Lesson I Part B

Author Atul Roy

Measures of Center and Spread

Now that we have learned a few ways of displaying and organizing a data, the following examples will help you review the methods of summarizing a data set with a few numbers.

Example 1.

When looking at a transcript the GPA (the grade point average) is a single number that catches the attention of the most of the people.

Recall the method of calculation of GPA.

Suppose that the following table shows the grades of a student during a semester.

Course	Credits	Grade	Grade Points
Accounting	3	A	4
Comp I	3	D	1
Precalculus	3	B	3
Programming	3	B	3
Psychology	3	B	3

Since all the courses are equally weighted, the GPA is $\frac{4+1+3+3+3}{5} = 2.8$

This measure of center is called Mean.

To calculate the mean of a set of numerical data, add all the observations and divide the sum by the total number of observations.

That is if the data is $x_1, x_2, \ldots, x_n$ **, the mean is** $\frac{x_1 + x_2 + \ldots + x_n}{n}$ **denoted by** $\overline{x}$ **if the data is a sample.**

Typically, mean of a population is denoted by μ

Note that

a) The mean takes into account each observation in the data

but

b) The mean is affected by extreme observations as is visible in the above example.

Example 2.

Another measure to express the central value is called Median.

The median is the middle entry when the data is arranged in numerical order. For example for the data

Course	Credits	Grade	Grade Points
Accounting	3	A	4
Comp I	3	D	1
Precalculus	3	B	3
Programming	3	B	3
Psychology	3	B	3

If we arrange the grade points in order,
1 3 3 3 4
Median is 3.

If the number of observations in a data is even, then the mean of the middle two entries when the data is in order, for example, if the data set is

4, 7, 11, 15, 19, 21,29, 43

the median is $\frac{15+19}{2} = 17$

The home prices are often reported in terms of the median, for example

http://www.mc-mncppc.org/research/data_library/real_estate_development/housing/hc14l.shtm

Median Housing Sales Price by Planning Area, 2002 Montgomery County, MD					
Planning Area	New Single-Family Detached	Existing Single-Family Detached	New Townhouse	Existing Townhouse	All Single-Family Types
Aspen Hill	NA	$2,267,000	$284,052	$205,000	$249,900
Bennett	NA	$315,000	NA	NA	$330,000
Bethesda	$928,747	$565,000	NA	$382,640	$565,000
Clarksburg	$444,134	$410,000	$281,460	NA	$373,818
Cloverly	$730,915	$325,000	NA	$169,900	$349,000
Damascus	NA	$250,000	NA	$133,500	$225,000
Darnestown	NA	$469,500	NA	$261,500	$418,700
Dickerson	NA	NA	NA	NA	NA
Fairland	$448,465	$277,000	NA	$165,000	$185,000
Gaithersburg City	$511,548	$334,500	$339,765	$198,500	$289,900
Gaithersburg Vicinity	NA	$290,000	$199,558	$159,900	$180,000
Germantown	$405,375	$348,500	$255.53	$177,000	$236,000
Goshen	NA	$327,500	NA	NA	$349,000
Kemp Mill	NA	$265,000	NA	$179,900	$260,000
Kensington/ Wheaton	$452,461	$244,000	$317,086	$194,500	$239,900
Lower Seneca	NA	NA	NA	NA	NA
Martinsburg & Vicinity	NA	NA	NA	NA	NA
North Bethesda	NA	$415,450	NA	$408,000	$417,500
Olney	$665,145	$355,000	NA	$19,380,050	$305,000
Patuxent	$536,343	$379,950	NA	NA	$436,188
Poolesville	NA	$275,000	NA	$118,500	$246,500
Potomac	NA	$710,000	NA	$385,125	$635,000
Rock Creek	$495,240	$365,000	NA	$245,000	$374,208
Rockville	$573,838	$275,000	$327,474	$302,000	$324,075
Silver Spring	NA	$346,500	NA	$263,000	$331,500
Takoma Park*	NA	$255,000	NA	$139,000	$250,000
Travilah	$682,465	$451,000	NA	$250,000	$440,000
White Oak	$525,270	$288,000	$269,325	$204,500	$275,000
County	$481,286	$340,000	$277,978	$185,500	$282,918

Therefore the use of mean and median depends on the context. Please look at many examples in the text.

