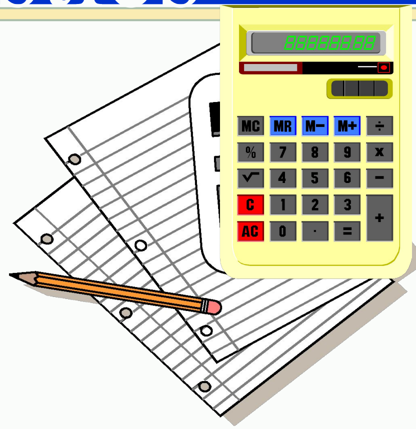
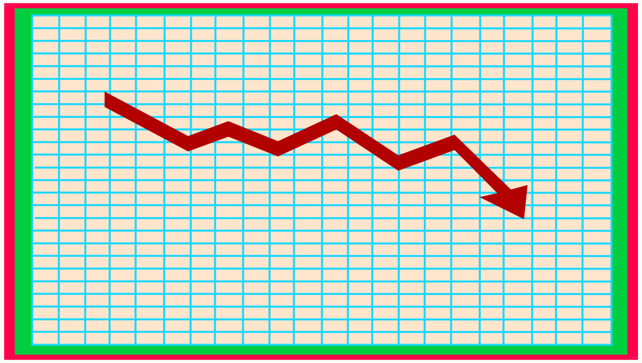
Chapter 2

Descriptive Statistics

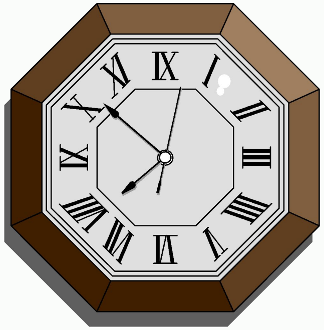


Elementary Statistics

Larson Farber



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

Frequency Distributions

- **Decide on the number of classes** (For this problem use 5)
- **Calculate the Class Width**
  - **(125 - 67) / 5 = 11.6 Round *up* to 12**
- **Determine Class Limits**
- **Mark a tally in appropriate class for each data value**

Class Limits		Tally	f 3
67	78	ǀǀǀ	
79	90	卌	5
91	102	卌 ǀǀǀ	8
103	114	卌 ǀǀǀǀ	9
115	126	卌	5

$\sum f = 30$

Do all lower class limits first.

Other Information

Midpoint: (lower limit + upper limit) / 2

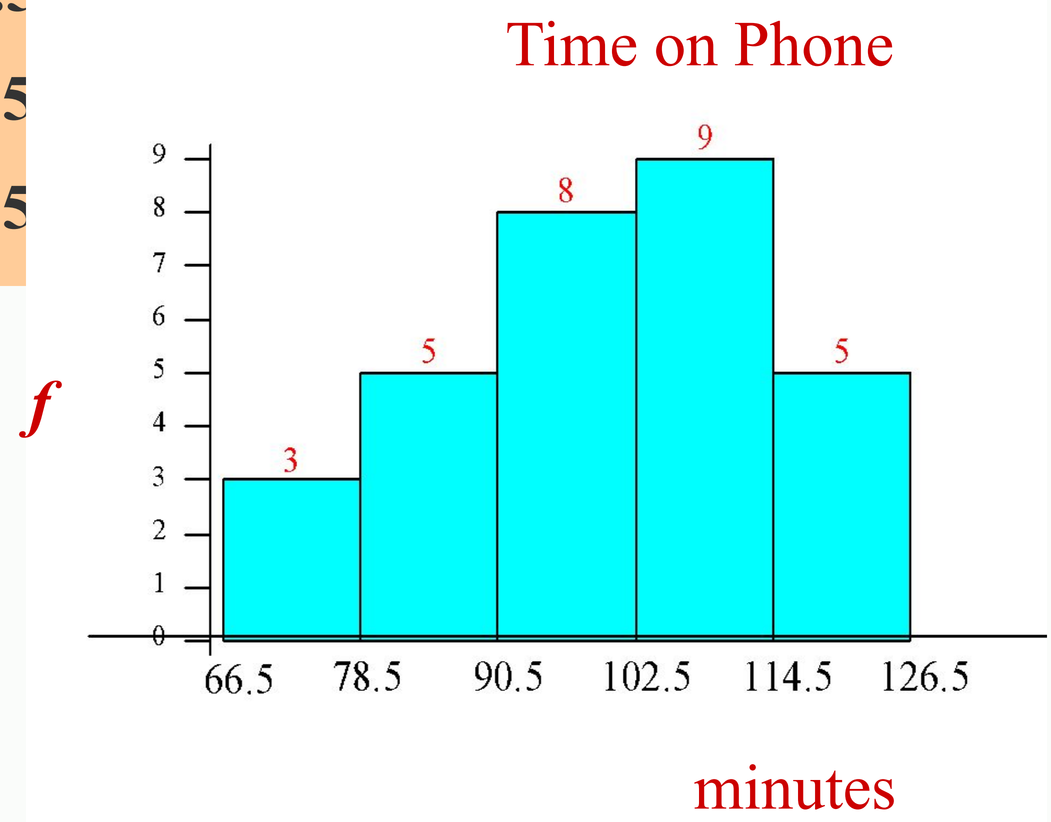
Relative frequency: class frequency/total frequency

