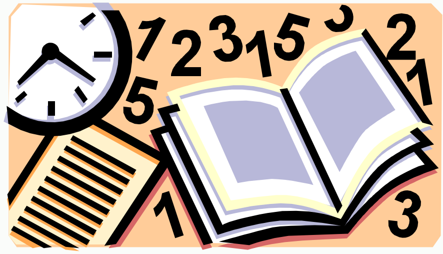
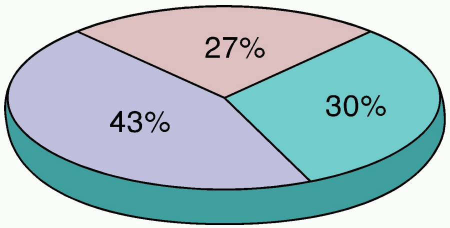
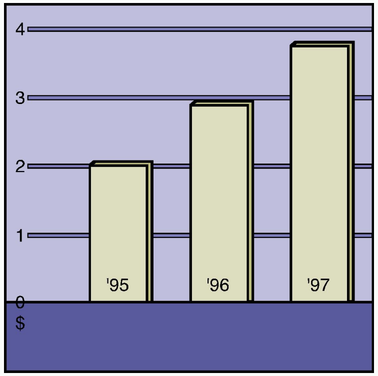
2 Descriptive Statistics



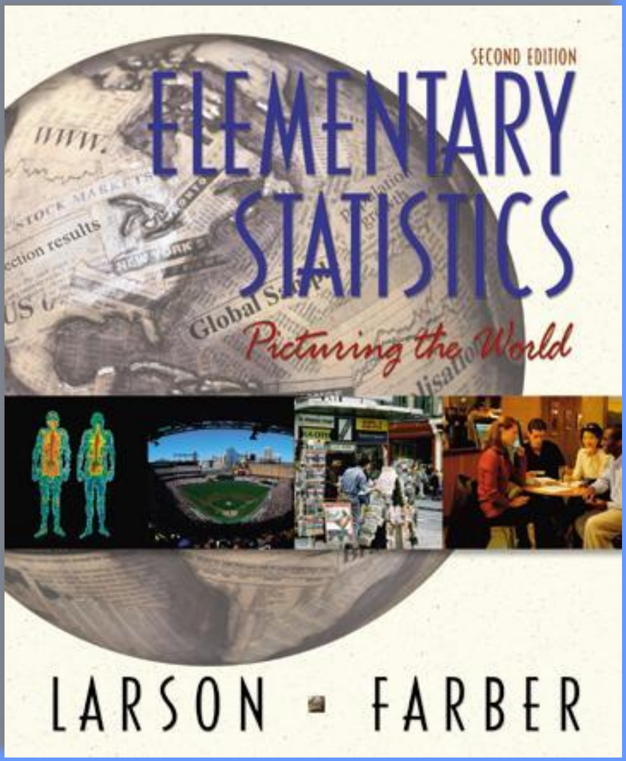


Elementary Statistics

Larson Farber

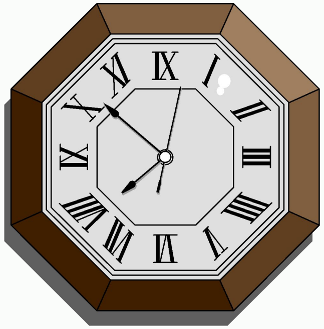


Section 2.1



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.

Key values:

Minimum value = 67
Maximum value = 125

Steps to Construct a Frequency Distribution

1. Choose the number of classes

Should be between 5 and 15. (For this problem use 5)

2. Calculate the Class Width

Find the range = maximum value – minimum. Then divide this by the number of classes. Finally, round up to a convenient number. (125 - 67) / 5 = 11.6 Round *up* to 12

3. Determine Class Limits

The lower class limit is the lowest data value that belongs in a class and the upper class limit it the highest. Use the minimum value as the lower class limit in the first class. (67)

4. Mark a tally | in appropriate class for each data value.

After all data values are tallied, count the tallies in each class for the class frequencies.

Construct a Frequency Distribution

Minimum = 67, Maximum = 125
Number of classes = 5
Class width = 12

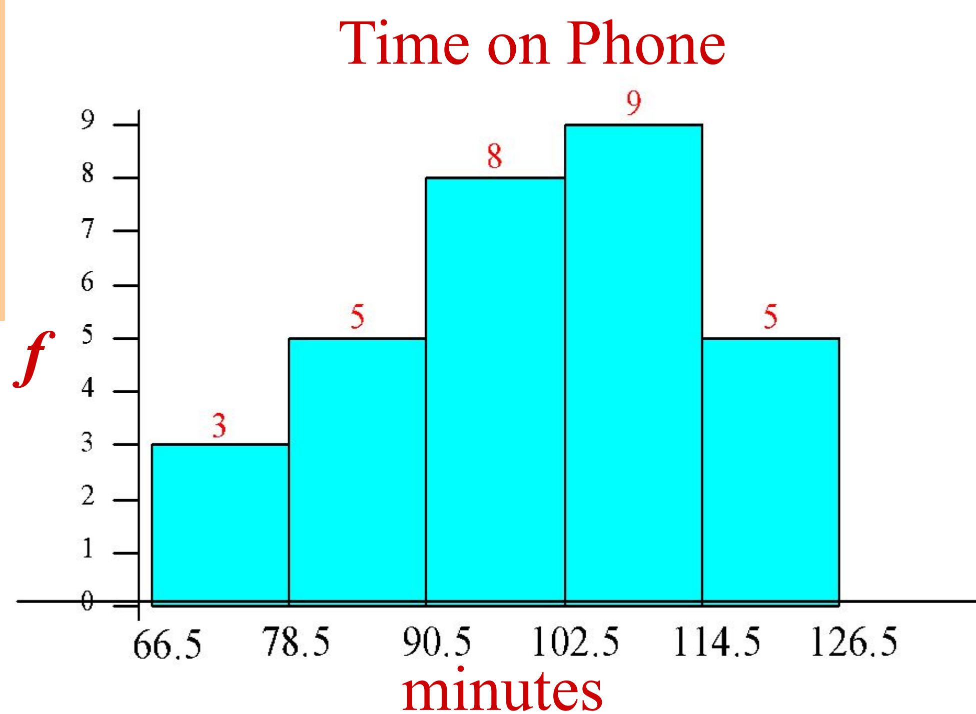
Class Limits		Tally	f
67	78	𝍷𝍷𝍷	3
79	90	𝍸	5
91	102	𝍸 𝍷𝍷𝍷	8
103	114	𝍸 𝍷𝍷𝍷𝍷	9
115	126	𝍸	5

$\sum f = 30$

Do all lower class limits first.