Some Examples of the Use of Mean and Median:

If a business wanting to start a service company in a residential area may want to look at the median income (for paying ability) of the residents because the mean may be pulled higher by a few high income people.

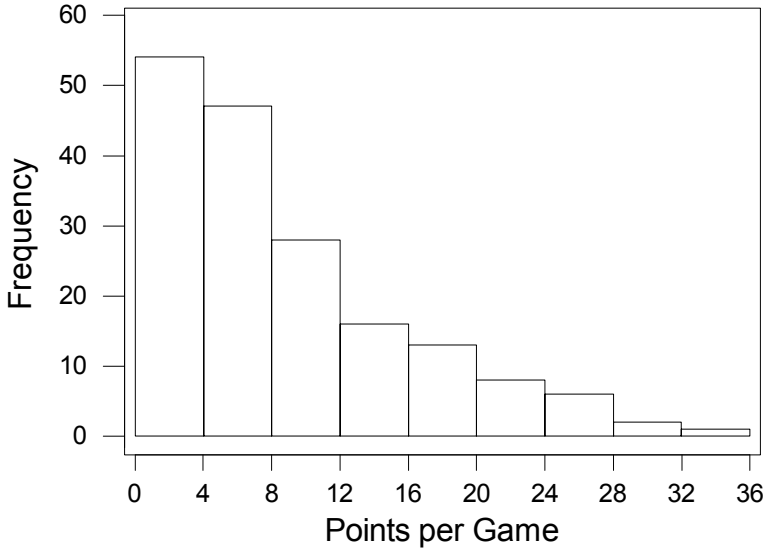
For homebuyers, generally the median price of the home is reported (look at the file median homeprices) but if a summary has to be used for tax revenues it has to be the mean.

For the information regarding the time that it takes the employees to fill out a lengthy but important survey,
median should be considered to see the time limit for the majority of the employees.

Mean should be considered to see how much total time will be lost.

The mean will be pulled towards the extreme values.

As an example recall the example 2 from the Lesson 1, which displayed the points per game from nba.com in the following histogram.