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

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

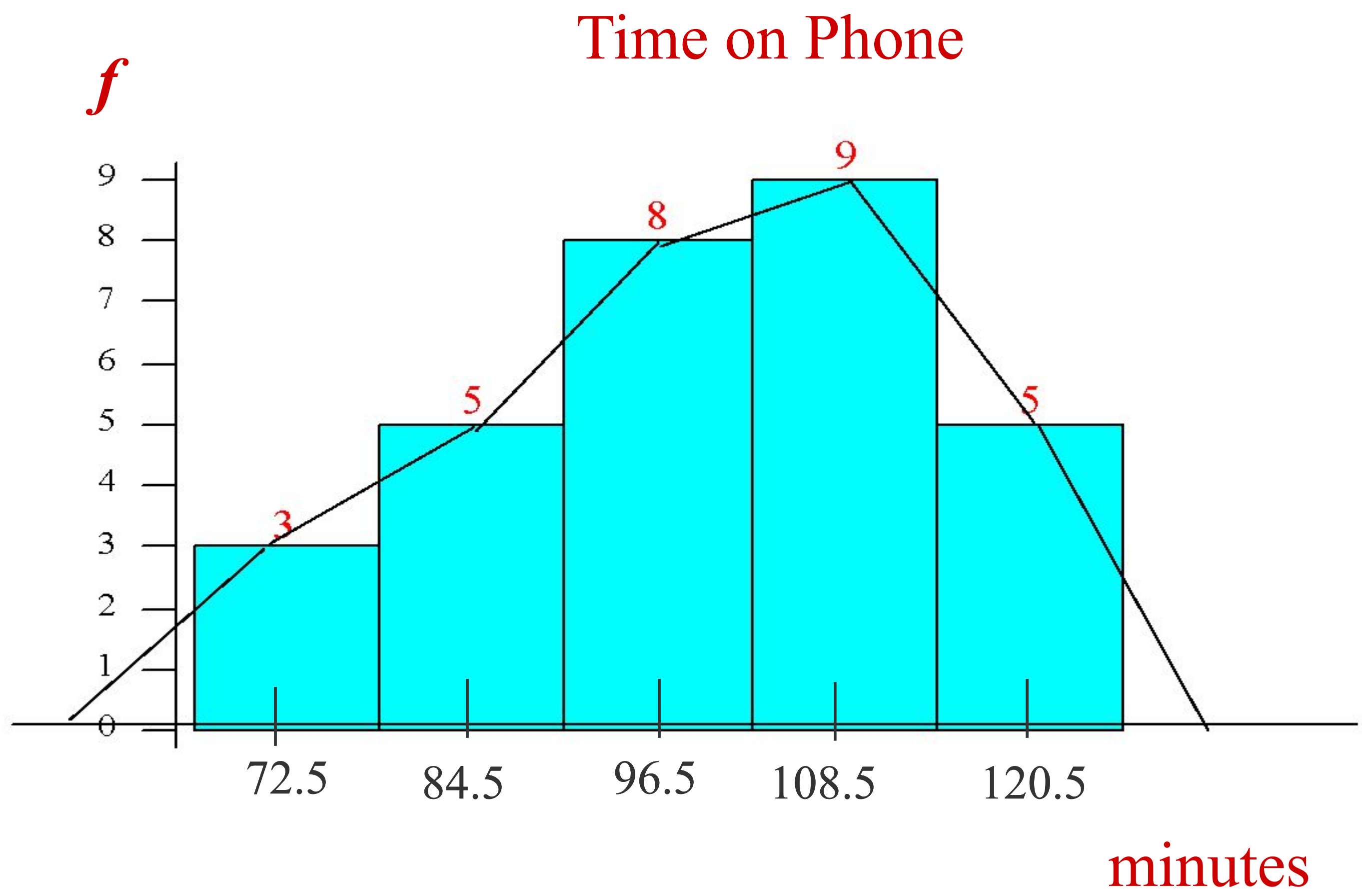
Frequency Histogram