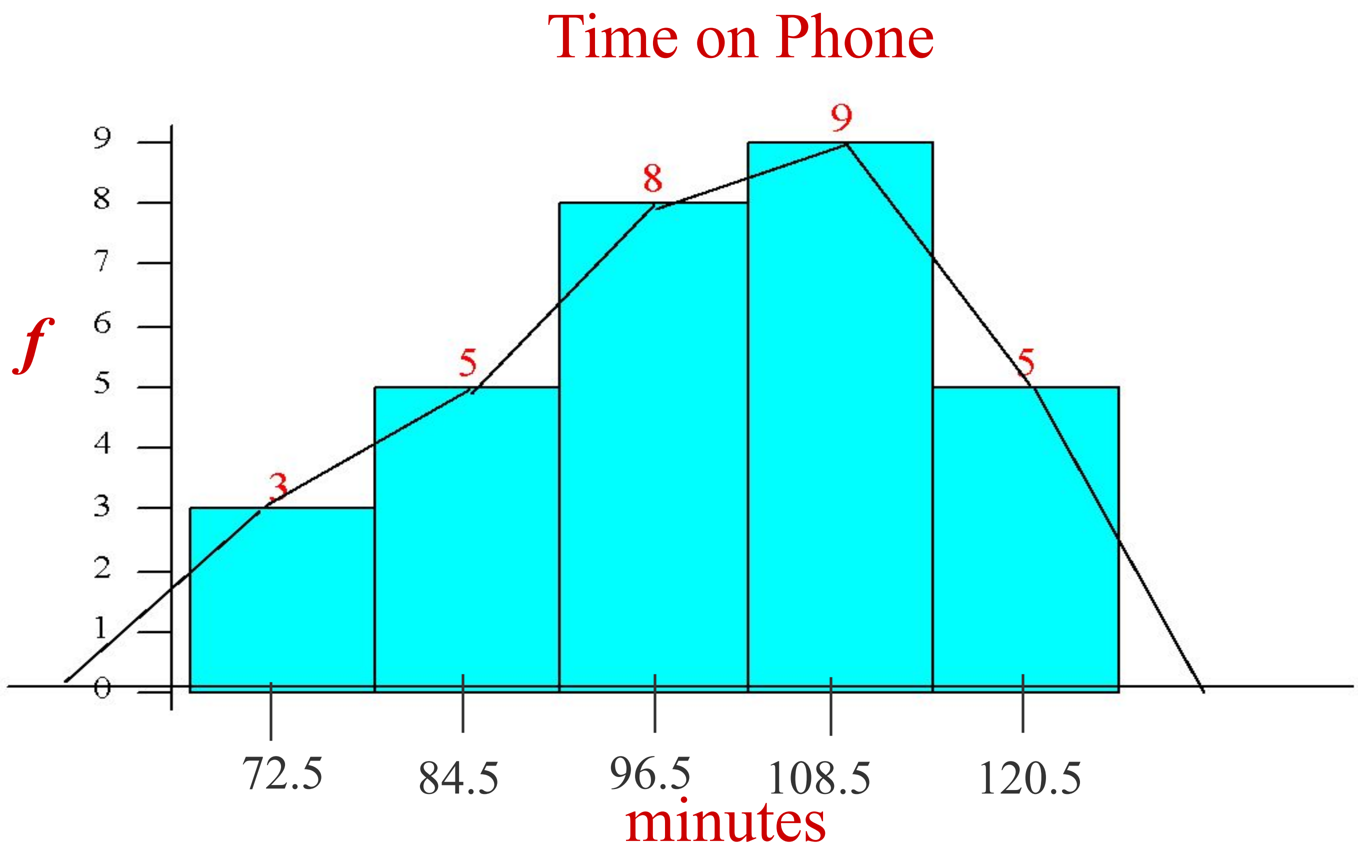
Frequency Histogram

Class	f	Boundaries
67 – 78	3	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



Frequency Polygon

Class	f
67 - 78	3
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.

Other Information

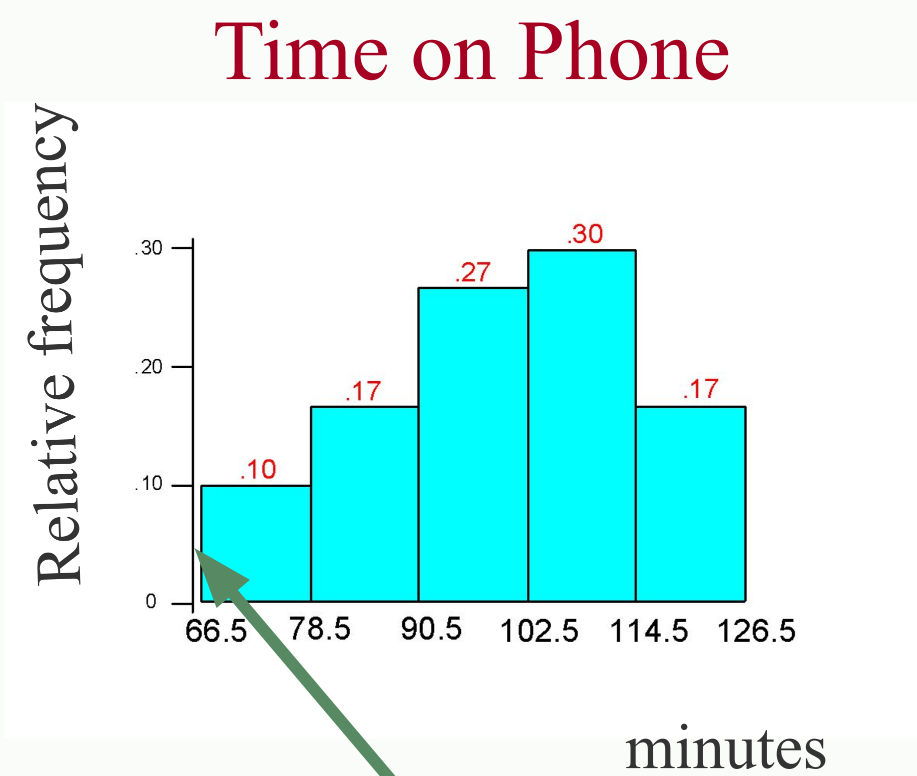
Midpoint: (lower limit + upper limit) / 2