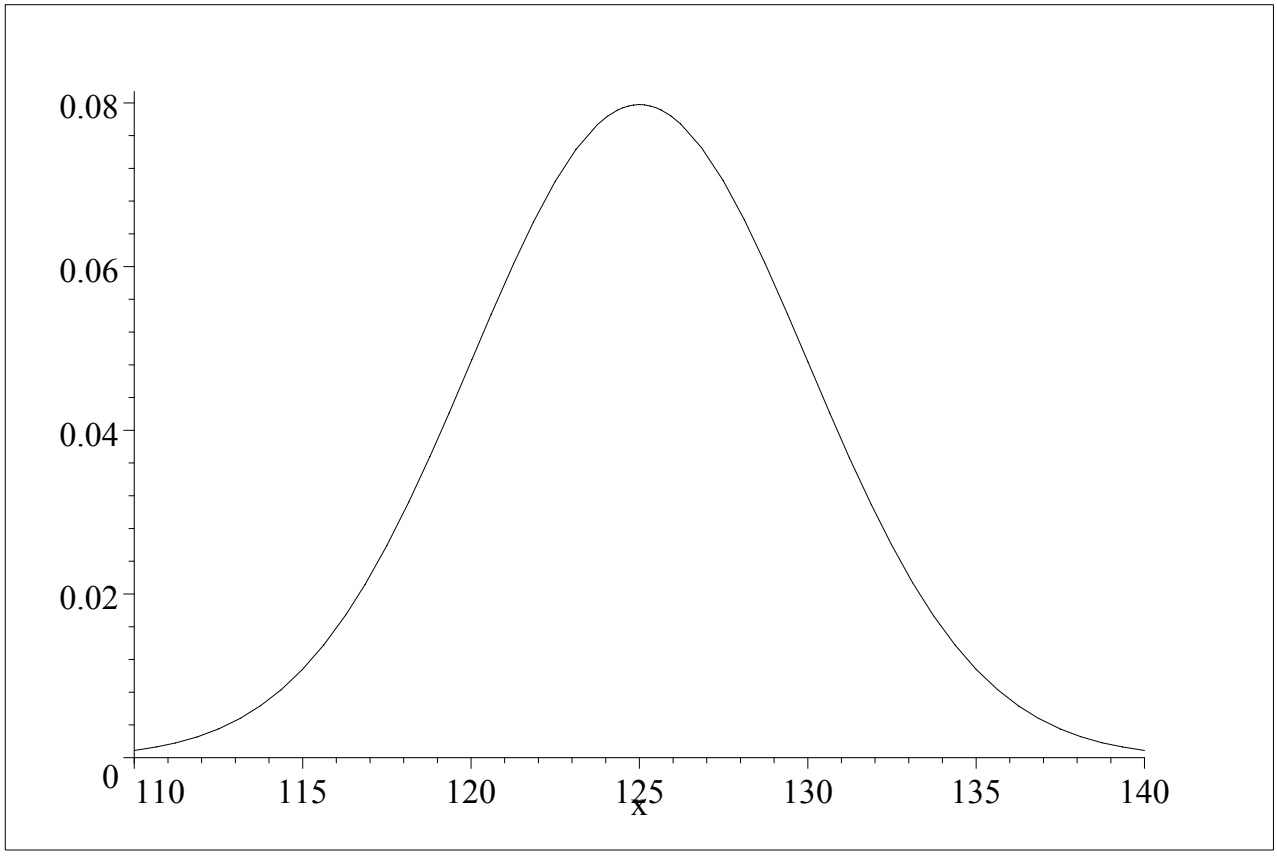


For this data set the median is 6.1 and the mean is 8.801 which is larger than 6.1.

The median must be looked at in the summary of a skewed data set
For a symmetric distribution the mean and median are the same, as shown in the example below.

$$\frac{1}{5\sqrt{2\pi}} e^{-\left((x-125)^2/50\right)}$$

I have written this expression just to obtain the graph, you are not required to work with such expressions.

