huber-data-3

Welcome to Lesson 3! In this lesson, we will learn what DISPERSION is and what statistical values are needed in order to best describe the spread of data. Further, you will learn all about a box-and-whisker plot which is a plot commonly used by statisticians when determining variability.

While measures of central tendency are used to estimate the middle values of a dataset, measures of dispersion are important for describing the spread of the data.

The term dispersion refers to degree to which the data values are scattered around an average value. Dispersion is synonymous with other words such as variability and spread.

Why is it important to analyze the spread of a particular set of data? Two different samples may have the same mean or median, but different levels of variability or vice versa.

Therefore, it is important to describe both the _____ and _____ of a data set.

1. median, variability
2. central tendency, dispersion
3. middle, mean
4. spread, variability

central tendency, dispersion

In this lesson, we will discuss the three statistical values most commonly used to describe the dispersion or variability of a data set. Variability is a fancy term used to classify how variable or spread out the data is.

The first descriptive statistic that can describe the variability of a data set is known as the RANGE. The range is the difference between the maximum and minimum values of the data set.

To demonstrate how you can use R to determine the range of a data set we will refer back to the cars93 data set from the previous lesson.

Type in the R-command 'range(cars93$price)' to determine the exact values for the minimum and maximum prices of cars in the data set.

```
range(cars93$price)
```

```
## [1]  7.4 61.9
```

The second important measure of variability is known as VARIANCE. Mathematically, VARIANCE is the average of the squared differences from the mean. More simply, variance represents the total distance of the data from the mean.

In R, you can use the command 'var(data)' to easily calculate the variance of a particular set of data. Try calculating the variance for the data 'cars93$price'.

```
var(cars93$price)
```

```
## [1] 132.3984
```

The values for variance and standard deviation are very closely related. The standard deviation can be calculated by taking the square root of the variance where as the variance can be calculated by

squaring the standard deviation.

To statisticians, the standard deviation is a more conventional measure of variability because it is expressed in the same units as the original data values.

Similar to variance, you can use the R-command 'sd(data)' to calculate the standard deviation of a particular set of data. Try calculating the standard deviation for the data 'cars93$price'.

```
sd(cars93$price)
```

```
## [1] 11.50645
```

The standard deviation is very important when analyzing our data set. A small standard deviation indicates that the data points tend to be located near the mean value, while a large standard deviation indicates that the data points are spread further from the mean.

Three important measures of variability are which of the following:

1. mean, median, range
2. spread, mean, central tendency
3. variance, dispersion, spread
4. range, variance, standard deviation

range, variance, standard deviation

A BOX PLOT, also called a BOX-AND-WHISKER PLOT, is used to summarize the main descriptive statistics of a particular data set and this type of plot helps illustrate the concept of variability. A box plot is used to visually represent the MINIMUM, FIRST QUARTILE (Q1), MEDIAN, THIRD QUARTILE (Q3), and MAXIMUM of a data set.

Here I have created a box plot to represent the price data for each of the three car types: large, midsize, and small. You'll notice that each of the 3 figures is composed of a closed 4-sided "box" with "whiskers" on the top and bottom, hence the name box-and-whisker plot.

The height of each box is referred to as the INTERQUARTILE RANGE (IQR). The more variability within the data, the larger the IQR. On the other hand, less variability within the data means a smaller IQR. The bottom of the box in the box plot corresponds to the value of the first quartile (Q1), and the top of the box corresponds to the value of the third quartile (Q3). To calculate the value of the IQR, simply subtract the value of Q1 from that of Q3.

The whiskers, or lines, that extend above and below each box represent roughly the upper 25% and lower 25% of data points, respectively. The only exception is when there are outliers, which I'll explain shortly.

Let's take a closer look at how quartiles are calculated. We start by sorting the data from least to greatest, just like when calculating the median. The first quartile (Q1), also known as the 25th PERCENTILE (since 25% of the data points fall at or below this value), is simply the median of the first half of the sorted data. Likewise, the third quartile (Q3), also known as the 75th percentile, is the median of the second half of the sorted data.

As shown in this plot, the blue horizontal line illustrates how to find the value for the first quartile. The green horizontal line illustrates how to find the value for the third quartile. The interquartile range is the range of data values that is contained in between these two lines.

Look again at our box plot of price vs. car type. You may be thinking to yourself, 'What is that circle above the box plots for the midsize cars, and why is there no circle above the box plot for the large

cars?' Those circles represent OUTLIERS in the data set.

An OUTLIER is an observation that is unusual or extreme relative to the other values in the data set. Outliers are useful in identifying a heavy skew in a distribution, and may signify a data collection or data entry error to a scientist. There are many different conventions for identifying outliers within a data set.

When looking at the box plot, which car types appear to have outliers?

1. small, midsize
2. midsize, large
3. small, large
4. small, midsize, large

small, midsize

As you can see in the box plot, the data for prices of 'midsize' cars vary from around 15 to 62, encompassing a range of approximately 50. This is a great deal larger than the variation for 'small' cars which range from around 5 to 15, encompassing a range of approximately 10. Therefore, since the range is much greater for 'midsize' cars, prices of 'midsize' cars have much higher variability in comparison to the prices of 'small' cars.

Now it is your turn! Is the variability of prices of cars of type 'large' higher or lower than that of cars of type 'small'?

1. higher
2. lower
3. the same

higher

You have officially mastered the concept of dispersion and have fully learned how to read and interpret a box-and-whisker plot. These are two very valuable tools used everyday in descriptive statistics. Congratulations!