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	115.5 -126.5



Frequency Polygon

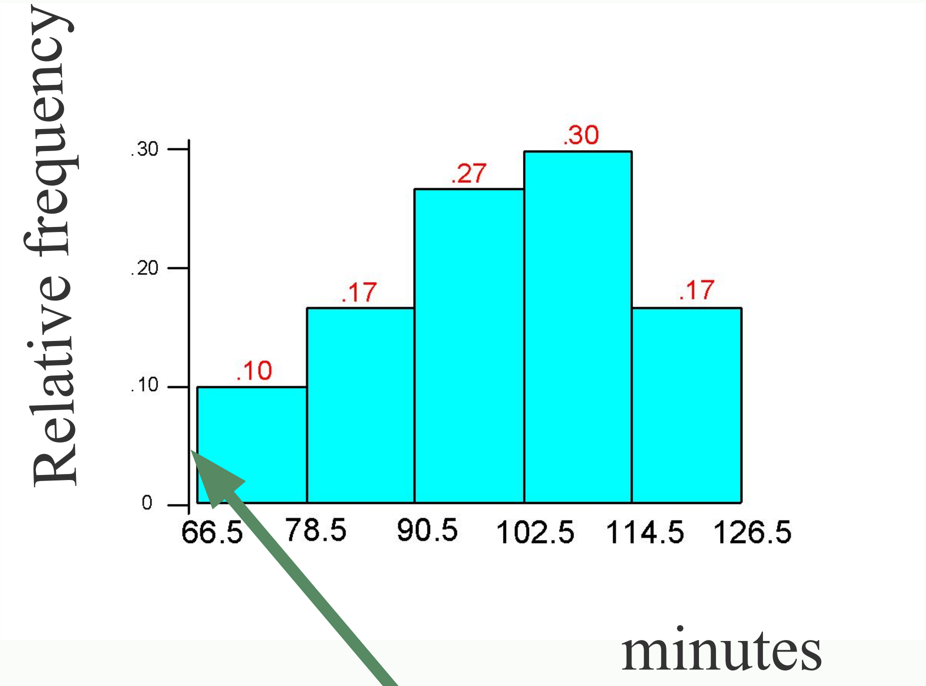
Class	*f*
67 - 78	3
79 - 90	5
91 - 102	8
103 -114	9
115 -126	5