Relative frequency: class frequency/total frequency

Cumulative frequency: Number of values in that class or in lower.

Class	f	Midpoint	Relative frequency	Cumulative Frequency
		(67+ 78)/2	3/30	
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

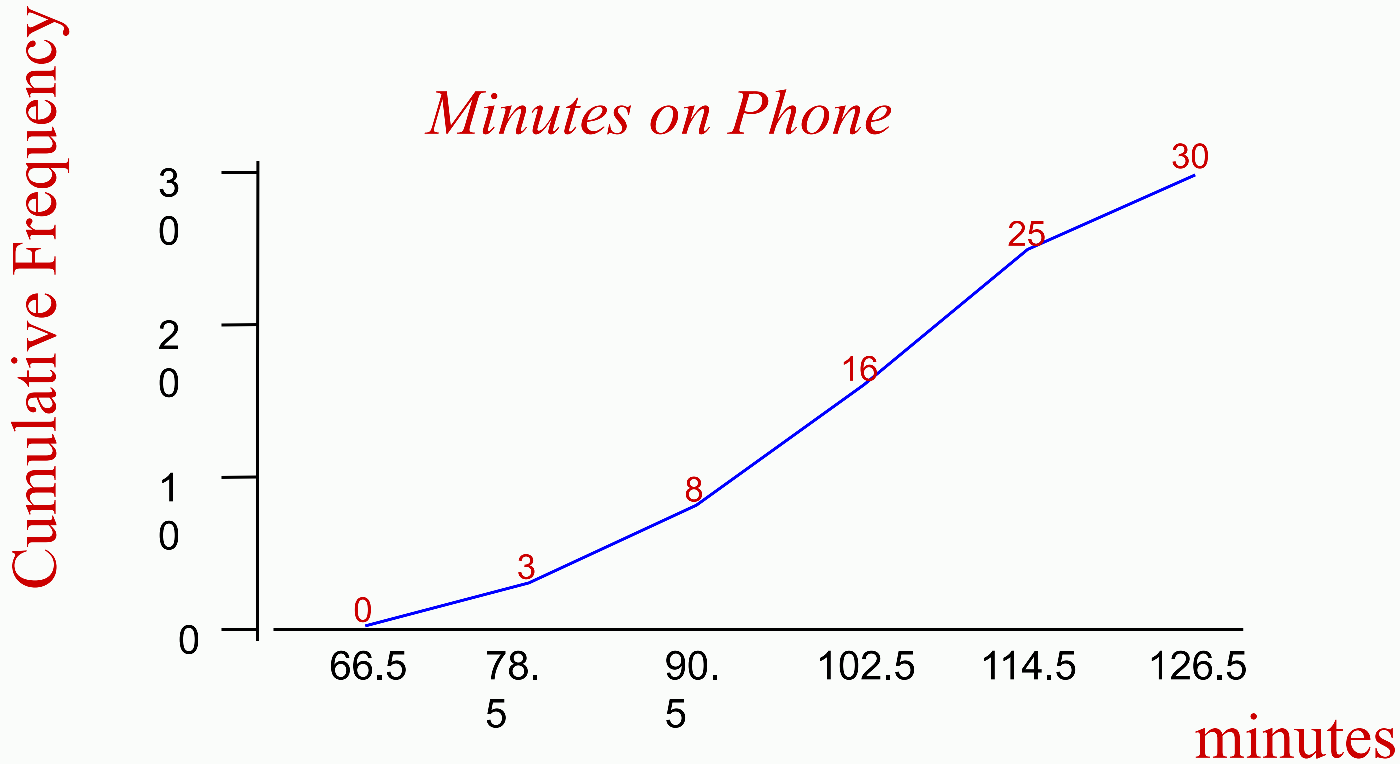
Relative Frequency Histogram



Relative frequency on vertical scale

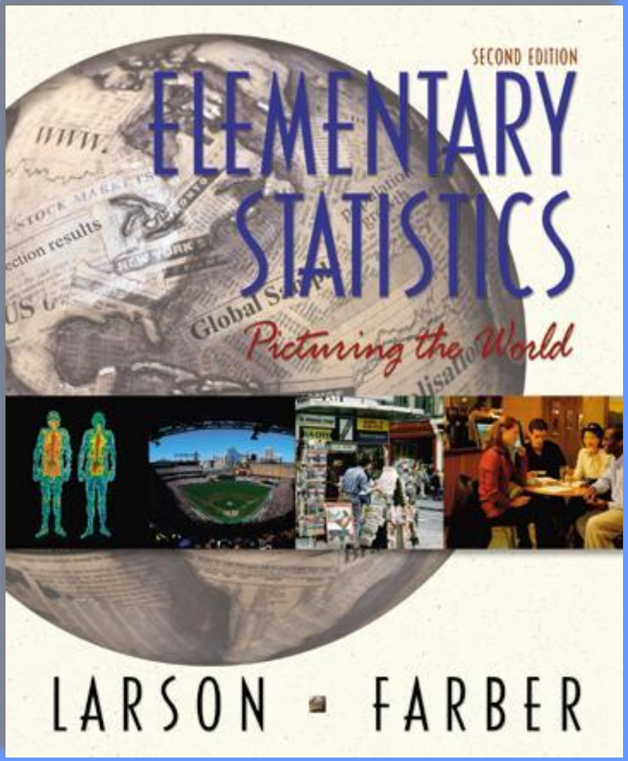
Ogive

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

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

102 124 108 86 103 82

Stem	Leaf		
6 |			
7 |			
8 |	6	2	
9 |			
10 |	2	8	3
11 |			
12 |	4		

To see complete display, go to next slide.