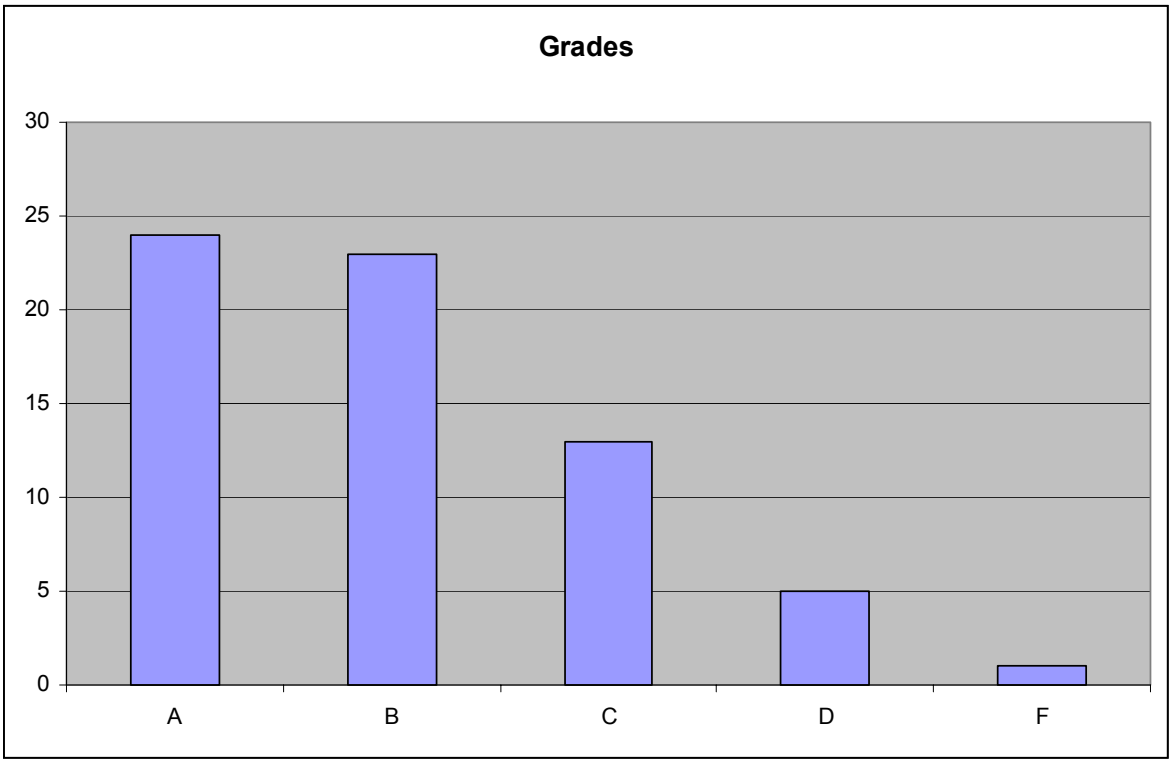


Example 3.

Mode is used for the most frequent value in the data. This is especially useful to express the average of a categorical or qualitative data.

Example:

The following display shows the grade distribution of the students in the STAT 200 courses that I (Atul Roy) have delivered on line till this summer.



The mode is A or the modal grade is A.

On a sad note only 66 of the 100 students who registered finished the courses. My face to face statistics courses have about 95% completion rate. I hope that this restructuring will help more students complete the course.

The above data is 7 years old, I have much better completion rate now, shall share it with you soon.

Measures of Spread

Now that we have seen three measures of center, let us understand some measures of spread dispersion of a numerical data set.

One measure that we shall learn is called standard deviation and gives a measure of spread around the mean.

The standard deviation is calculated in the following manner.

recall from the text that the formula for the standard deviation of a sample is

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

Suppose a sample data set is

8, 11, 12, 15, 17, 21

note that $\overline{x} = \frac{8 + 11 + 12 + 15 + 17 + 21}{6} = 14$

x	$(x - \overline{x})$	$(x - \overline{x})^2$
8	$8 - 14 = -6$	$(-6)^2 = 36$
11	$11 - 14 = -3$	$(-3)^2 = 9$
12	$12 - 14 = -2$	$(-2)^2 = 4$
15	$15 - 14 = 1$	$1^2 = 1$
17	$17 - 14 = 3$	$3^2 = 9$
21	$21 - 14 = 7$	$7^2 = 49$

$\sum (x - \overline{x})^2 = 36 + 9 + 4 + 1 + 9 + 49 = 108$

$s = \sqrt{\frac{108}{6 - 1}} = 4.6476$

Standard Deviation is a very meanigful measure of spread for a normal distribution.

Note that just like mean, even though the standard deviation takes into account each value in the data, it is affected by extreme values in the data, as shown in the following examples.

In the following list, note the way that the value of standard deviation changes as only one number is replaced by a much smaller or much larger value.

						st dev
81	79	87	90	87	85	4.119061
						st dev
31	79	87	90	87	85	22.58982
						st dev
81	79	87	90	87	156	29.35757

Emprical Rule

If the data shows approximately a normal distribution with mean μ and standard deviation σ then

Approximately 68% of the the observations in the data are within one standard deviation of the mean,
that is between $\mu - \sigma$ and $\mu + \sigma$.

Approximately 95% of the the observations in the data are within TWO standard deviations of the mean, that is between $\mu - 2\sigma$ and $\mu + 2\sigma$.