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

Relative Frequency Histogram

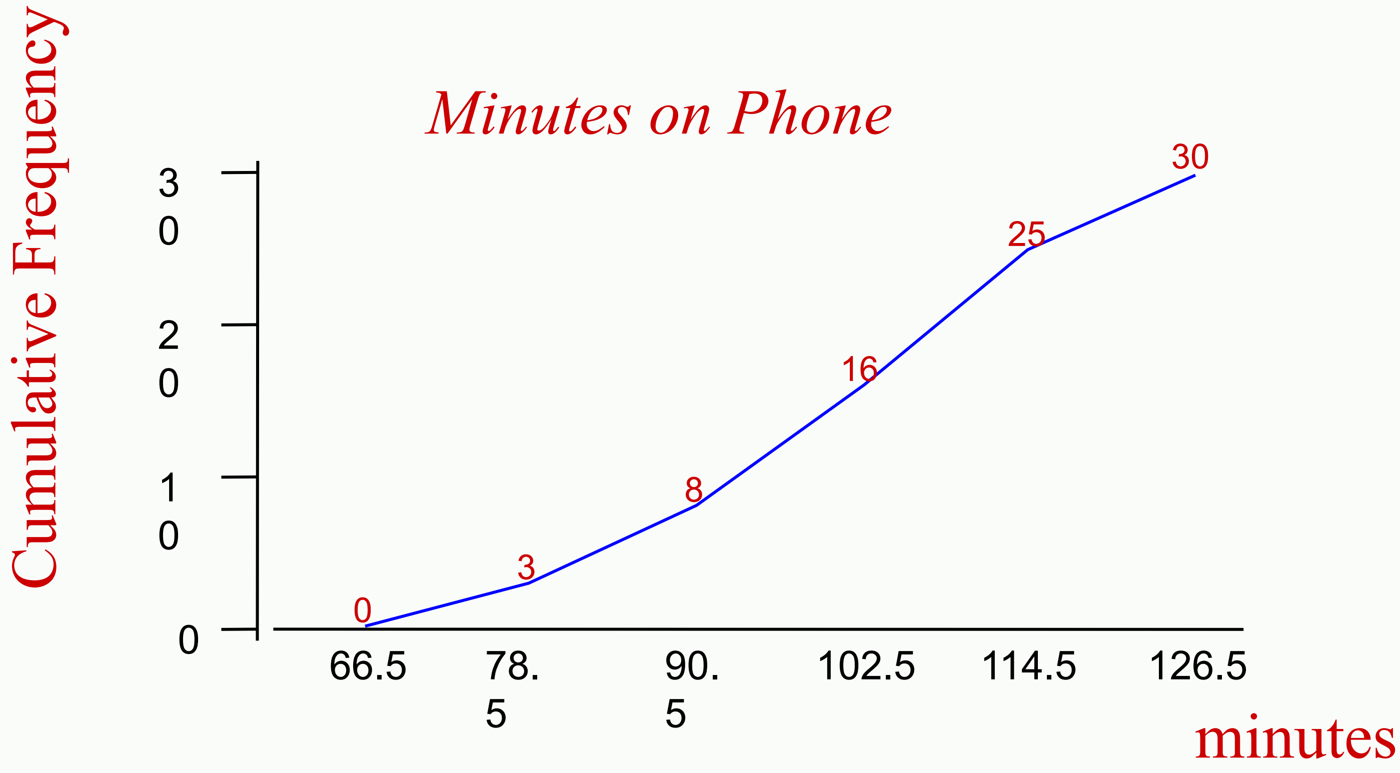
Time on Phone



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*.



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

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 → 8 | 2

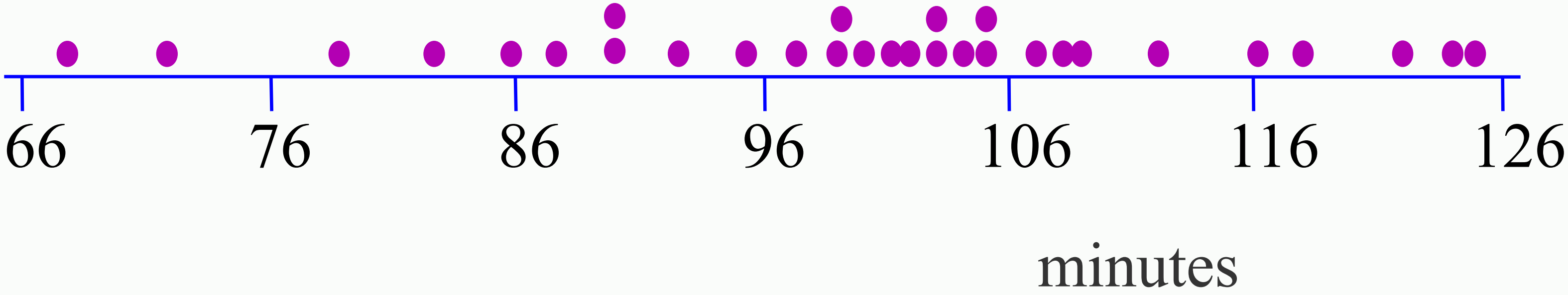
2nd line digits 5 6 7 8 9 → 8 | 5 6 7 7

1st line digits 0 1 2 3 4 → 11 | 2

2nd line digits 5 6 7 8 9 → 11 | 6 8

Dotplot

Phone



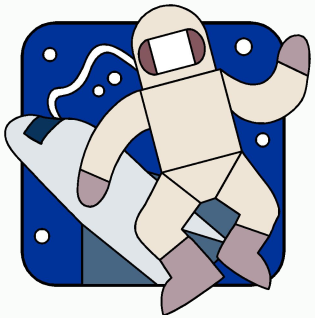
Pie Chart

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

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

The 1995 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.

