STAT 107 – Fall 2020

Measures of Center & Variability

Ha Khanh Nguyen (hknguyen)

The Mean (The Average)

- The mean is responsive to ALL data points and can be greatly affected by extreme values.
- Formula for Sample Mean:

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

  - x_1 = the 1st observation in our data set
  - x_2 = the 2nd observation… all the way to the last data point collected.

- Sample Mean vs. Population Mean:
  - Keep in mind that sample mean and population mean are two different things.
  - The sample mean represents the limited data we have collected and is notated $\bar{x}$ **(x-bar)**
  - The population mean represents the entire population and is notated **μ (mu)**

The Median (The Midpoint)

- The median represents the value of the middle or average person/unit. If you arrange all observations in numerical order, then the median is the point such that 50% of the data falls at or below that value and 50% falls at or above that value!
- Consider the following observations. Circle the Median

12 17 21 22 23 25 25 27 32

- What if a data set has 10 numbers instead of 9? Then take the average of the middle two numbers as the middle point!

12 17 21 22 23 25 25 27 32 35

- FYI: median is typically represented as a symbol by just a capital "M" regardless of whether it's a sample statistic or population parameter.

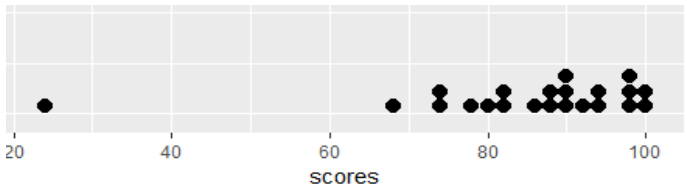
Mean vs. Median

- The mean is responsive to all data points. Changing just one value, especially on either extreme, will change the mean!
- Notice that when the lowest number goes down further, the median remains the same, but the mean goes down.

$$\begin{array}{lll} 13, 45, 78, 96 & : & \bar{x} = 58 \quad and \ M = 61.5 \\ 2, 45, 78, 96 & : & \bar{x} = 55.25 \ and \ M = 61.5 \end{array}$$

- For this reason, the median is an indicator of where the middle of your data is and is fairly resistant to extreme values. The mean is a representation of how all of your data balances and will get pulled more toward one side if there are more extreme values on that side.

Practice: Consider the following scores for a recent test.



Scores: 24, 68, 74, 74, 78, 80, 82, 82, 86, 88, 88, 90, 90, 90, 92, 94, 94, 98, 98, 98, 100, 100

Do you think the mean and median are different in this example? Which one will be lower?

Quartiles – Measures of Position

- Quartiles are three numbers that partition a data set into 4 approximately equal parts. In other words, 25% of the data falls between the first two values, 25% falls between values 2 and 3, 25% between values 3 and 4, and 25% from values 4 and 5.
- Q_1 is the 25th percentile. Also think of it as the median of the *lower* half of the dat.
- Q_2 is the 50th percentile. Which also makes it the **median** of the entire set of data!
- Q_3 is the 75th percentile, also the median of the *upper* half of the data.
- 5-Number Summary
  - **The 5-number summary**: (Minimum data point, Q_1, Q_2, Q_3, maximum data point)
  - This is a common set of numerical summaries to quickly get a sense of where your data falls.
  - 5-number summaries are also the basis for boxplots (which we will encounter in an upcoming section!)
  - Let's find the 5-number summary of test score data ____, ____, ____, ____, ____

Measures of Variability

- Consider the test scores data again: It's clear that not everyone has the same test score. There is clearly quite a bit of variation in the data! But ***how*** *do we* ***measure*** *that variation?*
- Let's look at different measures of variation and how they can help us answer these kinds of questions!
  - **Range – Distance from Maximum to Minimum**
    - Simply take the Maximum value – the Minimum Value
    - What is the range from the test score data? _____
    - Range is best used as a quick way to see the span of your data, but it isn't very good at telling you what is going on with most of your data points.
    - Range is also a problematic measure if you have an extreme value in your data.
  - **IQR (Inter-Quartile Range) – Distance from the 75th to 25th percentile**
    - IQR is considered a more *robust* measure of variability, meaning that it is unlikely to be thrown off by an extreme value.
    - The IQR of your data is calculated by taking $Q_3 - Q_1$
    - Find the IQR from the test score data:_____

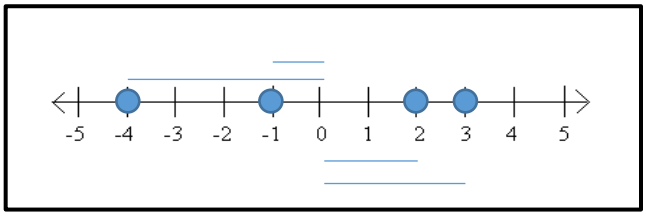
- Recognizing **Extreme Values** and choosing whether or not to label them as **Outliers**
  - An extreme value would be a general term for any data value that is unusually far or removed from most of the data.
  - One common way for identifying extreme values is the use of upper and lower fences.
  - **Upper and Lower Fences as Outlier-Identification Criteria**
    - Lower Fence = Q_1 - 1.5(Q_3- Q_1)
    - Upper Fence = Q_3 + 1.5(Q_3- Q_1)

Practice: Are any of our test score data points extreme values using the fence criteria?

- **Mean absolute deviation**
  - Notice that measures like Range and IQR do not actually account for every observation.
  - Consider a different way to measure variability: average distance from the mean.

Consider the following dataset representing the heights in inches of 10 high-school boys. What is the average distance from the mean in this dataset?

Rodrigo	70
Stan	65
Jeremy	73
Casey	68
David	62
Nick	69
Evan	71
Mickey	65
J. T.	70
Morgan	72



Standard Deviation and Variance

- The **Variance** for a dataset is:

$$\sigma^2 = \frac{\sum(x_i-\mu)^2}{n} = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}$$

  - This formula has certain nice properties in higher level statistics; it represents the *average squared distance* from the mean.
- **Standard Deviation** represents the *"typical deviation"* or "*average distance*" from the mean. This formula is more complex than variance, but produces an easier number to interpret!
- Standard deviation is calculated:

$$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n}} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}}$$

- Why both?
    - Standard deviation is a more practical measure to directly *interpret* since it is scaled in the units we are measuring, while variance is a simpler measure (no square root operation needed) and is often used when measuring variation within other calculations.
- **σ and σ²** represent population parameters for standard deviation and variance while **s and s²** represent sample statistics.
- **SPECIAL NOTE:** The formula for the sample statistics s and s^2 have a different denominator: instead of "n", it is "n-1"

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \text{ and } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Practice: Researchers decide to complete an experiment where some participants are asked to eat a fiber supplement with their breakfast every day while others are asked to eat a placebo bar. The researchers measure the participants' cholesterol levels to determine if the additional fiber has any affect on the participants' cholesterol.
Is this categorical data, discrete data, or continuous data?

The 5-number summary for the fiber supplement group is as follows:

119 138 145 166 196

What percentage of the fiber supplement group has a cholesterol level between 138 and 196?

Use this 5-number summary to determine the upper and lower fences. Are there any data values in this sample that seem to be extreme values?

It turns out that the highest measurement, 196, was actually a recording mistake. The correct measurement for that individual is 199. Which of the following summary statistics would change?

The Mean

The Standard Deviation

The Median

The IQR

To-do:

- Finish Lab 4, commit and push the lab using git commands!
- Complete HW 3 on Prairie Learn!

This lecture notes were adapted from "Course Notes for STAT 100: Statistics" by Kelly Findley & Ha Khanh Nguyen.