Stem-and-Leaf Plot

Key: 6 | 7 means 67

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

Stem-and-Leaf with two lines per stem

Key: 6 | 7 means 67

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

1st line digits 0 1 2 3 4

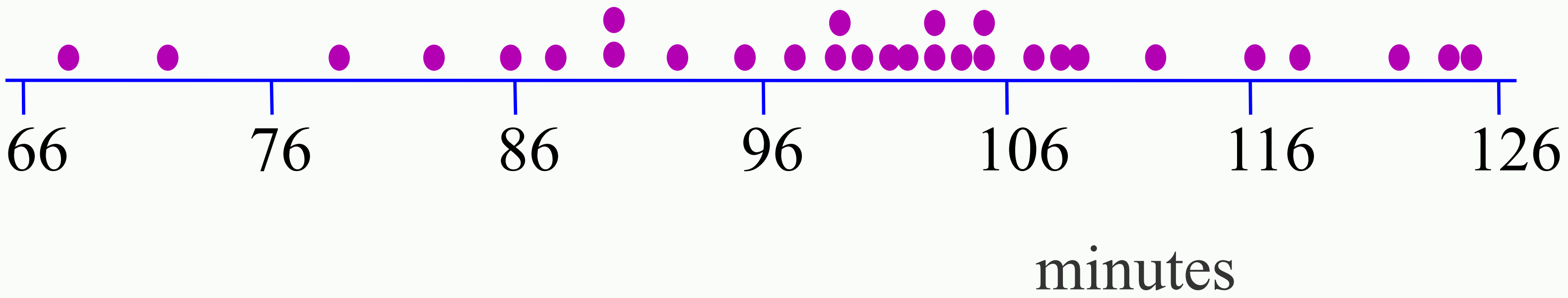
2nd line digits 5 6 7 8 9

1st line digits 0 1 2 3 4

2nd line digits 5 6 7 8 9

Dotplot

Phone

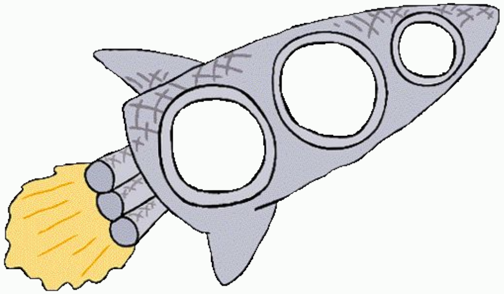


Pie Chart

- Used to describe parts of a whole
- Central Angle for each segment $\frac{\text{number in category}}{\text{total number}} \times 360^{o}$

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

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



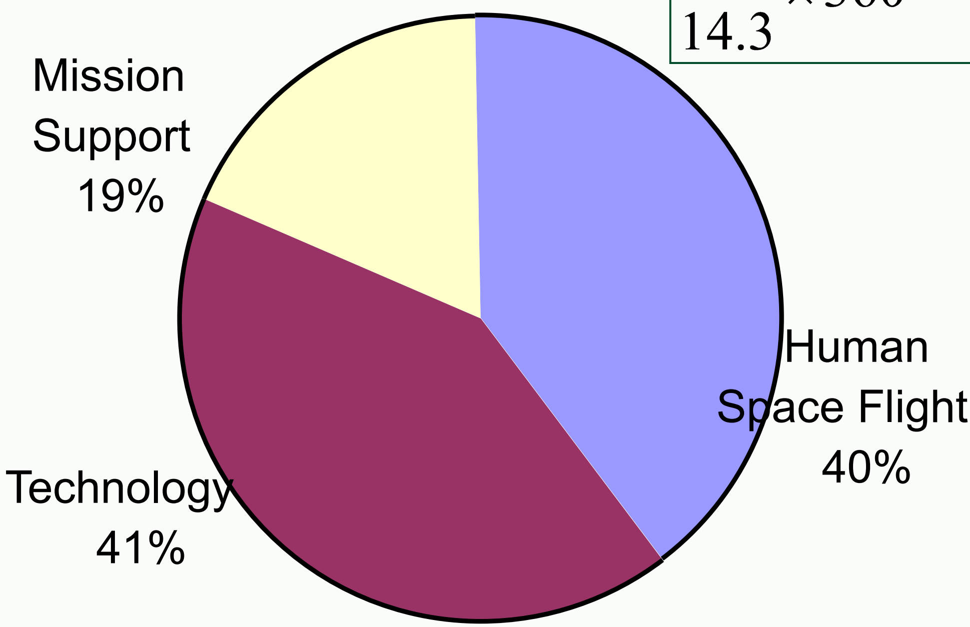
Construct a pie chart for the data.

Pie Chart

	Billions of $	Degrees
Human Space Flight	5.7	143
Technology	5.9	149
Mission Support	2.7	68
Total	14.3	360

$$\frac{5.7}{14.3} \times 360^\circ = 143^\circ$$

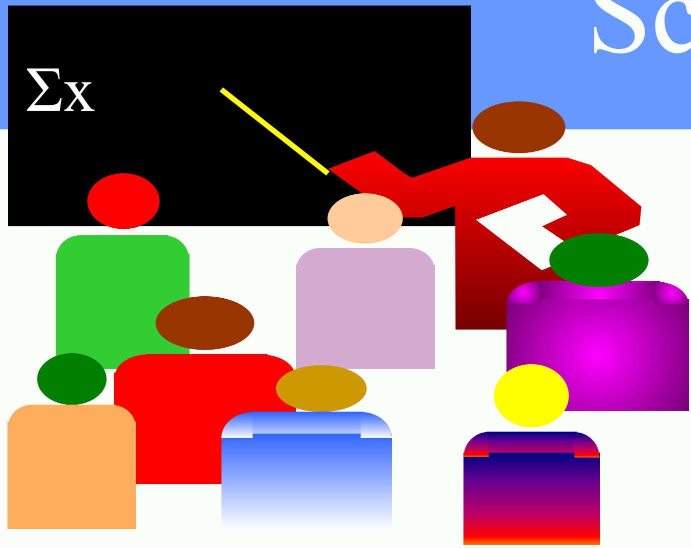
$$\frac{5.9}{14.3} \times 360^\circ = 149^\circ$$