Approximately 99.7% (or practically all of the data) of the the observations in the data are
within THREE standard deviations of the mean, that is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

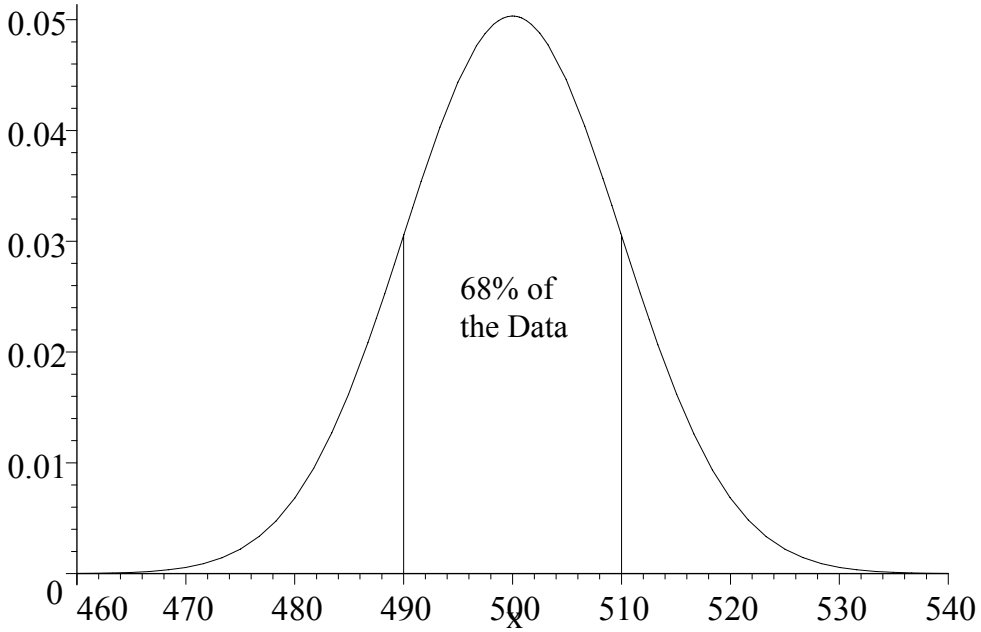
Example:

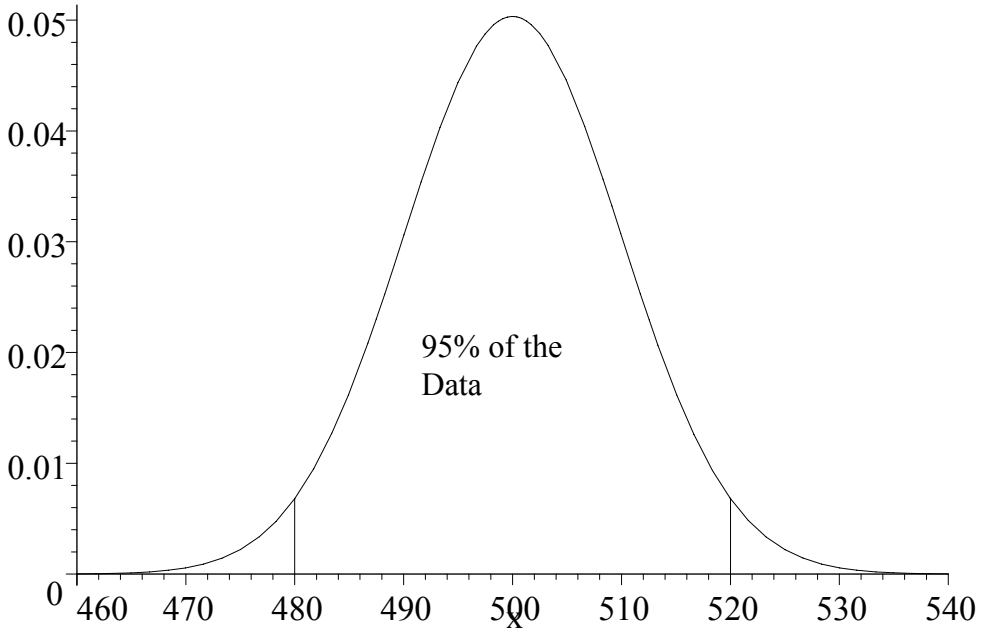
Assume that the contents of a sparkling water bottle show a normal distribution with mean
500 ml and standard deviation 10 ml.