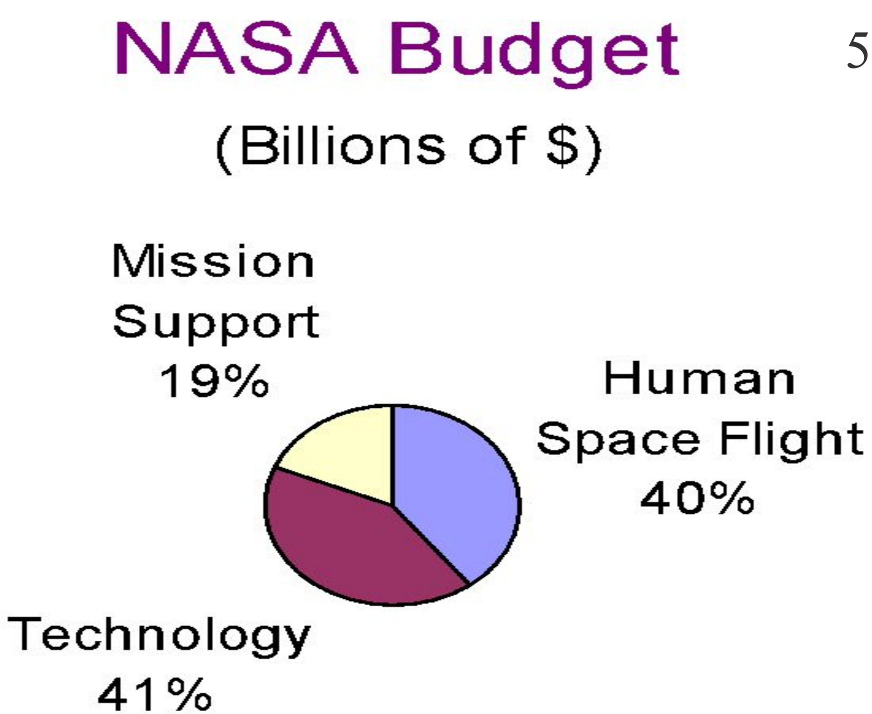
Pie Chart

	Billions of $	Angle(deg.)
Human Space Flight	5.7	143
Technology	5.9	149
Mission Support	2.7	68
	14.3	

Total

$5.7/14.3*360^o = 143^o$

$5.9/14.3*360^o = 149^o$



Measures of Central Tendency

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

For a population:

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

For a sample:

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

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

Mean:

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

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

Mean:

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

Median: Sort data in order

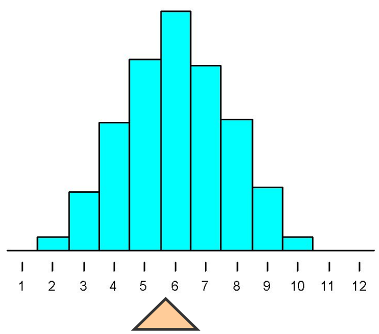
0 2 2 2 3 4 4 6

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

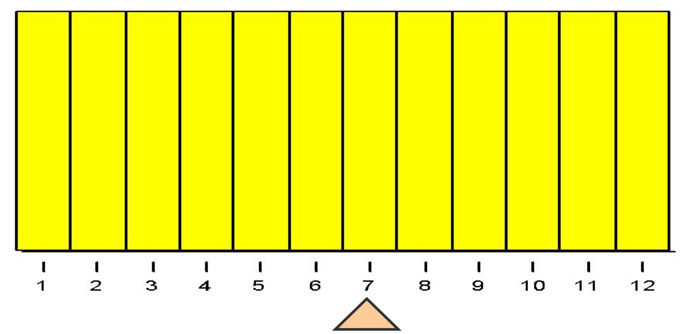
Mode: The mode is 2 since it occurs the most.

Shapes of Distributions

Symmetric

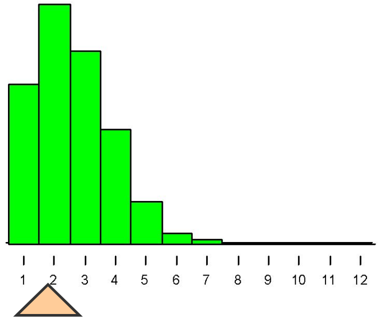


Uniform



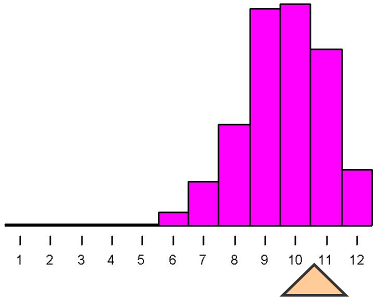
Mean = median

Skewed right



Mean > median

Skewed left



Mean < median

Descriptive Statistics

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 only uses 2 numbers from a data set.

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	-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

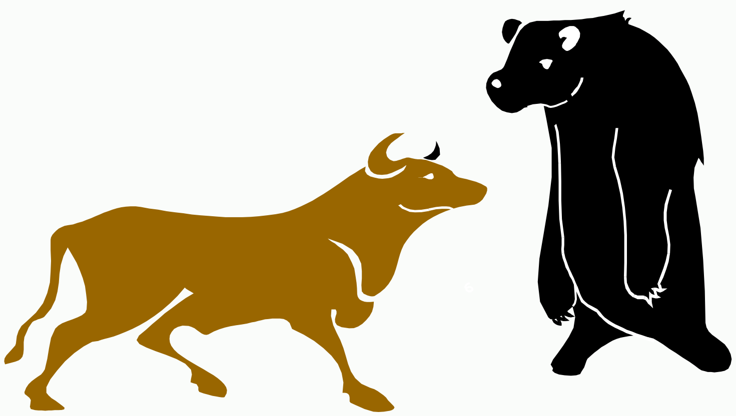
56 - 61.5

56 - 61.5

57 - 61.5

58 - 61.5

$\mu = 61.5$



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

The sum of the deviations is always zero.