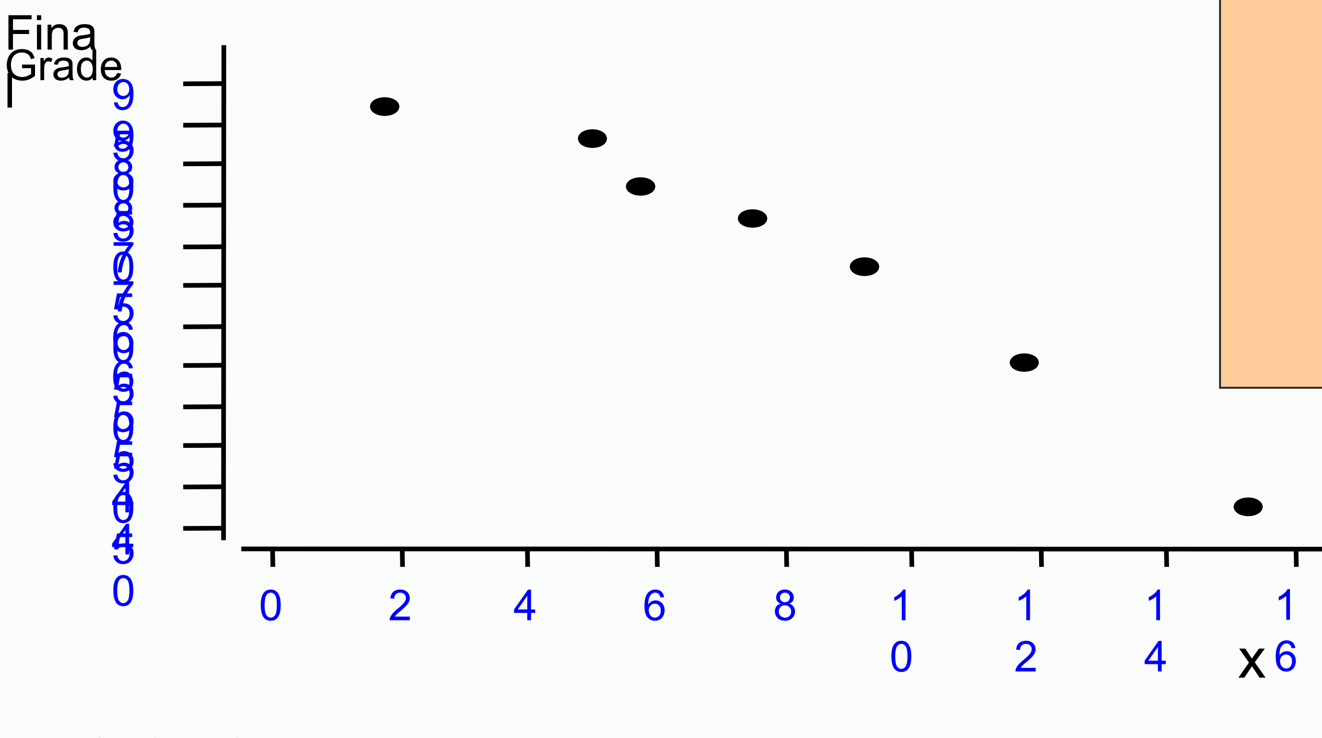
NASA Budget

(Billions of $)

Scatter Plot

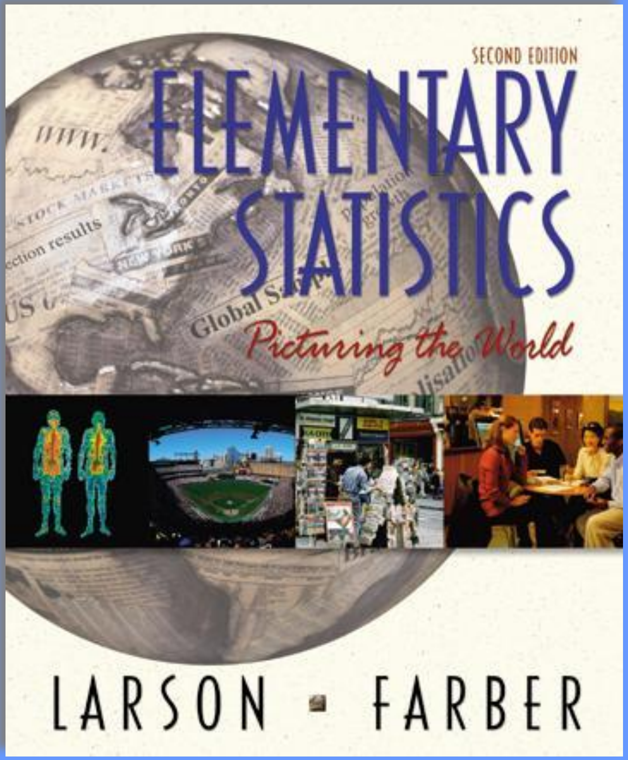


Absences	Grade
x	y
8	78
2	92
5	90
12	58
15	43
9	74
6	81



Section 2.3

Measures of Central Tendency



Measures of Central Tendency

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

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

The mean incorporates every value in the data set.

