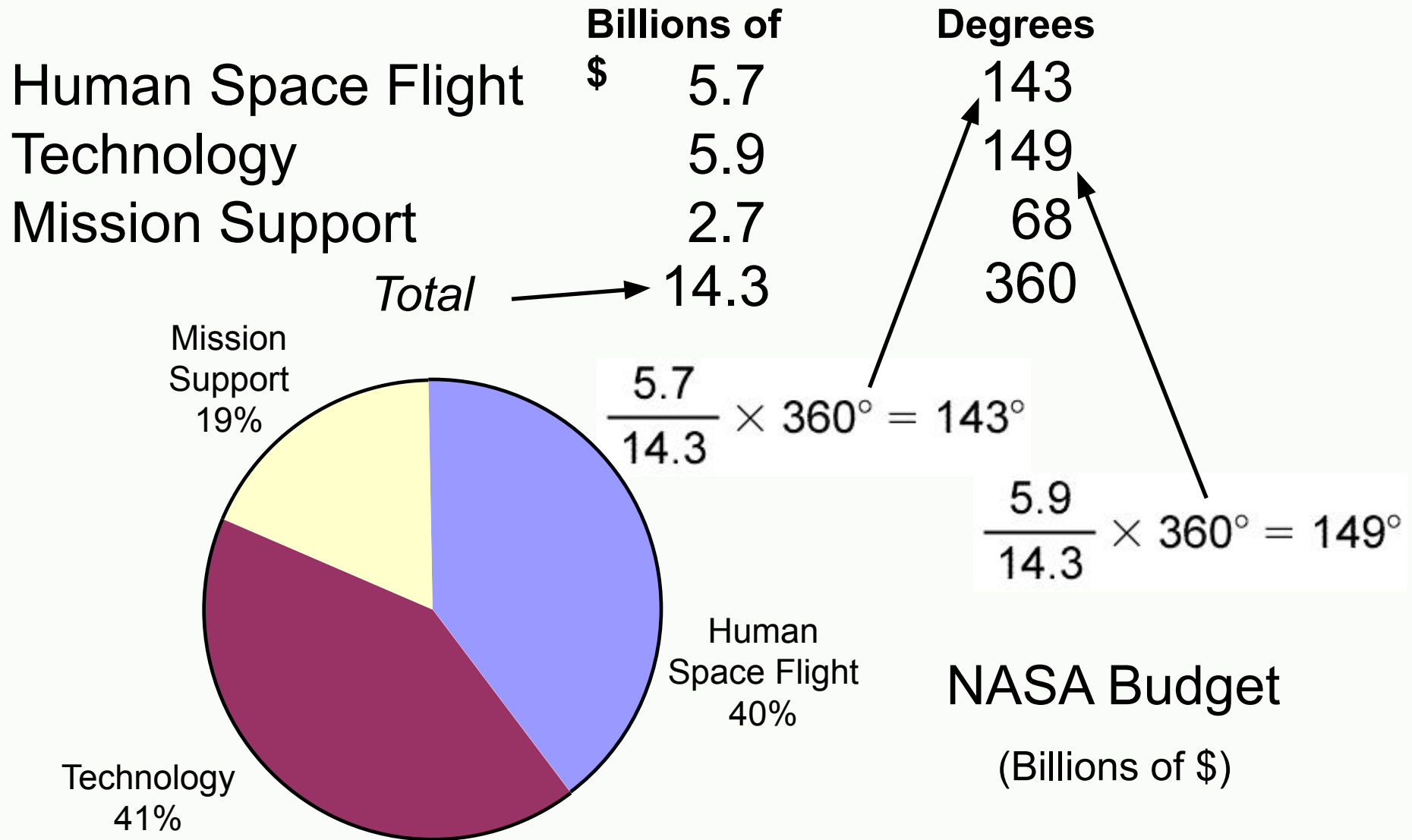
$$\frac{\text{number in category}}{\text{total number}} 360^\circ$$

NASA budget (billions of \$) divided among 3 categories.

	Billions of \$
Human Space Flight	5.7
Technology	5.9
Mission Support	2.7

Construct a pie chart for the data.