$x - \mu$

Population Variance

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

Stock A	$x - \mu$	$(x - \mu)^2$
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

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

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

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 Standard Deviation

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

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

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

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

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

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

Calculate the measures of variation for Stock B

Summary

Range = Maximum value - Minimum value

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

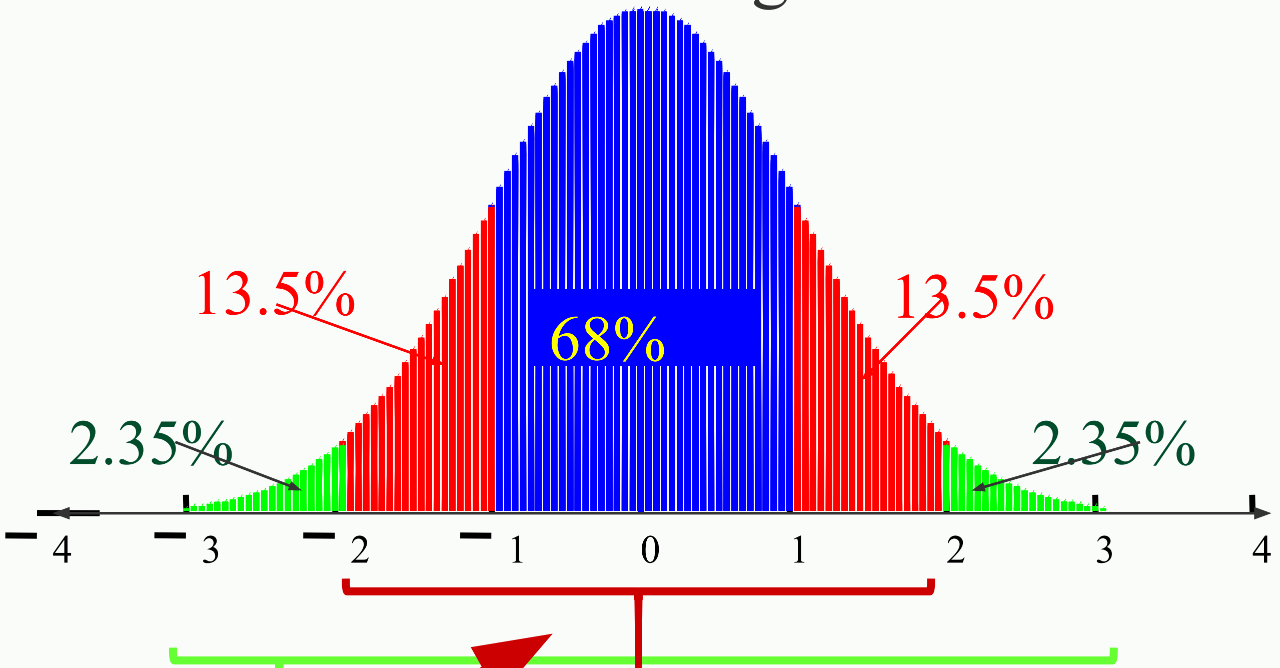
Population Standard Deviation $$\sigma = \sqrt{\sigma^2}$$

