COLLECT AND COMPARE DATA

This unit is about data gathering methods and making judgments based on various types of samples. To evaluate and improve the data collection process, the types of data and methods used must be carefully considered. When processing data, the measures of central tendency will be examined.

Surveys and Samples

Central Tendency of Data

Comparing Data

Measures of Variation

Surveys and Samples

When trying to find important information from or about a large group of people, as often done by marketing firms or political groups, samples of information are selected to see trends or opinions of the larger group. The different ways to select these groups can be controversial; therefore, it is important to know how these groups are formed and how the selection of data differs.

A convenience sample is a selection of people who are available at the moment. This is the quickest and most cost effective sample, but the least reliable.

A survey response is a pre-selected group of people who volunteer the data themselves.

A random sample is a sample where each item or person in the population is as likely to be chosen as any other. This type holds the best chance for the sample to be unbiased. Each member of the population has the same chance of being selected for the sample.

A representative sample is a random sample selected from subgroups or different categories within the population itself.

If you want to ask a general question to a large group of people to find their opinion or to decide a certain issue that would affect the whole group (something political perhaps), then you need a plan to get a fair representative response for the whole group. If the large group is too large to ask every member, then you can choose a sample of the group. You would only canvass this small group and use their responses to represent the whole large group.

Now let us consider the actual size in numbers. If the large group were 10,000 people, we might sample 10 people, 100 people, or 1000 people. Think about doing just that for each number. Ten would be too few, while 1000 would be too large. One hundred seems reasonable. Think of ways this survey can be done fairly and think about the job description of a person who does this type of work for a living. To find interesting background information about the big business of sampling, use the Internet to search for different types of opinion polls and the companies that collect them. (Zogby, Gallup, and US Census are examples.)

Central Tendency of Data

There are several ways to describe the central tendency of a group of data. We will discuss three: mean, median, and mode. Let's use the example of a typical classroom in high school and describe the age of the occupants in the room.

Example: Listed below are the ages of the students in the classroom along with the teacher's age.

16, 16, 15, 15, 15, 16, 15, 14, 18, 47

Put the data in order from least to greatest to analyze it.

14, 15, 15, 15, 15, 16, 16, 16, 18, 47

Mean is the average of the data.

$$\text{Mean}: \quad \frac{14+15+15+15+15+16+16+16+18+47}{10} = 18.7$$

Mean = 18.7 years

Median is the middle number of the data when it is in order from least to greatest or vice versa.

*When there is an odd number of numbers in the data (listed in order), the middle number is the median. If there is an even number of numbers, the median is the average of the two middle numbers of the data.

14, 15, 15, 15, 15, 16, 16, 16, 18, 47

$$\text{Median}: \quad \frac{15+16}{2} = 15.5$$

Median = 15.5 years

*In this example, there is an even number of numbers in the data set, so the median is the average of the two middle numbers.

Mode is the number that occurs most often in the data.

14, 15, 15, 15, 15, 16, 16, 16, 18, 47

Mode = 15 years

*The number that occurs most often (four times) in this data set is 15.

Now let's see what happens when the teacher's age is removed.

14, 15, 15, 15, 15, 16, 16, 16, 18, ~~47~~

$$\text{Mean}: \frac{14+15+15+15+15+16+16+16+18}{9} \approx 15.6$$

*15.6 is rounded to the nearest tenth.

Median: 14, 15, 15, 15, 15, 16, 16, 16, 18 = 15

Mode: 14, 15, 15, 15, 15, 16, 16, 16, 18, 47 = 15

Mean = 15.6 years Median = 15 years Mode = 15 years

View the chart below and compare how the mean, median, and mode were affected by removing the "large data point" that was the teacher's age.

Measure of Central Tendency	With Teacher's Age	Without Teacher's Age
Mean	18.7	15.6
Median	15.5	15
Mode	15	15

How was the mean affected? The mean was affected greatly. After the removal of the teacher's age, the mean became more representative of the entire group of data. The mean dropped from 18.7 to 15.6.

How was the median affected? The median was slightly affected and became a little lower after the teacher's age was removed. The median dropped from 15.5 to 15.

How was the mode affected? There was no change in the modes.

*Notice that the median and the mode were less affected by a single large (or small) data point in comparison to the mean. When reporting data, sometimes the median is given rather than the mean because it is more representative of the whole group of data.