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

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

Median: Sort data in order

0 2 2 2 $\boxed{3}$ 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

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

Median: Sort data in order

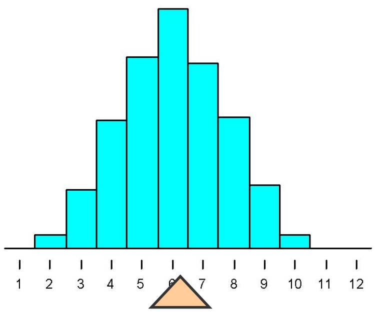
0 2 2 $\boxed{2 \quad 3}$ 4 4 6

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

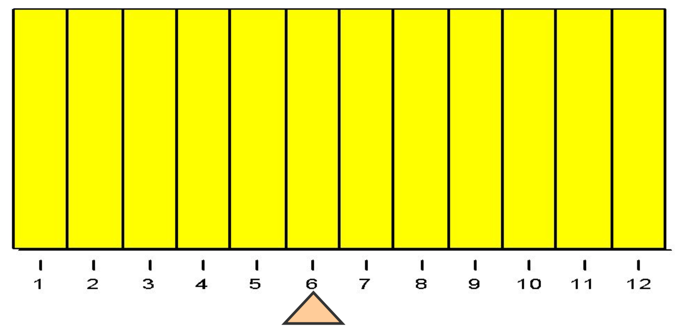
Mode: The mode is 2 since it occurs the most.

Shapes of Distributions

Symmetric

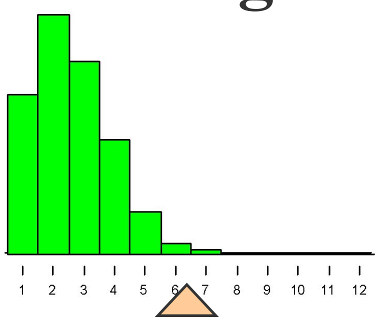


Uniform



Mean = Median

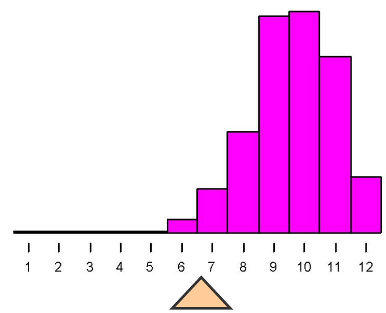
Skewed right



Mean is right of median

Mean > Median

Skewed left



Mean is left of median.

Mean < Median

Outliers

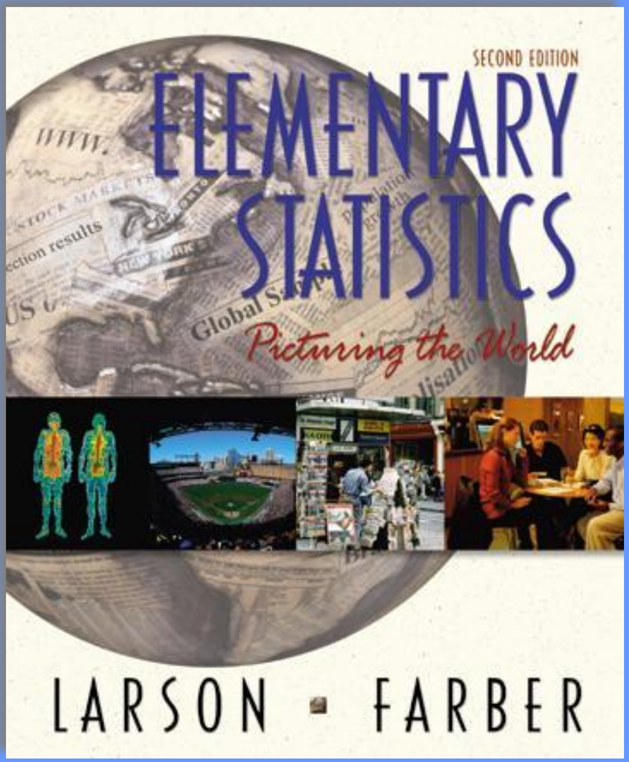
What happened to our mean, median and mode when we removed 40 from the data set?

40 is an **outlier**

- An outlier is a value that is much larger or much smaller than the rest of the values in a data set.
- Outliers have the biggest effect on the **mean**.

Section 2.4

Measures of Variation



Measures of Variation

- Range = Maximum value - Minimum value
- Variance is the sum of the deviations from the mean divided by $n - 1$.
- Standard deviation is the square root of the variance.

- **Example:** A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results are shown below.
- Brand A: 10, 60, 50, 30, 40, 20
- Brand B: 35, 45, 30, 35, 40, 25

Find the mean and range for each brand, then create a stack plot for each. Compare your results.

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

Stock A:
Mean = 61.5
Median =62
Mode= 67

Stock B:
Mean = 61.5
Median =62
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 2 numbers from a data set.

To Calculate Variance & Standard Deviation:

1. Find the **deviation,** the difference between each data value, x, and the mean, $\bar{x}$.

2. Square each deviation.

3. Find the sum of all squares from step 2.

4. Divide the result from step 3 by $n-1$, where n = the total number of data values in the set.

Deviations

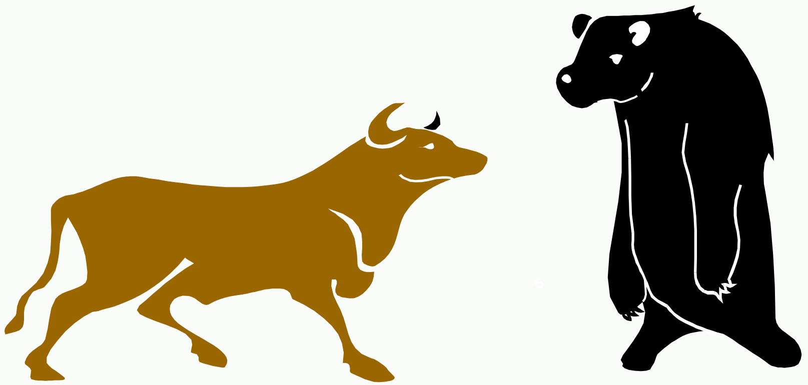
Stock A	Deviation
56	-5.5
56	-5.5
57	-4.5
58	-3.5
61	-0.5
63	1.5
63	1.5
67	5.5
67	5.5
67	5.5

56 - 61.5

56 - 61.5

57 - 61.5

$$\bar{x} = 61.5$$



$$\sum(x - \bar{x}) = 0$$

The sum of the deviations is always zero.

