Descriptive Statistics

What are descriptive statistics

Descriptive statistics are measures that summarize a set of data. They help to provide quick information about the contents of the data set and can provide some direction for further analysis or indicate potential issues with the data. Measures include:

- Measures of Central Tendency:
    - Mean
    - Median
    - Mode
- Variability:
    - Variance
    - Standard Deviation
- Percentile/Quartile

Mean

The mean looks at all the values in an array and comes up with a representative number. The mean might not be a value in the array

Pros:

- Considers all data points equally

Cons:

- Can be influenced by outliers

Calculation:

Sum(array)/Count(array)

Mean

In the starwars data set:

1. In a cell, calculate the **SUM** of the **height** column
2. In another cell, **COUNT** the number of values in the **height** column
3. In a third cell, divide the **SUM** (cell from #1) by the **COUNT** (cell from #2)

4. In another cell use the **AVERAGE**() function and compare the result to your calculation in #**3**

Median

The median orders all the values in the array and selects the one closest to the middle. The median may be a value in the array.

Pros:

- Not as affected by outliers

Cons:

- May not capture important fluctuations in the data

Calculation:

1. Sort data from lowest to highest
2. Select value in the middle or average of two middle most values

Median

1. Delete the empty row between the column titles and data
2. Sort the **height** from lowest to highest
3. Add a new column to the left of **height** called **height_order**
4. Fill **height_order** with incrementing numbers, starting with 1 at the top
5. Find the midpoint by dividing the **COUNT** by 2 (from #2 in the **mean** calculation)
6. Use **VLOOKUP** to find 1 or 2 middlemost values (if the midpoint calculated in #**5** is a decimal, use whole numbers to find the 2 heights just above and just below that midpoint)
7. If there are 2 middlemost values, find the median by calculating the **AVERAGE** of the 2; if just 1 middle value, that is the median
8. In another cell use the **MEDIAN**() function and compare the result to **#6** or **#7**

Mode

The mode identifies the most frequent value(s) in the array. The mode(s) will be a value in the array

Pros:

- If data range is discrete, can identify peaks and valleys across the distribution

Cons:

- Not always a useful measure (size, misleading, too many/no mode)

Calculation:

1. Count number of occurrences for each unique value

Mode

1. Use **Conditional Formatting** to highlight duplicates in the **height** column

2. Find the value(s) with the highest number of duplicates - (look carefully!)

3. In another cell use the **MODE**() function and compare the result to **#2**

Variance and Standard Deviation

Variance describes generally how far values are from the mean. Standard Deviation is the square root of variance, useful in that it returns a value that is in the same units as the mean.

Calculation:

1. Calculate the **AVERAGE** of the array
2. Subtract each value from the **AVERAGE** to get the difference between each value and the mean
3. Square the differences - you want absolute distance from the mean so values won’t cancel each other out
4. Find the variance by calculating the **AVERAGE** of squared differences
5. Find the standard deviation by calculating **SQRT** of mean squared differences

Variance and Standard Deviation

1. Add a new column to the right of **height** called **diff_mean**
2. Subtract the height from the **MEAN (**remember to use an absolute reference to point to the cell where you calculated the mean**)**, copy the formula down the column
3. Add a new column to the right called **diff_mean_sqd**
4. Square the values from **diff_mean**
5. Find the variance by calculating the **AVERAGE**() for **diff_mean_sqd**
6. In another cell use the **VARP**() function on the **height** column and compare the result to #**5**
7. Find the standard deviation by calculating the **SQRT**() of **#5**
8. In another cell use the **STDEVP**() function on the **height** column and compare the result to **#7**

Percentile/Quartile

A percentile is the value below which a certain percentage of the sorted data fall. Quartiles are specific percentile cutoffs:

- First quartile = 25%
- Second quartile = 50%
- Third quartile = 75%
- Fourth quartile = 100%

Calculation:

1. Sort the data from lowest to highest
2. Select value at position that corresponds to the percentage of the total number of observations (round up). Example: Value at position 80, if looking for 80th percentile in 100 observations.

Percentile/Quartile

1. Find the 90th percentile of height.
    a. Calculate 0.9***COUNT** of observations in the **height column**
    b. Round up (can just do this manually, no need to use a function)
    c. Use **VLOOKUP** to find **height** at the **height_order** from **#2**
2. Find the First Quartile. Repeat steps **1-3** using 0.25, instead of 0.9
3. In another cell use the **PERCENTILE.INC**(**height**, 0.9) function and compare the result to #**3**
4. In another cell use the **QUARTILE.INC**(**height,** 1) function and compare the result to #**4**