Be careful when making decisions about reporting the results of calculated measures of central tendency. Choose the measure that demonstrates an **authentic center** of the data.

Comparing Data

When comparing two sets of data, we can often notice something that is different with the two lists even though the data we are comparing will have similar measures of central tendency (averages).

Example: Examine the test scores of Bill and Dale, and then answer the questions below.

Bill's Scores:	55%	75%	95%
Dale's Scores:	73%	75%	77%

(a) Who is scoring average grades?

First, look at the mean.

Bill's average (mean): $\frac{55+75+95}{3}=75$

Dale's average (mean): $\frac{73+75+77}{3}=75$

Bill's mean score is 75%...so is Dale's.

Now look at the median.

Bill's Scores:	55%	75%	95%
Dale's Scores:	73%	75%	77%

Bill's median score is 75%...so is Dale's.

(b) Can the difference in the students' scores be described some other way?

To see a difference, consider the spread.

How far are Bill's scores spread from the lowest score to the highest score?

$95 - 55 = 40$ 40 points

How far are Dale's scores spread from the lowest score to the highest score?

$77 - 73 = 4$ 4 points

(c) So, which student has performed "more average"?

Dale has shown more consistency in scoring average grades; thus, we can conclude that he has performed "more average" overall.

This is just one way to try to report and interpret data with some consistency.

Measures of Variation

The study of **quartiles** is another way to help learn about the nature and tendency of data.

Quartiles divide data that is arranged in order from least to greatest into four equal parts.

The ***median***, sometimes referred to as the Second Quartile, separates the data in half.

The ***Lower Quartile (LQ),*** sometimes referred to as the First Quartile, is the median of the first half of the data.

The ***Upper Quartile (UQ)***, sometimes referred to as the Third Quartile, is the median of the second half of the data.

The ***range*** of the data is the difference between the highest and lowest data values and is determined by subtracting these values.

The ***Interquartile Range*** is the range of the middle half of the data and is determined by subtracting the lower quartile from the upper quartile **(UQ – LQ).**

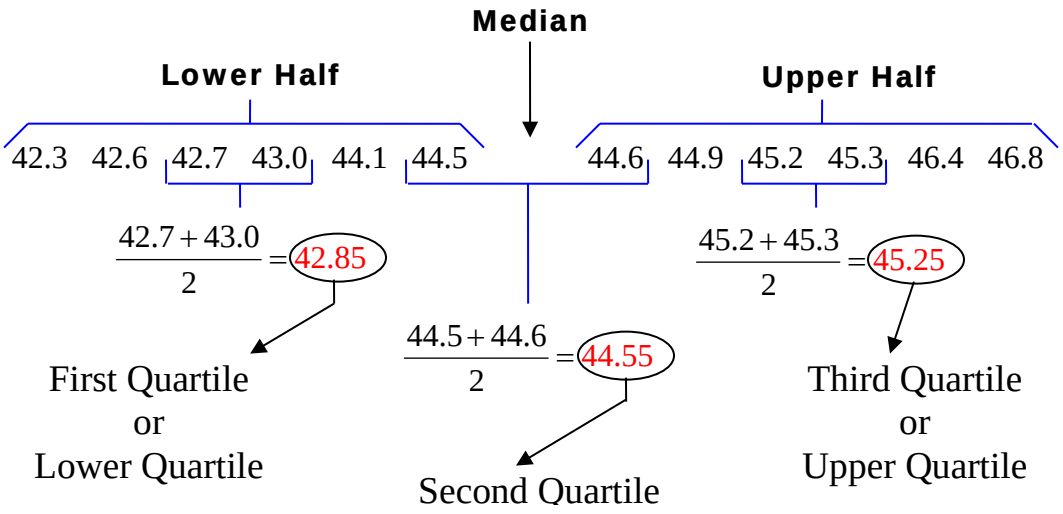
Many companies analyze data using "measures of variation" to determine the promotion and implementation of their product.

Listed below are several examples where the "measures of variation" are calculated. The data is given in order from least to greatest.

*Note: The first step in determining quartiles is to put the **data in order** from least to greatest.

Example 1: Calculate the measures of variation for the following set of data.