According to the Empirical Rule

About 68% of the bottles will have contents between $500 - 10 = 490$ ml and $500 + 10 = 510$ ml.
About 95% of the bottles will have contents between $500 - 2 \times 10 = 480$ ml and $500 + 2 \times 10 = 520$ ml.
About 99.7% of the bottles will have contents between $500 - 3 \times 10 = 470$ ml and $500 + 3 \times 10 = 530$ ml.

Or graphically





For a non-normal distribution, often a five number summary of

the data

consisting of

Lowest Observation, First Quartile, the Median, Third Quartile, Highest Observation are used.

The quartiles are explained below and the five number summary will be revisited.

Quartiles:

The first quartile Q_1: is the the value for a data set so that 25% of the values of data are below Q_1. In other words it is the median of the lower half of the data.

Suppose the following data shows the scores of students on the first exam in a calculus class consisting of 11 students.

54 67 75 77 81 84 85 85 88 92 96

↑ ↑

Q_1 Median

The third quartile Q_3: is the value for a data set so that 75% of the values of data is below Q_3. In other words it is the median of the upper half of the data.

that is

54 67 75 77 81 84 85 85 88 92 96

↑ ↑ ↑

Q_1 Median Q_3

Sometimes you have to take the mean of two values to determine quartiles, as shown in the following example,

The following table shows average monthly temperatures for San Diego, California

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Â°C	8.7	9.3	10	12.1	14.9	18.6	22.7	22.8	20.1	16	11.3	8.6	14.6
Â°F	47.7	48.7	50	53.8	58.8	65.5	72.9	73	68.2	60.8	52.3	47.5	58.3

If we would like to determine the quartiles of these values in °F, we shall arrange the values in order

47.5 47.7 48.7 50 52.3 53.8 58.8 60.8 65.5 68.2 72.9 73

↑ ↑ ↑

Q_1 median Q_3

therefore

$$Q_1 = \frac{48.7 + 50}{2} = 49.35°\text{F}$$

$$Q_3 = \frac{65.5 + 68.2}{2} = 66.85°\text{F}$$

The Five Number Summary of a data set

Lowest Value, Q_1, Median, Q_3, Highest Value

This is a very useful description for a skewed data set.

Inter Quartile Range= $Q_3 - Q_1$

The Inter Quartile Range gives the range of the middle half of the data.

Example:

Let us look at the five number summary and a display of that Five Number Summary of the infant mortality data as shown below. The data shows the infant mortality rate per 1000 live birth in the year 1987 for 24 countries with the 1999 population of 10 million or more people.

```
IMR
    4.1    5.1    5.1    5.1    5.5    5.6    5.8    6.0    6.2    6.3
    6.3    6.4    6.7    7.1    7.6    7.8    9.5   10.0   12.8   14.4
   16.1   16.5   18.1   18.4
```