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

Sample Standard Deviation $$s = \sqrt{s^2}$$

Empiricl Rule 68- 95- 99.7% rule

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



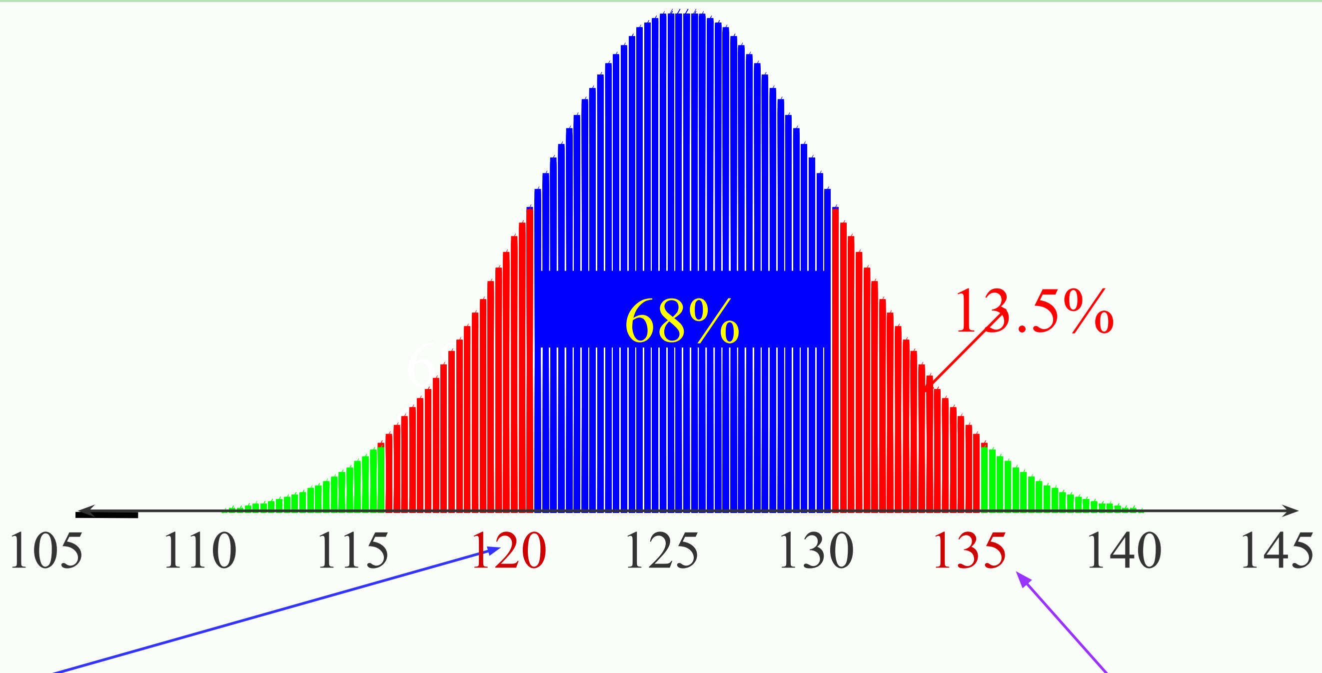
About 68% of the data lies within 1 standard deviation of the mean

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 is 1 standard deviation below the mean and \$135 thousand is 2 standard deviation above the mean.

68% + 13.5% = 81.5%

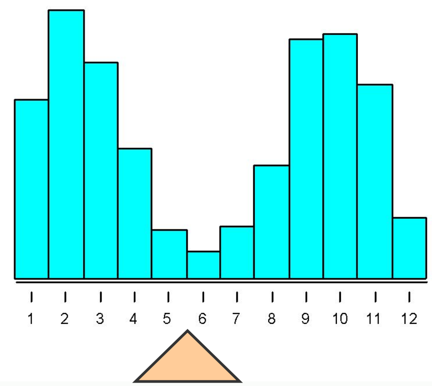
So, 81.5% of the homes have a value between \$120 and \$135 thousand .

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/k²**.

μ =6

σ =3.84



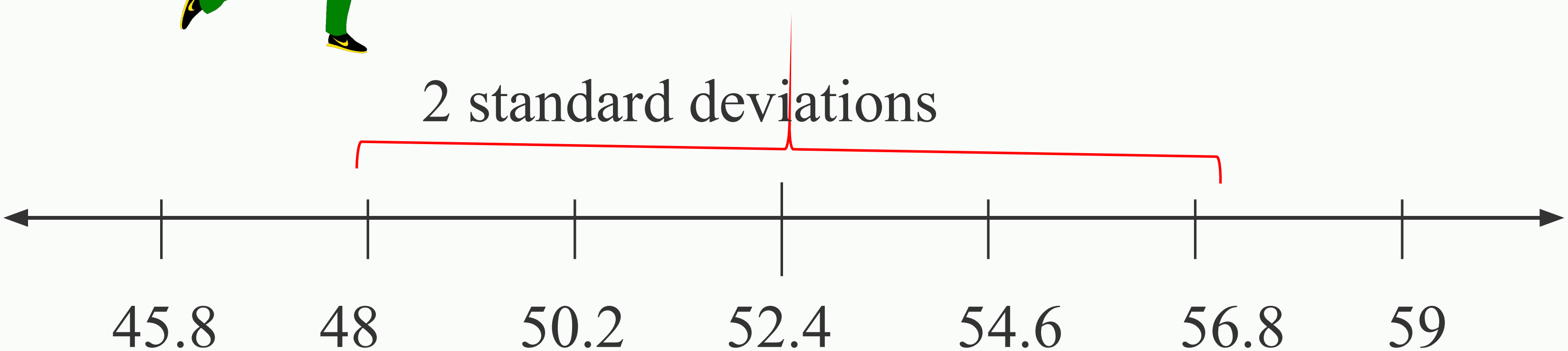
For k = 2, ***at least*** 1-1/4 = 3/4 or 75% of the data lies within 2 standard deviation of the mean.

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.



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

Grouped Data

To approximate the mean of data in a frequency distribution, treat each value as if it occurs at the midpoint of its class. x = Class midpoint.