Variance

Variance: The sum of the squares of the deviations, divided by n -1.

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
		188.50

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

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

Sum of squares

Standard Deviation

Standard Deviation The square root of the variance.

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

$$\sigma = \sqrt{20.94}$$

The standard deviation is 4.58.

Summary

Range = Maximum value - Minimum value

Variance

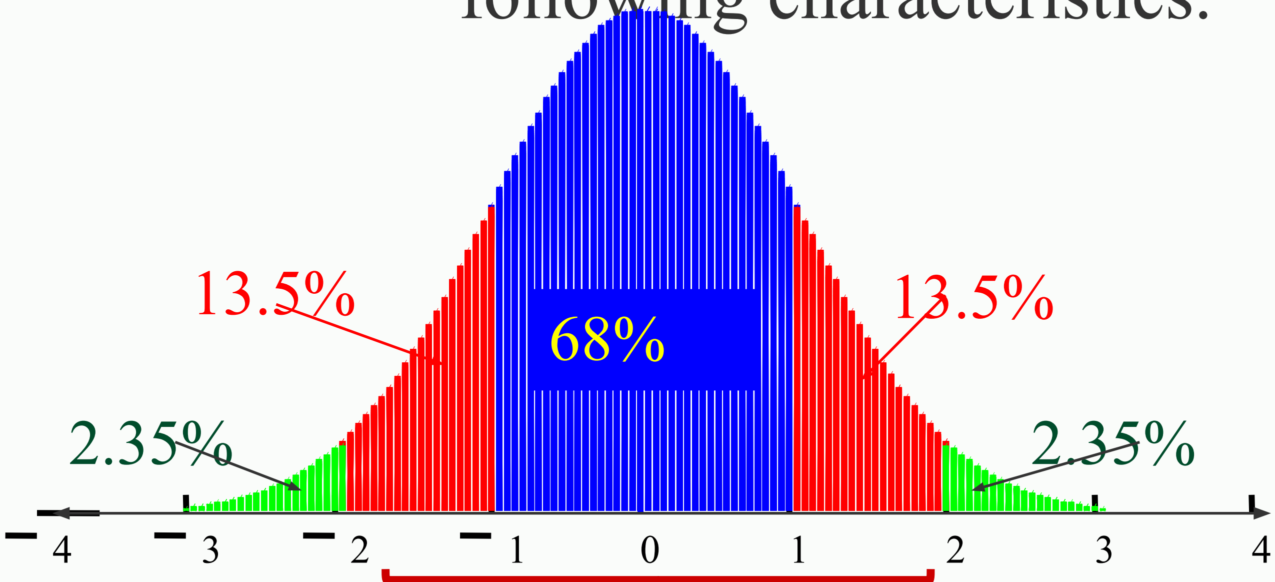
$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Standard Deviation

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

Empirical Rule (68-95-99.7%)

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



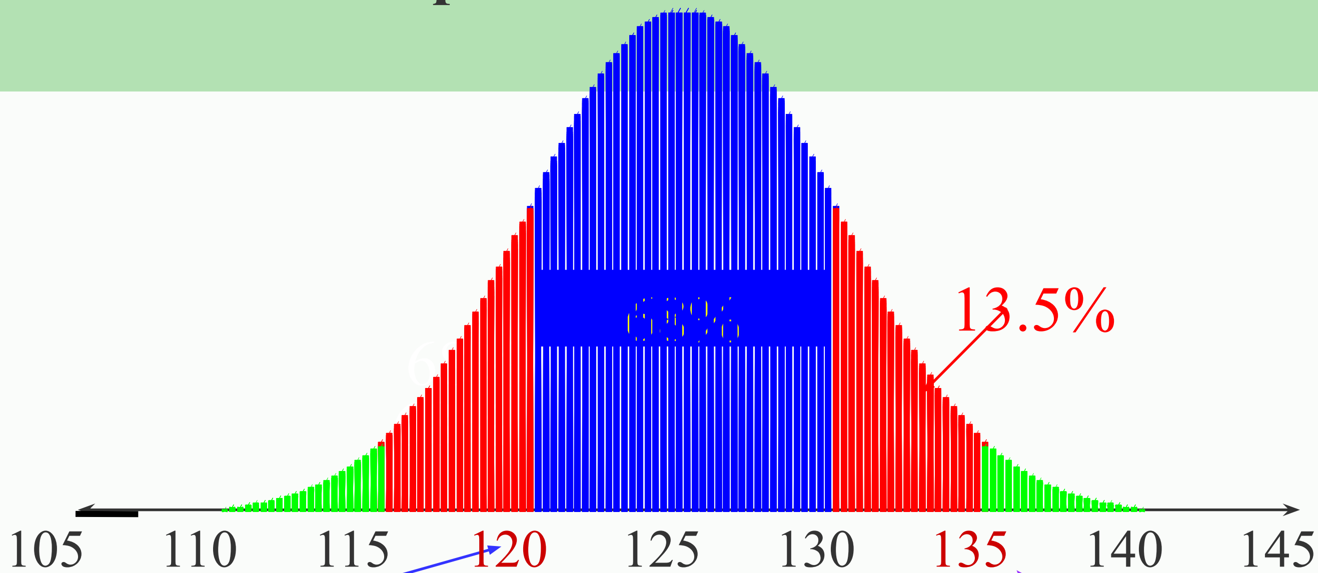
About 68% of the data lies within 1 standard deviation of the mean

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 street is $125 thousand with a standard deviation of $5 thousand. The data set has a bell shaped distribution. Estimate the percent of homes between $120 and $135 thousand



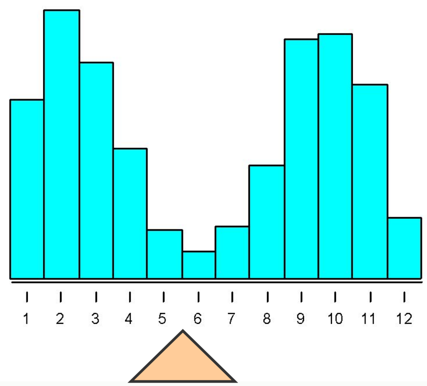
$120 thousand is 1 standard deviation below the mean and $135 thousand is 2 standard deviation above the mean.68% + 13.5% = 81.5%

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 lies within 2 standard deviation of the mean.