Pie Chart



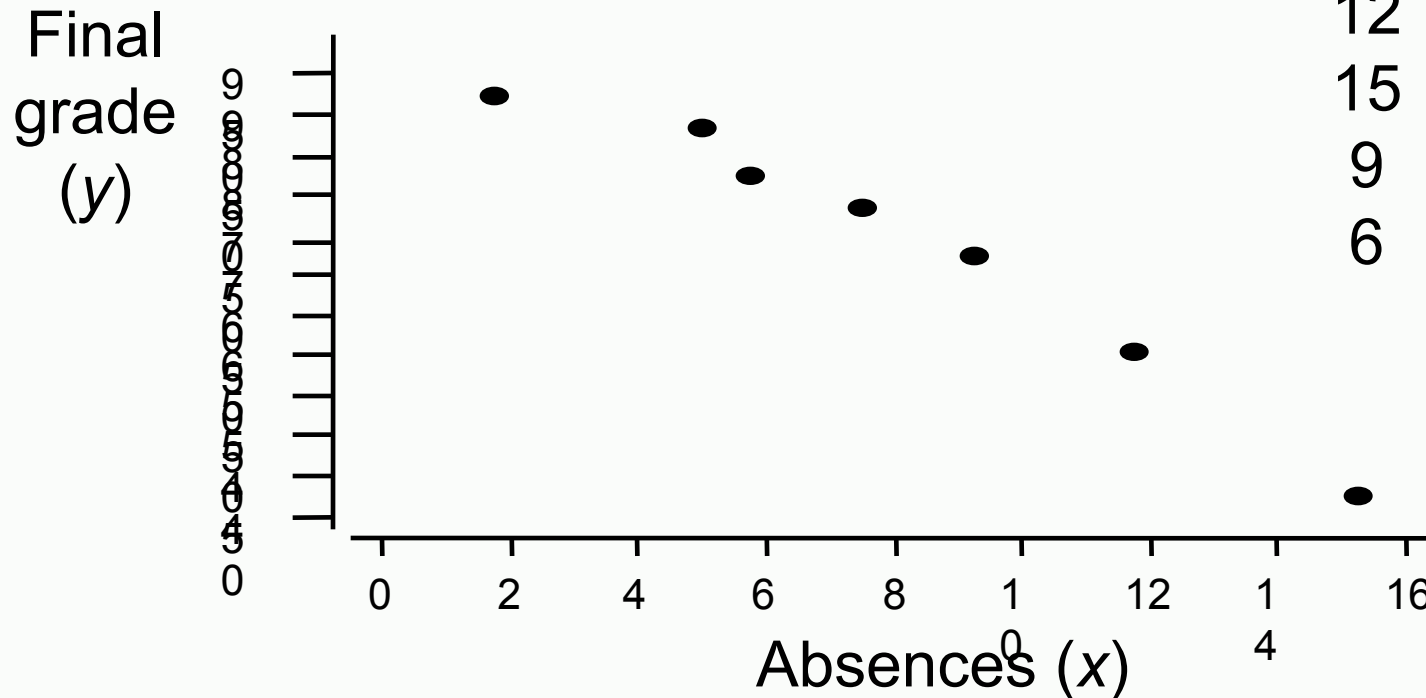
Pareto Chart

- A vertical bar graph in which the height of the bar represents frequency or relative frequency**
- The bars are in order of decreasing height**
- See example on page 53**

Scatter Plot

- Used to show the relationship between two quantitative sets of data

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

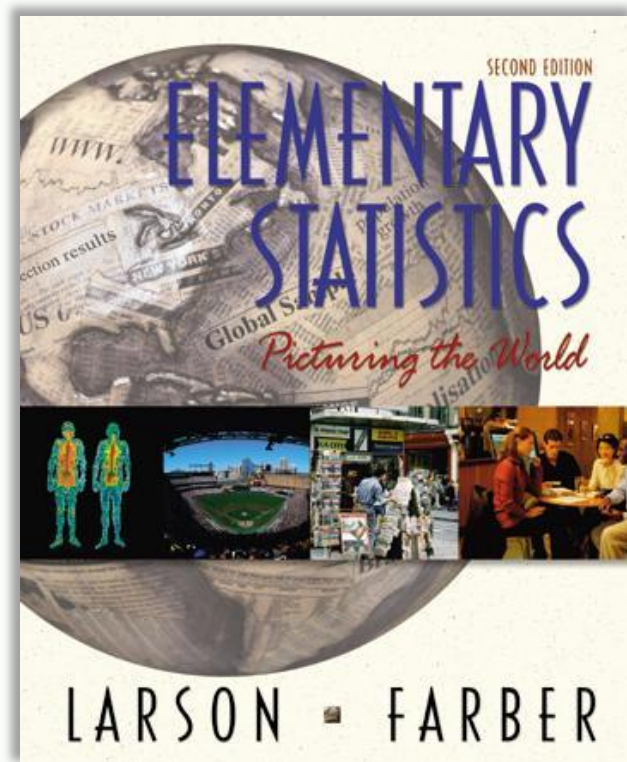


Time Series Chart / Line Graph

- Quantitative entries taken at regular intervals over a period of time
- See example on page 55

Section 2.3

Measures of Central Tendency



Measures of Central Tendency

Mean: The sum of all data values divided by the number of values

For a population:

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

For a sample:

$$\bar{x} = \frac{\sum x}{n}$$

Median: The point at which an equal number of values fall above and fall below

Mode: The value with the highest frequency

An instructor recorded the average number of absences for his students in one semester. For a random sample the data are:

2 4 2 0 40 2 4 3 6

Calculate the mean, the median, and the mode

An instructor recorded the average number of absences for his students in one semester. For a random sample the data are:

2 4 2 0 40 2 4 3 6

Calculate the mean, the median, and the mode

Mean: $\bar{x} = \frac{\sum x}{n}$ $\sum x = 63$ $n = 9$ $\bar{x} = \frac{63}{9} = 7$

Median: Sort data in order

0 2 2 2 3 4 4 6 40

The middle value is 3, so the median is 3.

Mode: The mode is 2 since it occurs the most times.

Suppose the student with 40 absences is dropped from the course. Calculate the mean, median and mode of the remaining values. Compare the effect of the change to each type of average.

2 4 2 0 2 4 3 6

Calculate the mean, the median, and the mode.

Mode: The mode is 2 since it occurs the most times.

Suppose the student with 40 absences is dropped from the course. Calculate the mean, median and mode of the remaining values. Compare the effect of the change to each type of average.

2 4 2 0 2 4 3 6

Calculate the mean, the median, and the mode.

Mean: $\bar{x} = \frac{\sum x}{n}$ $\sum x = 23$ $n = 8$ $\bar{x} = \frac{23}{8} = 2.875$

Median: Sort data in order.

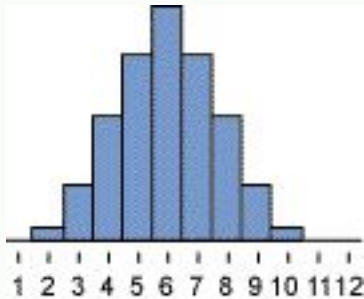
0 2 2 2 3 4 4 6

The middle values are 2 and 3, so the median is 2.5.

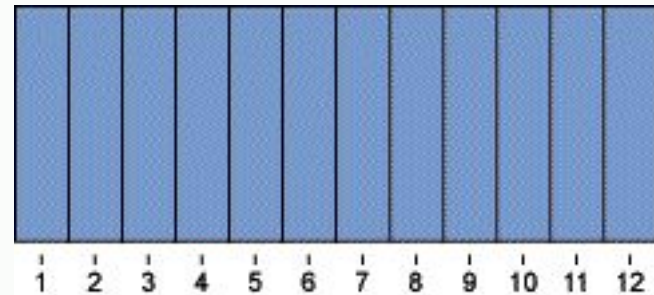
Mode: The mode is 2 since it occurs the most times.

Shapes of Distributions

Symmetric

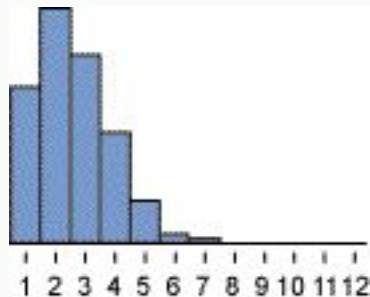


Uniform



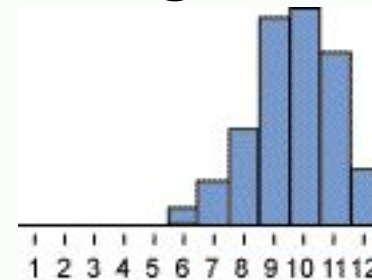
Mean = Median

**Skewed right
positive**



Mean > Median

**Skewed left
negative**



Mean < Median

Weighted Mean

A **weighted mean** is the mean of a data set whose entries have varying weights

$$\bar{X} = \frac{\sum (x \cdot w)}{\sum w}$$

where w is the weight of each entry

Weighted Mean

A student receives the following grades, A worth 4 points, B worth 3 points, C worth 2 points and D worth 1 point.

If the student has a B in 2 three-credit classes, A in 1 four-credit class, D in 1 two-credit class and C in 1 three-credit class, what is the student's mean grade point average?

