2/4/2023

Descriptive Statistics
Measures of Center, Spread, Location

Measures of Center
You can think of these as "typical" values: if you had to summarize the data in a single value, what would you pick?

Mean, Median, Mode

Mean = average
Add all the values and then divide by the total number of values.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

When we talk about the population mean (rather than the sample mean), we use the Greek letter mu: μ.

Median: the 50% point of the data; half the data is the same value or smaller than the median, and half the data is the same value or larger than the median.

To calculate the median by hand:

1) Sort the data in order (from smallest to largest)
2) Take the total number of observations n and calculate: $\frac{n+1}{2}$; this is the location in the sorted list where the media is located
    a. If the n is odd, then this formula gives you a whole number. Count through the list until you find the position of that value and then the number at that position is the median
    b. If the n is even, you get a ½ value , find the values on either side of that position, and then average the two values you get to get the median.

Typically, the median is used when the data is significantly skewed.

Mode: the most common value in the data set; continuous data often doesn't have a mode. In continuous data we may talk about the "modal class" – in a histogram, the bar with the most values (the tallest one) is called the modal class: give the range that the bar represents.

If you have more than one value that appears most frequently (multiple modes), then the practice is to report 1 mode, or 2 modes, but if there are 3 or more, report no mode.

Relationship between mean, median, mode:
If the distribution is symmetric: they are all about the same value (exactly symmetric, then exactly the same)
If the distribution is left skewed: mean is smallest, then median, then mode
If the distribution is right-skewed: mean the biggest, then median, then mode.

Measures of location and spread

Quartiles: are the ¼ and ¾ points of the data
After we find the median, then we divide the data in half at the median, and then the median of the bottom half is the first quartile, and the median of the top half is the third quartile.

First quartile: ¼ of the data is smaller (or equal to) the value at that location
Third quartile: ¾ of the data is smaller (or equal to) the value at that location

Trouble comes with: when the median is a value in the dataset, do we include in the bottom/top the half?

The standard we are going to follow is to include the median in both bottom and top half. The median the value where half of the data is equal to or less than the value. Use the inclusive version of the quartile formula in Excel.

Deciles, Percentiles
The deciles are the 10% marks: 10%, 20%, 30%, etc. up to 90% (somewhat uncommon)
Percentiles divide the data into 100 groups: at the percentage levels.
50^{th} percentile = median
25^{th} percentile = first quartile
75^{th} percentile = third quartile
30^{th} percentile = 30% of the data is below or equal to that value
99^{th} percentile = 99% of the data is below or equal to that value

When you get really close to 100% or 0%, carry more decimal places.
0^{th} percentile: smallest possible value (min)
100^{th} percentile: the largest possible value (max)

Calculate a percentile by hand:
1) Sort data from smallest to largest
2) If I want to find the percentile of a given value in the data set: location (position value) of that number and divide by the total number of observations. (then round as needed).

Where is a particular percentile in some data:
Multiply the percentile (as a decimal) by the total observations: give me the location in the sorted list of that percentile.

5-number summary:
Minimum, 1^{st} quartile, median, 3^{rd} quartile, maximum

Measures of Spread
Range, the Standard Deviation, IQR

Range: the difference between the maximum and minimum values.

	Salary
Minimum	$10,100
1st Quartile	29975
Median	$53,700

third Quartile	77025
Maximum	$168,800

Range: 168,800-10,100 = 158,700

IQR : Interquartile Range
The difference between the third quartile and the first quartile (middle 50% of the data)

IQR: 77025 – 29975 = 47,050

Standard deviation and variance

Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

1) Find the mean
2) Subtract the mean from every observation
3) Square all those values
4) Add them up
5) Divide by one less than the total number of observations

The standard deviation is the square root of the variance, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

These formulas are for sample data, and we will use these almost exclusively : if you aren't sure, use these versions.

For the population variance and standard deviation:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

If the problem does not say "population" don't use this formula.