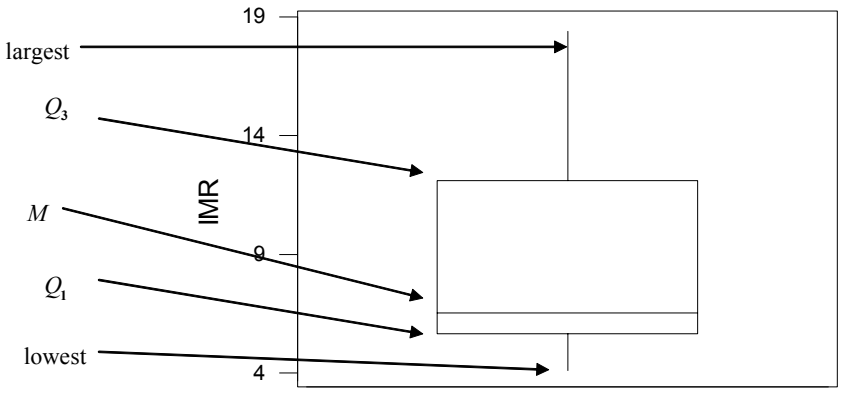
I leave for you to verify that

Lowest=4.1, $Q_1 = 5.65$**, Median=6.55,** $Q_3 = 11.4$**, Highest=18.4**

(note that some packages have slightly different ways of calculating the quartiles.)

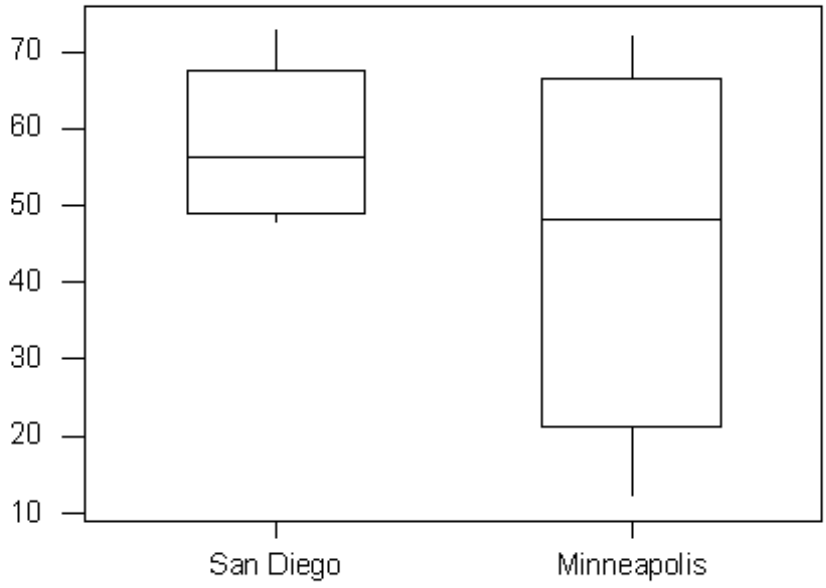
This five number summary can be displayed in a boxplot as shown below. Note the data ranging from Q_1 to Q_3 is enclosed in the box.

Such a plot is also called a box and whisker plot.



You may note that the boxplot shows that the distribution is skewed to the right (or upper side.) Look at the list of the countries covered by this data and relate that to the shape of this boxplot.

Boxplots are very useful for the comparison of distributions. The following display shows the side by side boxplots of the average monthly temperatures in San Diego and Minneapollis.



A numerical method to designate potential outliers.

Remember that IQR = $Q_3 - Q_1$

If the value of an observation is less than $Q_1 - 1.5 \times (\text{IQR})$
or greater than $Q3 - 1.5 \times (\text{IQR})$**, designate that observation as an outlier.**

