Intro to descriptive statistics in WBS Primer - Quiz notes

What is the data type of "the number of students in a Data Science batch"? discrete data

Discrete data refers to counts or numbers that can only take certain specific values and cannot be meaningfully divided into smaller increments. This contrasts with continuous data, which can take any value within a range and can be subdivided into finer and finer increments.

For example, the number of students in a data science batch is discrete because you can count individual students (1, 2, 3, etc.), but you cannot have a fraction of a student. This number is a whole, specific value and doesn't make sense in fractions or decimals.

What is not a measure of central tendency? most

"Most" is not a measure of central tendency. *Measures of central tendency are statistical metrics used to identify a single value that is typical or representative of a set of data*.

Rectangular data is organized in rows and columns. True

True, *rectangular data is indeed organized in rows and columns*.

- **Rows:** Each row typically represents an *individual record or observation*. For instance, in a dataset of students, each row might represent a different student.
- **Columns:** Each column represents a *different variable or attribute*. For the student dataset example, columns might include variables like age, grade, major, etc.

This structure, resembling a rectangle, is why it's referred to as "rectangular data." It's the standard format used in spreadsheets.

What is the mode: [1, 1, 3, 7, 8]? 1

In the dataset `[1,1,3,7,8]`, the mode is indeed 1.

The mode is defined as the value that appears most frequently in a data set.

In this case, the number 1 appears twice, while all other numbers (3, 7, and 8) appear only once. Therefore, 1 is the mode, as it is the most frequent number in this dataset.

What do you need to calculate the range? The minimum and the maximum value.

To calculate the range of a dataset, *you need the minimum and the maximum values in that dataset.*

The range gives you an idea of the *spread or dispersion of the values in your dataset*. It shows how wide the span of values is, from the smallest to the largest.

The variance is the square root of the standard deviation. False

Standard Deviation is the square root of the Variance. The standard deviation gives you a sense of *how spread out the values in your data set are around the mean*, and *it's derived from the variance, not the other way around*.

1. **Variance**: *This measures how spread out a set of numbers is*. To find it, you first calculate the average (mean) of the numbers. Then, for each number, you subtract the mean and square the result. Finally, you average these squared differences.
2. **Standard Deviation**: *This is a measure of the amount of variation or dispersion in a set of values*. To find the standard deviation, you take the square root of the variance. This step helps to bring the units back to the original units of the data, making it easier to interpret.

What is not true about the lower quartile? It is half the value of the mean.

- **Lower Quartile (Q1)**: *This is a measure of central tendency that divides a data set into four equal parts*:
  - The *lower quartile is the median of the lower half of the data set*. It marks the 25th percentile, meaning that 25% of the data points in the set are below this value.
- **Mean**: This is the average value of a data set, calculated by adding up all the numbers and then dividing by the count of the numbers.

What is not part of a boxplot? Mean.

The **mean** (average) of the data is not typically shown in a standard boxplot. The focus of a boxplot is more on the median and the spread of the data (through quartiles and IQR), rather than the mean.

Here's what a typical boxplot includes:

1. **Median**: This is the *middle value* of the dataset, which is highlighted as a line within the box.
2. **Quartiles**:
    - The lower quartile (Q1) is the median of the lower half of the dataset.
    - The upper quartile (Q3) is the median of the upper half of the dataset. These quartiles form the edges of the box in the boxplot.
3. **Interquartile Range (IQR)**: This is the range between the first and third quartiles (Q3 - Q1), representing the middle 50% of the data.
4. **Whiskers**: These lines extend from the quartiles to the minimum and maximum values in the data, excluding outliers.
5. **Outliers**: Points that fall outside of the whiskers. They are often indicated as dots outside the whiskers.

Outliers are typically all values that are more extreme than the Interquartile Range. False.

- **Outliers**: These are data points that are *significantly different from most other data points in a dataset*. In the context of a boxplot, outliers are often determined based on their distance from the Q1 and Q3 (the quartiles). A common method to identify outliers is to look for:
    - Values that are more than 1.5 times the IQR below the first quartile (Q1).
    - Values that are more than 1.5 times the IQR above the third quartile (Q3).

This means that not all values outside the IQR are considered outliers. Only those that exceed the IQR by a significant amount (typically 1.5 times the IQR) are labeled as outliers. *This method helps to distinguish between regular variation within the data and extreme values that are significantly different from the rest*.

Which would be the x with the highest bar in a histogram? x= `[1,1,1,2,3,3,4,4,4,5]` . 1 and 4

In a histogram, t*he height of each bar represents the frequency (or count) of values within a certain range or specific value*.

In a right-skewed distribution, the mean is larger than the median. True

In a right-skewed distribution *the mean is typically larger than the median*. Here's why:

1. **Right-Skewed Distribution**: This is a type of distribution where most of the data points are concentrated on the left, with the tail extending to the right. In other

words, there are a few unusually large values in the data set.

2. Mean: Since the mean takes into account all data points, including the large values in the tail of a right-skewed distribution, it gets pulled towards the right.
3. Median: The median, being the middle value, is less affected by extreme values. In a right-skewed distribution, the median will be closer to the bulk of the data on the left.
   As a result, in a right-skewed distribution, the mean is generally greater than the median. This difference between the mean and median can actually be used as an indicator of skewness in a data set.

Which term does not describe the number or position of a distribution's peaks?
Univariate

- Univariate: This term refers to *data that consists of only one variable or attribute. It's about the type of data being analyzed, not the characteristics of its distribution such as peaks. For example, a dataset containing only the heights of a group of people is univariate*.
  In contrast, terms that do relate to the number or position of a distribution's peaks include:
- Unimodal: A distribution with one peak.
- Bimodal: A distribution with two distinct peaks.
- Multimodal: A distribution with more than two peaks.
  These terms focus on the shape of the distribution, particularly how many peaks (high points) it has, which can indicate the presence of different subgroups within the data or other features of the data's distribution.