42.3 42.6 42.7 43.0 44.1 44.5 44.6 44.9 45.2 45.3 46.4 46.8



Range = Highest Value – Lowest Value = 46.8 – 42.3 = 4.5
Interquartile Range = UQ – LQ = 45.25 – 42.85 = 2.4

In this example there is an even number of data points in the data (12); thus, the **Second Quartile (median)** is the average of the two middle numbers. (44.55)

There is also an even number of data points in the lower half of the data (6); thus, the **First Quartile** is the average of the two middle numbers in the lower half. (42.85)

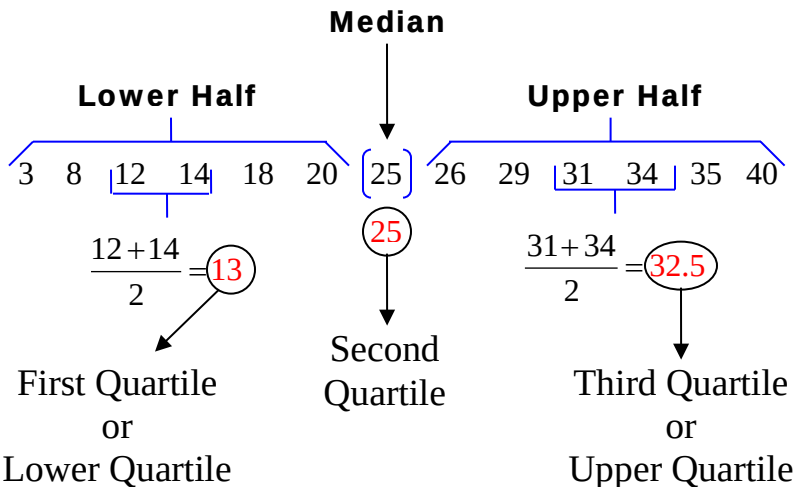
There is also an even number of data points in the upper half of the data (6); thus, the **Third Quartile** is the average of the two middle numbers in the upper half. (45.25)

The **range** is the difference between the highest and lowest data points. (4.5)

The **Interquartile Range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (2.4)

Example 2: Calculate the measures of variation for the following set of data.

3 8 12 14 18 20 25 26 29 31 34 35 40



Range = Highest Value – Lowest Value = 40– 3 = 37
Interquartile Range = UQ – LQ = 32.5 – 13 = 19.5

In this example there is an odd number of data points (13); thus, the **Second Quartile (median)** is the middle number. (25)

There is an even number of data points in the lower half of the data (6); thus, the **First Quartile** is the average of the two middle numbers in the lower half. (13)

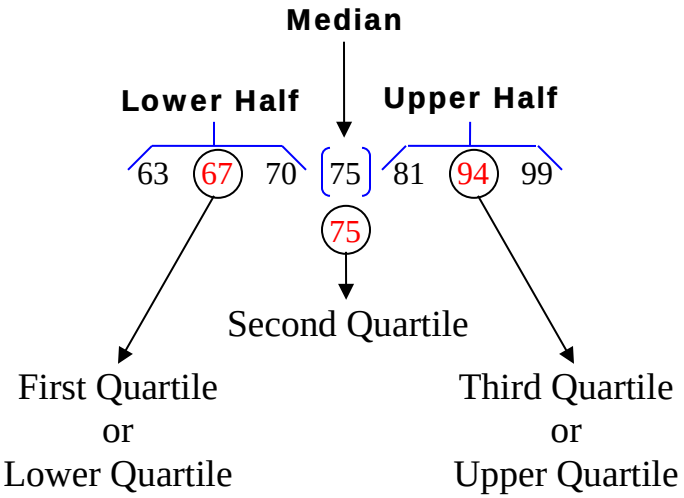
There is also an even number of data points in the upper half of the data (6); thus, the **Third Quartile** is the average of the two middle numbers in the upper half. (32.5)

The **range** is the difference between the highest and lowest data points. (37)

The **Interquartile Range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (19.5)

Example 3: Calculate the measures of variation for the following set of data.

63 67 70 75 81 94 99



Range = Highest Value – Lowest Value = 99 – 63 = 36
Interquartile Range = UQ – LQ = 94 – 67 = 27

In this example there is an odd number of data points (7) in the data; thus, the **Second Quartile (median)** is the middle number. (75)

There is an odd number of data points (3) in the lower half of the data; thus, the **First Quartile** is the middle number in the lower half. (67)

There is also an odd number of data points (3) in the upper half of the data; thus, the **Third Quartile** is the middle number in the upper half. (94)

The **range** is the difference between the highest and lowest data points. (36)

The **Interquartile Range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (27)