Mean of Grouped Data

The **mean of a frequency distribution** for a sample is approximated by

$$\bar{X} = \frac{\sum (x \cdot f)}{n}$$

where x are the midpoints, f are the frequencies and n is $\sum f$

Mean of Grouped Data

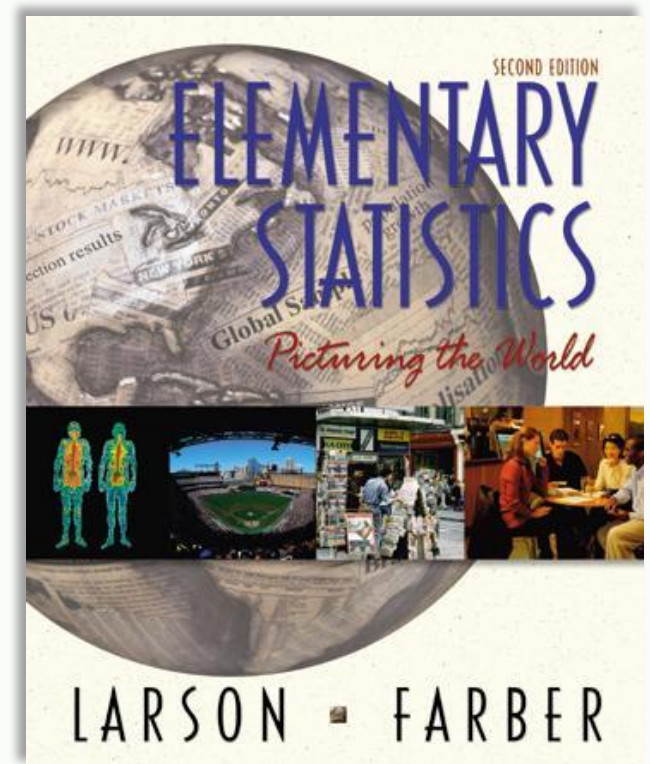
The heights of 16 students in a physical ed. class:

Height	Frequency
60-62	3
63-65	4
66-68	7
69-71	2

Approximate the mean of the grouped data

Section 2.4

Measures of Variation



Two Data Sets

Closing prices for two stocks were recorded on ten successive Fridays. Calculate the mean, median and mode for each.

Stock A			Stock B
	56	33	
	56	42	
	57	48	
	58	52	
	61	57	
	63	67	
	63	67	
	67	77	
	67	82	
	67	90	

Two Data Sets

Closing prices for two stocks were recorded on ten successive Fridays. Calculate the mean, median and mode for each.

Stock A			Stock B
	56	33	
	56	42	
	57	48	
	58	52	
	61	57	
	63	67	
	63	67	
	67	77	
	67	82	
	67	90	
Mean = 61.5			Mean = 61.5
Median = 62			Median = 62
Mode = 67			Mode = 67

Measures of Variation

Range = Maximum value – Minimum value

Range for A = $67 - 56 = \$11$

Range for B = $90 - 33 = \$57$

The range is easy to compute but only uses two numbers from a data set.

Measures of Variation

To calculate measures of variation that use every value in the data set, you need to know about deviations.


The **deviation** for each value x is the difference between the value of x and the mean of the data set.


In a **population**, the deviation for each value x is: $x - \mu$


In a **sample**, the deviation for each value x is: $x - \bar{x}$


Deviations

Stock A Deviation

56 -5.5  $56 - 61.5$

56 -5.5  $56 - 61.5$

57 -4.5  $57 - 61.5$

58 -3.5  $58 - 61.5$

61 -0.5

63 1.5

63 1.5

67 5.5

67 5.5

67 5.5

$$\mu = 61.5$$

$$\Sigma(x - \mu) = 0$$

The sum of the deviations is always zero.

Population Variance

Population Variance: The sum of the squares of the deviations, divided by N.

x	$x - \mu$	$(x - \mu)^2$
56	- 5.5	30.25
56	- 5.5	30.25
57	- 4.5	20.25
58	- 3.5	12.25
61	- 0.5	0.25
63	1.5	2.25
63	1.5	2.25
67	5.5	30.25
67	5.5	30.25
67	5.5	30.25
		<hr/>
		188.50

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

$$\sigma^2 = \frac{188.50}{10} = 18.85$$

↑
Sum of squares

Population Standard Deviation

Population Standard Deviation: The square root of the population variance.

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

$$\sigma = \sqrt{18.85} = 4.34$$

The population standard deviation is \$4.34.

Sample Variance and Standard Deviation

To calculate a sample variance divide the sum of squares by $n - 1$.

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

$$s^2 = \frac{188.50}{9} = 20.94$$

The sample standard deviation, s , is found by taking the square root of the sample variance.

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

$$s = \sqrt{20.94} = 4.58$$

Interpreting Standard Deviation

Standard deviation is a measure of the typical amount an entry deviates (is away) from the mean.

The more the entries are spread out, the greater the standard deviation.

The closer the entries are together, the smaller the standard deviation.

When all data values are equal, the standard deviation is 0.

