Basics of Statistics

By……. Sakeeb Sheikh

Measures of Central Tendency

- categories or scores that describe what is "average" or "typical" of a given distribution.

These include the mode, median and mean.

The Mode

- The mode is the category with the greatest frequency (or percentage).

Ex:

What is the mode of favourite flavours of Ice Cream?

Coconut = 22
Chocolate = 15
Vanilla = 7
Strawberry = 9

Answer would be coconut as highest frequency, NOT 22

The Median

- The median is the middlemost number.

- In other words, it's the number that divides the distribution exactly in half such that half the cases are above the median, and half are below.

- It's also known as the 50th percentile

The Median

- Note that finding the median requires first ordering all of the observations from least to greatest.

- For example, for the numbers 14, 6, 12, 18, 8, 4 ordering would be 4, 6, 8, 12, 14, 18 and its count is even number so median would be average of two middle numbers

- So the median is 10 (12 + 8 = 20; 20/2 = 10).

The Median

- One of the median's advantages is that it is not sensitive to outliers.

- An outlier is an observation that lies an abnormal distance from other values in a sample.

- For example,

  Distribution 1: 1, 3, 5, 7, 20
  Distribution 2: 1, 3, 5, 7, 20,000

- These two distributions have identical medians

The Mean

- The mean is typically refer to as "the average".
- It is the highest measure of central tendency
- The mean takes into account the value of every observation and thus provides the most information of any measure of central tendency.
- Unlike the median, however, the mean is sensitive to outliers.
- $\overline{X} = \frac{\sum X}{N}$

Measures of Variability

- We need to determine if the observations tend to cluster together or if they tend to be spread out. Consider the following example:

  Sample 1: {0, 0, 0, 0, 25}
  Sample 2: {5, 5, 5, 5, 5}

- Both of these samples have identical means (5) and an identical number of observations (n = 5), but the amount of variation between the two samples differs considerably.

Measures of Variability

Sample 1: {0, 0, 0, 0, 25}
Sample 2: {5, 5, 5, 5, 5}

- Sample 2 has no variability (all scores are exactly the same), whereas Sample 1 has relatively more (one case varies substantially from the other four).
- In this session, we will be going over four measures of variability: *the range, the inter-quartile range (IQR), the variance and the standard deviation.*

The Range

- The range is the difference between the highest and lowest scores in a data set and is the simplest measure of spread.

- We calculate range by subtracting the smallest value from the largest value. As an example, let us consider the following data set:

23	56	45	65	69	55	62	54	85	25

- The maximum value is 85 and the minimum value is 23. This gives us a range of 62 (85 – 23 = 62).

The Range

- Whilst using the range as a measure of variability doesn't tell us much.

- But it does give us some information about how far apart the lowest and highest scores are.

Quartiles and the Interquartile Range

- Quartiles basically means "quarter" or "fourth.“
- Finding the quartiles of a distribution is as simple as breaking it up into fourths.
- Each fourth contains 25 percent of the total number of observations.

Quartiles and the Interquartile Range

- Quartiles divide a rank-ordered data set into four equal parts.

Q1 is the "middle" value in the first half of the rank-ordered data set

Q2 is the median value of the data set

Q3 is the "middle" value of the second half of the rank-ordered data set

Q4 would technically be the largest value in the dataset, but we ignore it when calculating the IQR (we already dealt with it when we calculated the range).

Quartiles and the Interquartile Range

- Thus, the interquartile range is equal to Q3 minus Q1 (or the 75th percentile minus the 25th percentile, if you prefer to think of it that way).

As an example,

- Given numbers: 1, 3, 4, 5, 5, 6, 7, 11.
- Q1 is the middle value in the first half of the data set, Q1 = (3 + 4)/2 or Q1 = 3.5.
- Q3 is the middle value in the second half of the data set, Q3 = (6 + 7)/2 or Q3 = 6.5.
- The interquartile range is Q3 minus Q1, so the IQR = 6.5 - 3.5 = 3.

The Variance

- The variance is a measure of variability that represents on how far each observation falls from the mean of the distribution.

Ex: Given numbers represent my total monthly comic book purchases over the last five months : 2, 3, 5, 6, 9

- The first step in calculating the variance is finding the mean of the distribution. In this case, the mean is 5 (2+3+5+6+9 = 25; 25/5 = 5).

The Variance

- The second step is to subtract the mean (5) from each of the given observations (numbers):

  $2 - 5 = -3$
  $3 - 5 = -2$
  $5 - 5 = 0$
  $6 - 5 = 1$
  $9 - 5 = 4$

- Third, we square each of those answers to get rid of the negative numbers

  $(-3)^2 = 9$
  $(-2)^2 = 4$
  $(0)^2 = 0$
  $(1)^2 = 1$
  $(4)^2 = 16$

The Variance

- Fourth, we add them all together:

  $$9+4+0+1+16=30$$

- Finally, we divide by N-1, the total number of observations is 5, so $5 - 1 = 4$

  $$30/4 = \mathbf{7.5}$$

7.5 summarizes the amount of variability in our distribution.

The Variance

- The bigger the number, the more variability we have in our distribution.

- Please note: a variance can never be negative.

- If you come up with a variance that's less than zero, you've done something wrong.

The Standard Deviation

- In variance, when we square the numbers to get rid of the negatives (step 3), we also inadvertently square our unit of measurement.
- In other words, if we were talking about miles, we accidentally turned our unit of measurement into miles squared.
- If we were talking about comic books, we accidentally turned our unit of measurement into comic books squared (which, needless to say, doesn't always make a lot of sense).

The Standard Deviation

- In order to solve that problem, we calculate the standard deviation. The formula for the standard deviation looks like this:

$$Sx = \sqrt{S^2 x}$$

- In other words, calculating the standard deviation is as simple as taking the square root of the variance, reversing the squaring we did in the calculation of the variance.

- In our example, the standard deviation is equal to the $\sqrt{7.5}$ = **2.74**