For k = 3, ***at least*** 1-1/9 = 8/9= 88.9% of the data lies within 3 standard deviation of the mean.

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.

Mark a number line in standard deviation units.

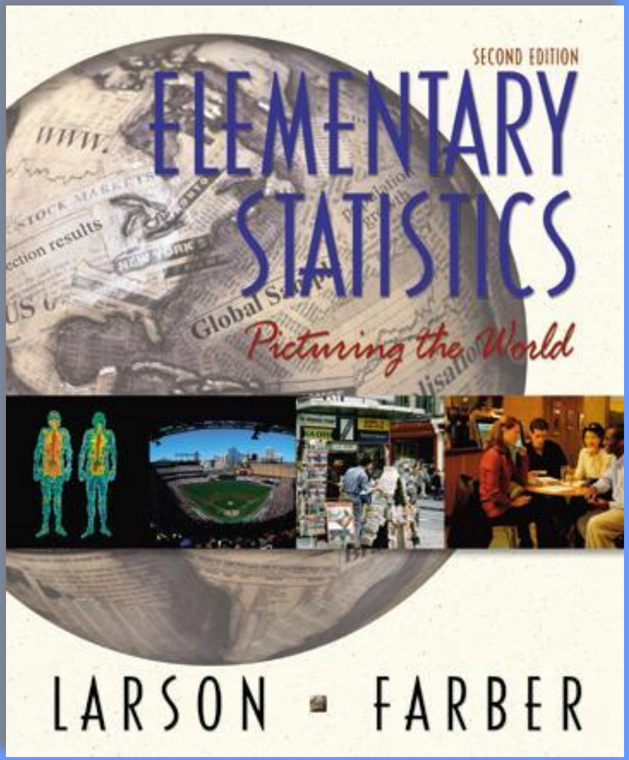
2 standard deviations

45.8 48 50.2 52.4 54.6 56.8 59

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

Section 2.5

Measures of Position



Quartiles

3 quartiles Q_1, Q_2 and Q_3 divide the data 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

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 .

Median Q2=

Q1= Q3=

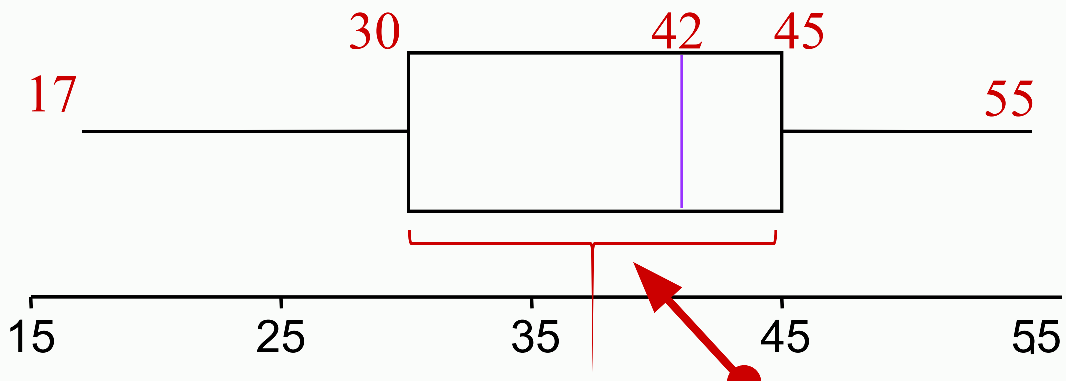
Interquartile Range (IQR)= Q3-Q1

IQR =

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



Interquartile Range = 45-30=15

Percentiles

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

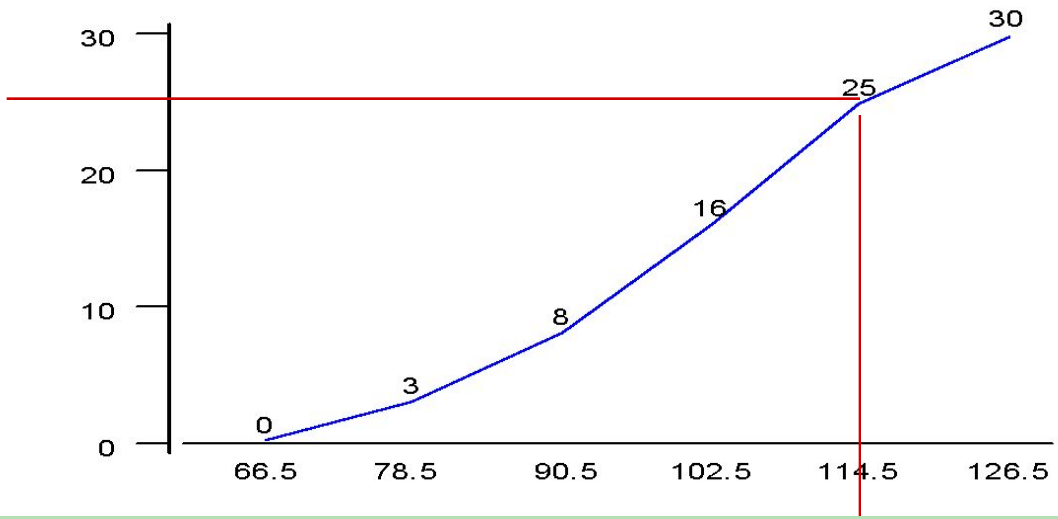
$P_{50} = Q_2$ = the median

$P_{25} = Q_1$

$P_{75} = Q_3$

A 63nd 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

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

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

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.