Summary

Range = Maximum value – Minimum value

Population Variance

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

Population Standard Deviation

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

Sample Variance

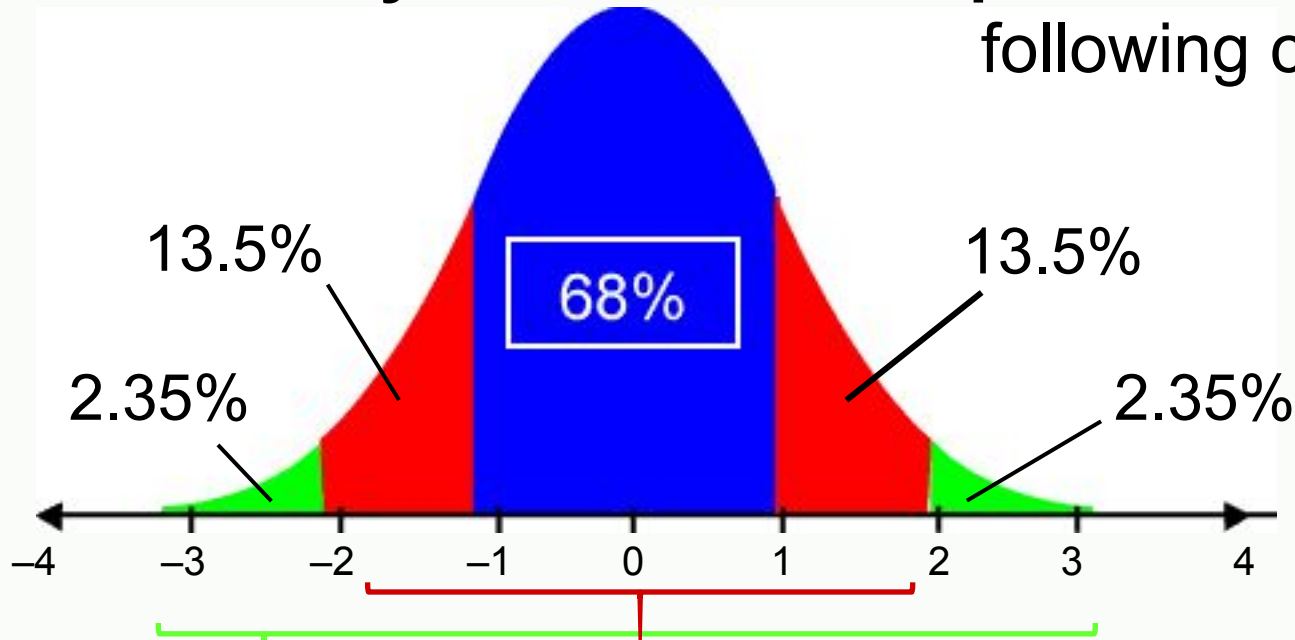
$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

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

Empirical Rule (68-95-99.7%)

Data with **symmetric bell-shaped** distribution have the following characteristics.



About **68%** of the data lies within 1 standard deviation of the mean

About **95%** of the data lies within 2 standard deviations of the mean

About **99.7%** of the data lies within 3 standard deviations of the mean

Using the Empirical Rule

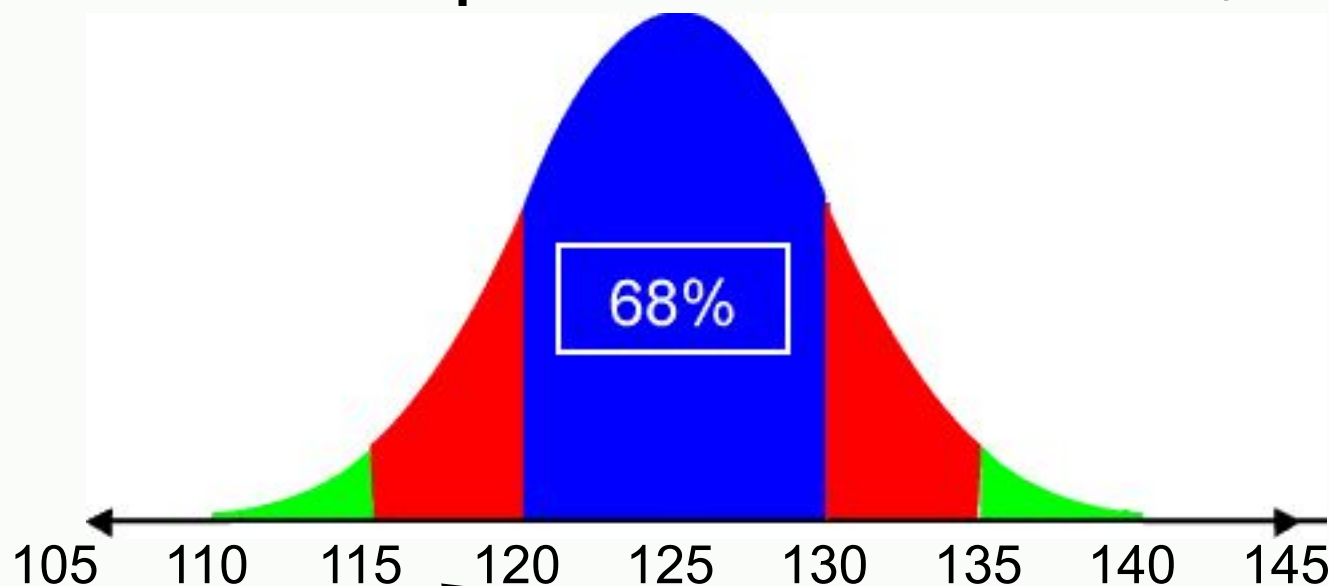
The mean value of homes on a certain street is \$125,000 with a standard deviation of \$5,000.