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

Class	f	Midpoint (x)	x f
67- 78	3	72.5	217.5
79- 90	5	84.5	422.5
91- 102	8	96.5	722.0
103-114	9	108.5	976.5
115-126	5	120.5	602.5
	30		**2991**

$$\bar{x} = \frac{2991}{30} = 99.7$$

Grouped Data

To approximate the standard deviation of data in a frequency distribution, use x = class midpoint.

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

$$\bar{x} = 99.7$$

Class	f	Midpoint	$(x-\bar{x})^2$	$(x-\bar{x})^2 * f$
67- 78	3	72.5	739.84	2219.52
79- 90	5	84.5	231.04	1155.20
91- 102	8	96.5	10.24	81.92
103-114	9	108.5	77.44	696.96
115-126	5	120.5	432.64	2163.2
	30			**6316.8**

$$s = \sqrt{\frac{6316.8}{29}} = \sqrt{217.8207} = 14.76$$

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

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 rank (27 +1)/2 = 14. The median = Q_2 = 42.

There are 13 values below the median.

Q_1 rank= 7. Q_1 is 30.

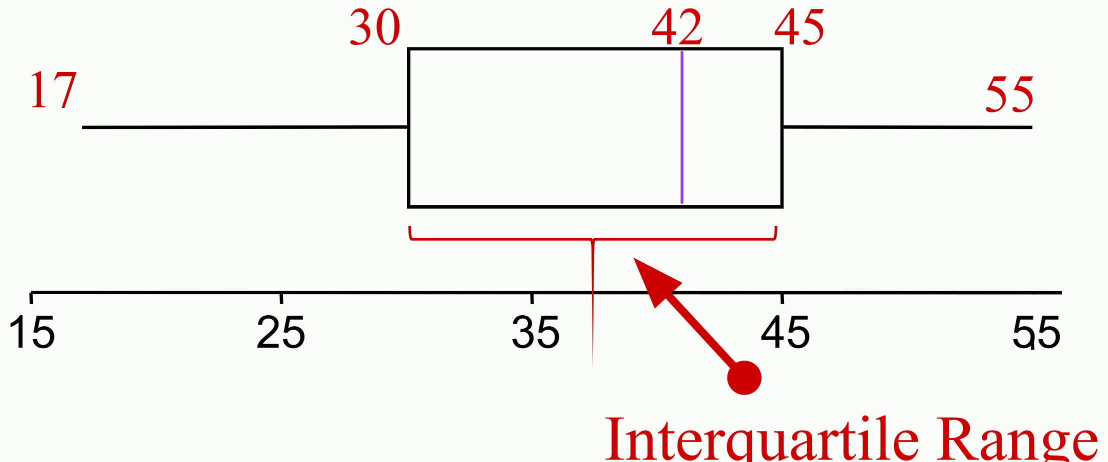
Q_3 is rank 7 counting from the last value. Q_3 is 45.

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

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



Percentiles

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

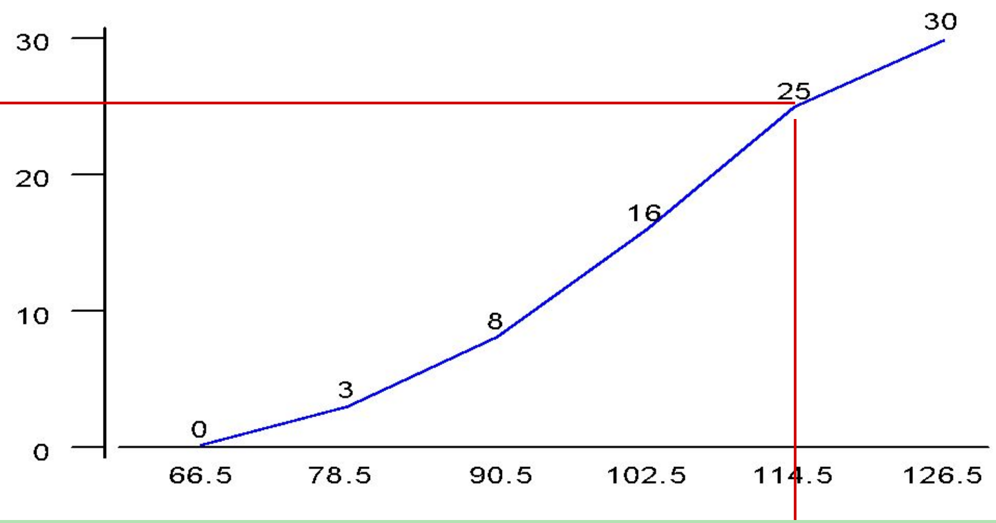
$P_{50} = Q_2 =$ the median

$P_{25} = Q_1$

$P_{75} = Q_3$

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} .