As an example, let us recall the example 3 from the lesson 1 regarding the points per game from the NBA playoffs of the year 2003.

```
Points Per Game
    0.0    0.3    0.4    0.5    0.5    0.5    0.6    0.7    0.8    0.8
    0.9    0.9    1.0    1.0    1.0    1.0    1.0    1.2    1.3    1.5
    1.8    1.8    1.8    1.9    2.0    2.1    2.2    2.3    2.3    2.5
    2.5    2.5    2.6    2.6    2.7    2.8    2.8    2.8    2.9    3.0
    3.0    3.0    3.0    3.0    3.1    3.2    3.3    3.3    3.5    3.6
    3.7    3.8    3.8    3.8    4.0    4.0    4.0    4.0    4.2    4.3
    4.3    4.4    4.5    4.6    4.7    4.7    4.8    4.8    4.9    5.0
    5.0    5.0    5.2    5.2    5.3    5.3    5.5    5.6    5.7    5.7
    5.7    5.8    5.8    5.9    6.0    6.0    6.1    6.1    6.5    6.5
    6.6    6.7    6.9    7.0    7.0    7.2    7.5    7.8    7.8    7.8
    7.8    8.0    8.0    8.3    8.5    8.7    8.9    9.1    9.2    9.2
    9.3    9.3    9.4    9.4    9.4    9.6    9.7    9.9   10.0   10.0
   10.2   10.4   10.8   11.2   11.3   11.4   11.5   11.5   11.6   12.7
   12.7   12.8   12.8   13.2   13.6   13.9   14.1   14.2   14.3   14.5
   14.7   14.8   14.8   15.3   15.8   16.1   17.2   17.3   17.4   17.8
   18.0   18.3   18.5   18.9   19.0   19.0   19.5   19.6   20.1   20.4
   22.0   22.5   22.8   23.1   23.5   23.7   24.7   24.8   25.3   27.0
   27.0   27.1   31.7   31.7   32.1
```

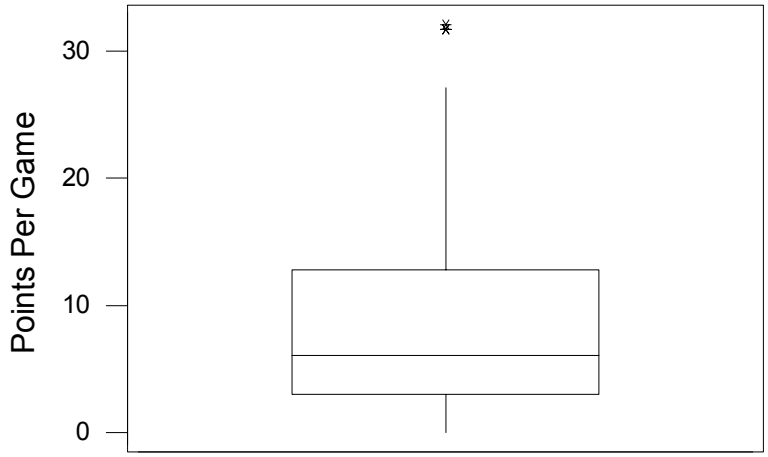
For this data note that the five number summary is

Lowest=0.000 Q_1 **=3.000 M=6.100** Q_3 **=12.800 Highest=32.1**

$Q_3 - Q_1 = 12.8 - 3 = 9.8$ **IQR**

$3 - 1.5 \times 9.8 = -11.7$ **of course there is nothing lower than this.**
$\mathbf{12.8} + \mathbf{1.5} \times \mathbf{9.8} = \mathbf{27.5}$

This summary is displayed by boxplot, the upper whisker is stopped at 27.5 and the values above that shown with astericks.



.....

Chebyshev's Theorem:

The Chebycheff's Theorem says
For a distribution with mean μ **and standard deviation** σ
at least $\left(1 - \frac{1}{k^2}\right)100$ **percent of the data is between** $\mu - k\sigma$ **and** $\mu + k\sigma.$

AN APPLICATION:

Suppose that the past data of the length of a meeting shows $\mu = 90$ **minutes and** $\sigma = 10$ **minutes.**

and we would like to examine a statement "most of the meetings went over two hours."

Two hours = 120 minutes

$120 - 90 = 3 \times 10$ **0r** $3 \times \sigma$

According to Chebycheff,

at least $\left(1 - \frac{1}{3^2}\right)100 = 88.888888\,9\%$ **of the meetings lasted between** $90 - 3 \times 10 = 60$ **minutes and** $90 + 3 \times 10 = 120$ **minutes.**

Therefore it is not possible that more than 50% went over 2 hours.