The data set has a bell shaped distribution.

Estimate the percent of homes between \$120,000 and \$135,000.

Using the Empirical Rule

The mean value of homes on a certain street is \$125,000 with a standard deviation of \$5,000. The data set has a bell shaped distribution. Estimate the percent of homes between \$120,000 and \$135,000.



\$120,000 is 1 standard deviation below the mean and \$135,000 is 2 standard deviations above the mean.

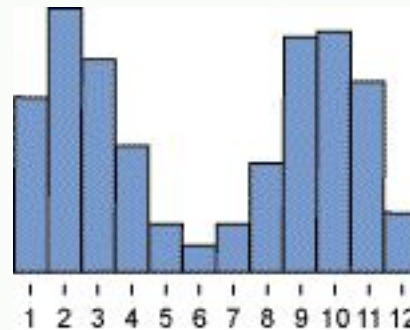
$$68\% + 13.5\% = 81.5\%$$

So, 81.5% have a value between \$120 and \$135 thousand.

Chebychev's Theorem

For any distribution regardless of shape the portion of data lying within k standard deviations ($k > 1$) of the mean is *at least* $1 - 1/k^2$.

$$\mu = 6$$
$$\sigma = 3.84$$



For $k = 2$, *at least* $1 - 1/4 = 3/4$ or 75% of the data lie within 2 standard deviation of the mean. At least 75% of the data is between -1.68 and 13.68.

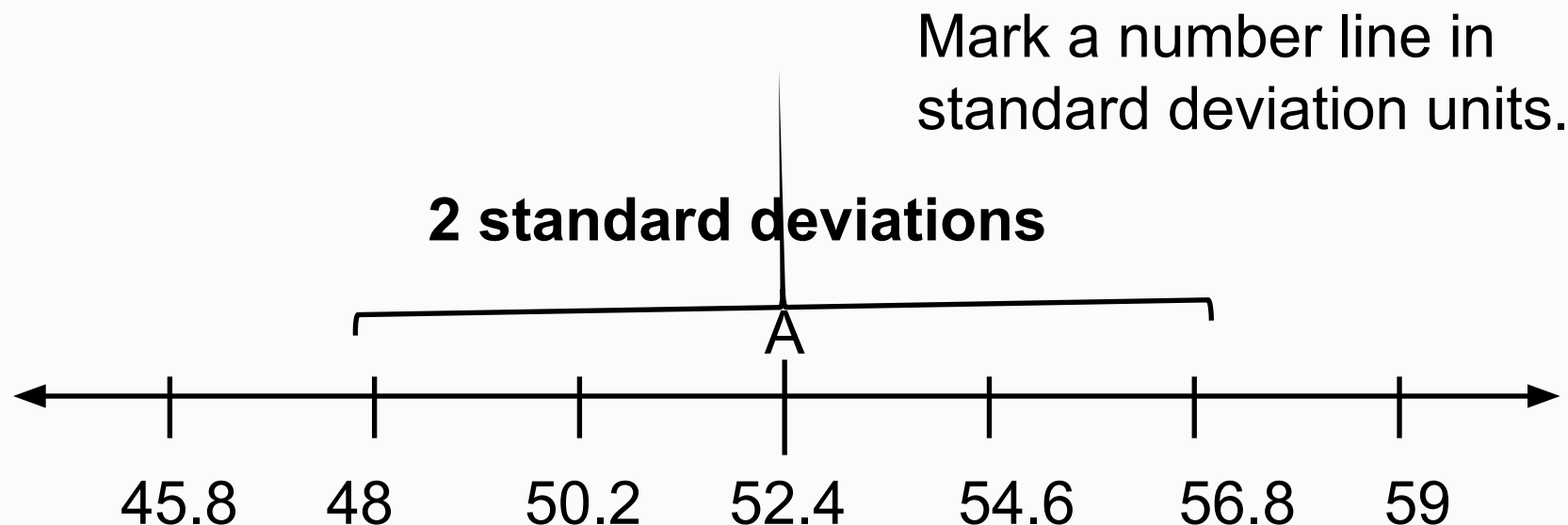
For $k = 3$, *at least* $1 - 1/9 = 8/9 = 88.9\%$ of the data lie within 3 standard deviation of the mean. At least 89% of the data is between -5.52 and 17.52.

