Descriptive Statistics

NASHVILLE SOFTWARE SCHOOL

Objective: To identify the types, and uses, of descriptive statistics tools

The meaning and uses of descriptive statistics

- Define descriptive statistics
- Describe how using descriptive statistics is an important part of the data exploration process
- Explain how descriptive statistics can be used to summarize large amounts of data

Descriptive statistics functions available in spreadsheets

- Measures of central tendency
- Measures of variability
- Percentile and Quartile functions

Descriptors of Central Tendency - Mean

SUM(array)/COUNT(array)

Mean – Often generically called the “average”. Among to most easily recognizable measures of central tendency thus able to be used when communicating to people with minimal statistics/analytics experience.

Pros:

- Weighs all elements of an array equally

Cons:

- Can be overly influenced by extreme values and outliers

Descriptors of Central Tendency - Median

The value of the nth item of a sorted array where n = LENGTH(array)/2

Median – Is a measure that describes the center of a sorted group. It is both recognizable and easily described, this makes it a good tool for communicating.

Pros:

- Not as affected by outliers

Cons:

- Data must be sorted
- May not detect meaningful fluctuations within the data

Descriptors of Central Tendency - Mode

The maximum frequency of all values in an array

Median – A measure used to describe the value(s) which occur most frequently. Most likely to match the value of an item chosen at random from an array.

Pros:

- Useful at identifying peak values
- If data is sorted can identify transitions between peaks and valleys of frequency

Cons:

- May not be useful if data is continuous
- May not be useful even if data is continuous

Variance

Variance – a quantitative description of how tightly, or loosely, items of an array are clustered around the mean. Not often used in isolation, used to find the standard deviation

Calculating the variance of an array:

- Calculate the mean of the array
- Subtract each value of the array from the mean to get the difference
- Square each difference
- Find the mean of those squared values

Standard Deviation

Standard Deviation – One of the most useful and widely used tools for statistical analysis. More difficult to explain than the measures of central tendency but often more powerful and informative.

Calculating the standard deviation of an array:

- Calculate the variance of the array
- Take the square root of the variance

Percentile and Quartiles

Percentile – The value in which the given percentage of data falls below when sorted.

Quartiles – Specific, equally-spaced, percentiles that divide the data into 4 equal parts.

To find the percentile:

1) Multiply the count of the array with the desired percentile

2) The product (rounded up) is which item of the sorted array holds the percentile value

3) An exception is made for the 50th percentile which is not rounded up, but instead the mean of the two values it falls between

Exercises

Note: Discard entries for which data is unavailable

1) Find the mean of all the character's heights.

2) Find the median height using a LOOKUP.

3) Use conditional formatting to highlight duplicate heights, use this to determine the mode.

4) Calculate the variance and standard deviation for height.

5) Find the 90th percentile value for height. Do the same for the 1st quartile (25th percentile).

Additional Resources

Step-by-step assistance for the above exercises.

Mean

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

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 meancalculation)

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

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

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

Percentiles

1. Find the 90th percentile of height.
   - a. Calculate 0.9*COUNT of observations in the height column (#2 from mean)
   - b. Round up (can just do this manually, no need to use a function)
   - c. Use a VLOOKUP to find the height at the height_order from b

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