Chebychev's Theorem

The mean time in a women's 400-meter dash is 52.4 seconds with a standard deviation of 2.2 sec. Apply Chebychev's theorem for $k = 2$.

Chebychev's Theorem

The mean time in a women's 400-meter dash is 52.4 seconds with a standard deviation of 2.2 sec. Apply Chebychev's theorem for $k = 2$.



At least 75% of the women's 400-meter dash times will fall between 48 and 56.8 seconds.

Standard Deviation of Grouped Data

Sample standard deviation = $s = \sqrt{\frac{\sum (x - \bar{x})^2 f}{n - 1}}$

x	f	xf	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 f$
-----	-----	------	---------------	-------------------	---------------------

f is the frequency, n is total frequency, $\bar{x} = \frac{\sum xf}{n}$

See example on pg 82

Estimates with Classes

When a frequency distribution has classes, you can estimate the sample mean and standard deviation by using the midpoints of each class.

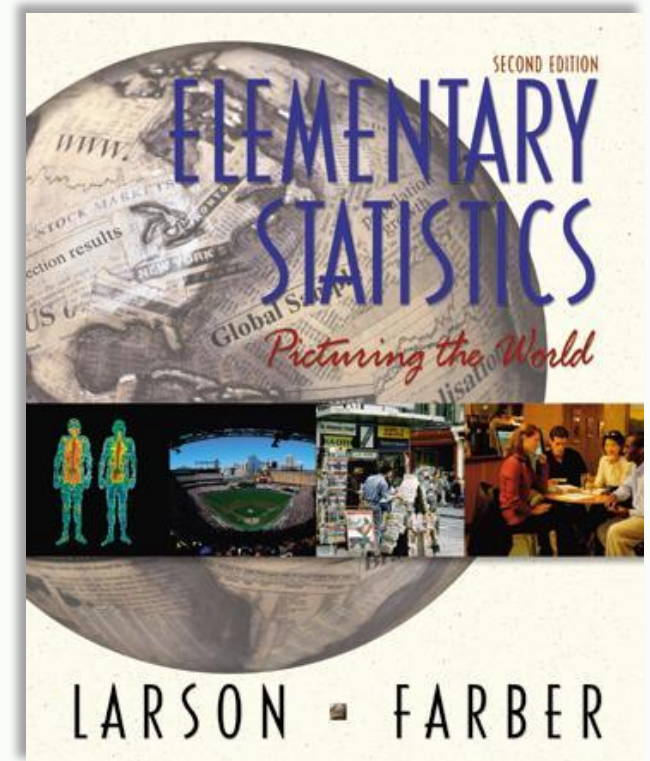
$$\bar{x} = \frac{\sum xf}{n}$$

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

x is the midpoint, f is the frequency, n is total frequency

Section 2.5

Measures of Position



Quartiles

Fractiles – numbers that divide an ordered data set into equal parts.

Quartiles (Q_1 , Q_2 and Q_3) - divide the data set into 4 equal parts.

Q_2 is the same as the median.

Q_1 is the median of the data below Q_2 .

Q_3 is the median of the data above Q_2 .

Quartiles

You are managing a store. The average sale for each of 27 randomly selected days in the last year is given. Find Q_1 , Q_2 , and Q_3 .

28 43 48 51 43 30 55 44 48 33 45 37
37 42 27 47 42 23 46 39 20 45 38 19
17 35 45

Finding Quartiles

The data in ranked order ($n = 27$) are:

17 19 20 23 27 28 30 33 35 37 37 38 39 42
42 43 43 44 45 45 45 46 47 48 48 51 55.

The median = Q_2 = 42.

There are 13 values above/below the median.

Q_1 is 30.

Q_3 is 45.

Interquartile Range (IQR)

Interquartile Range – the difference between the third and first quartiles

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

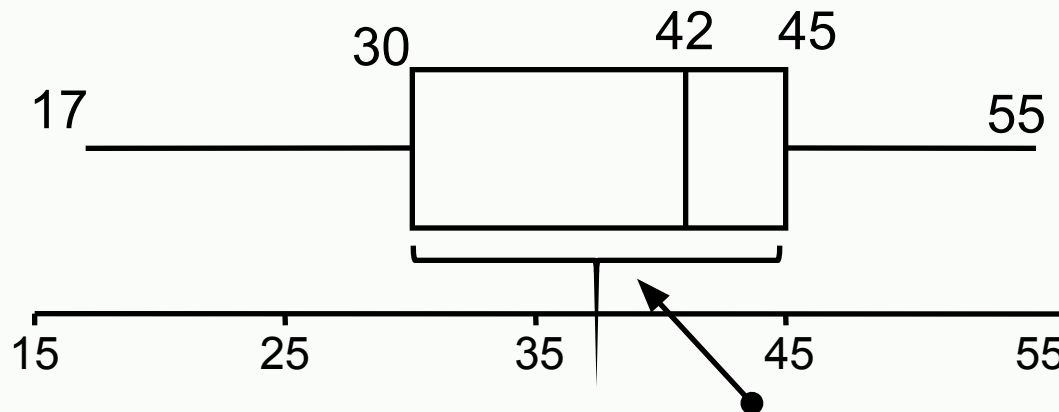
The Interquartile Range is $Q_3 - Q_1 = 45 - 30 = 15$

Any data value that is more than 1.5 IQRs to the left of Q_1 or to the right of Q_3 is an outlier

Box and Whisker Plot

A box and whisker plot uses 5 key values to describe a set of data. Q_1 , Q_2 and Q_3 , the minimum value and the maximum value.

Q_1	30
Q_2 = the median	42
Q_3	45
Minimum value	17
Maximum value	55



$$\text{Interquartile Range} = 45 - 30 = 15$$

Percentiles

Percentiles divide the data into 100 parts. There are 99 percentiles: $P_1, P_2, P_3 \dots P_{99}$.

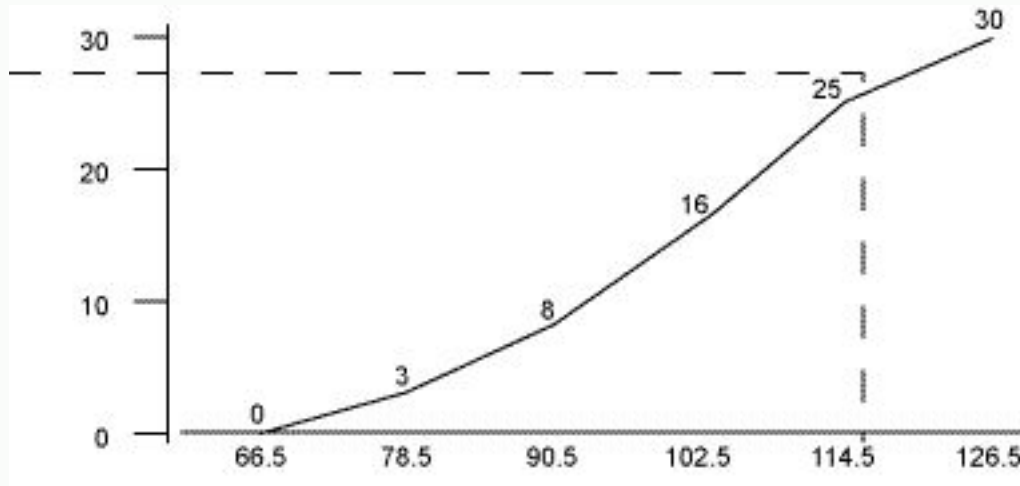
$$P_{50} = Q_2 = \text{the median}$$

$$P_{25} = Q_1$$

$$P_{75} = Q_3$$

A 63rd percentile score indicates that score is greater than or equal to 63% of the scores and less than or equal to 37% of the scores.

Percentiles



Cumulative distributions can be used to find percentiles.

114.5 falls on or above 25 of the 30 values.

$$25/30 = 83.33.$$

So you can approximate $114 = P_{83}$.

Standard Scores

Standard score or z-score - represents the number of standard deviations that a data value, x , falls from the mean.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{X - \mu}{\sigma}$$

Standard Scores

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

The test scores for a civil service exam have a mean of 152 and standard deviation of 7. Find the standard z-score for a person with a score of:

(a) 161

(b) 148

(c) 152

Calculations of z-Scores

(a)

$$z = \frac{161 - 152}{7}$$

$$z = 1.29$$

A value of $x = 161$ is 1.29 standard deviations above the mean.

(b)



$$z = \frac{148 - 152}{7}$$

$$z = -0.57$$

A value of $x = 148$ is 0.57 standard deviations below the mean.

(c)



$$z = \frac{152 - 152}{7}$$

$$z = 0$$

A value of $x = 152$ is equal to the mean.

Standard Scores

When a distribution is approximately bell shaped, about 95% of the data lie within 2 standard deviations of the mean. When this is transformed to z-scores, about 95% of the z-scores should fall between -2 and 2.

A z-score outside of this range is considered unusual and a z-score less than -3 or greater than 3 would be very unusual.