**Questions**

1. ***What is the difference between ordinal and nominal scale of measurement?***

The **nominal scale of measurement** is a qualitative measure that uses discrete categories to describe a characteristic of the research participants. For each participant, the researcher determines the presence, absence, and type of the attribute. Nominal scales of measurement may have two categories, such as citizen status (citizen/non-citizen), or they can have more than two categories, like religious affiliation (e.g., Agnostic, Buddhist, Jewish, Muslim) or marital status (e.g., divorced, married, single).

An **ordinal scale of measurement** rank-orders participants on some scale or attribute, but the difference between numbers does not convey fixed or equal differences. Thus, with ordinal data, we know that a one-unit increase in an ordinal scales represents “more,” but we don’t know how much more. For example, a group of participants can be rank-ordered from least to most politically active. We know that a person who is ranked as 5 is more politically active than a person who is ranked as 4, but not how much more politically active.

1. ***How to choose the number of intervals in histogram?***

There are few rules of thumb for choosing the number of intervals. They are discussed below where “K” is the number of bins and “n” is the number of data points.

1. Square-root choice: (used by Excel generated histograms)
2. Sturge’s formula: . It is derived from a binomial distribution and implicitly assumes an approximately normal distribution.

However, it is a recommended approach to try out different combinations before settling on a final value for the number of intervals.

1. ***How to choose measure of center between mean and median?***

As we know the mean is very sensitive to outliers (because it factors in their magnitude), while the median is resistant to outliers. The mean is an appropriate measure of center only for symmetric distributions with no outliers. In all other cases, the median should be used to describe the center of the distribution.

1. ***Why use IQR as measure of spread instead of range or standard deviation? When n(the number of data points) is odd, does the median value lie in the top 50% or bottom 50% of the values?***

IQR is measured without taking into account the mean value of the distribution. But the range and standard deviation takes the mean value into account. And mean value is susceptible to outliers. So, IQR is unaffected by outliers while range and standard deviation are affected by outliers. ***BONUS***: This is the reason IQR is the one of the most popular methods to identify and remove outliers.

When the number of data points is odd, the median does not lie in either the top 50% or bottom 50% of the values, it lies exactly between them.

1. ***Mention some possible use cases where coefficient of variation should not be used?***

Some cases where coefficient of variation should not be used are:

* Samples containing both positive and negative values as this might lead to a zero value.
* Comparing coefficients of variation between parameters using relative units can result in differences that may not be real. If we compare the same set of temperatures in Celsius and Fahrenheit (both relative units, where Kelvin and Rankine scale are their associated absolute values):

Celsius: [0, 10, 20, 30, 40]

Fahrenheit: [32, 50, 68, 86, 104]

The sample standard deviations are 15.81 and 28.46, respectively. The CV of the first set is 15.81/20 = 0.79. For the second set (which are the same temperatures) it is 28.46/68 = 0.42.

If, for example, the data sets are temperature readings from two different sensors (a Celsius sensor and a Fahrenheit sensor) and you want to know which sensor is better by picking the one with the least variance, then you will be misled if you use CV. The problem here is that you have divided by a relative value rather than an absolute.

Comparing the same data set, now in absolute units:

Kelvin: [273.15, 283.15, 293.15, 303.15, 313.15]

Rankine: [491.67, 509.67, 527.67, 545.67, 563.67]

The sample standard deviations are still 15.81 and 28.46, respectively, because the standard deviation is not affected by a constant offset. The coefficients of variation, however, are now both equal to 0.0539.

1. ***What should be the measure of center with IQR as measure of spread?***

With IQR as measure of spread the measure of center should be median.

1. ***Does zero covariance imply two random variables are independent?***

No. Covariance only captures linear relationships. It can also be that there exists a non-linear relationship between them; in which case the value of covariance will be zero.

1. ***What is the covariance of a random variable with itself?***

The covariance of a random variable with itself is the variance of that variable. Mathematically,

1. ***Is covariance sensitive to scale of the random variable?***

Yes, covariance is sensitive to scale of the random variable.

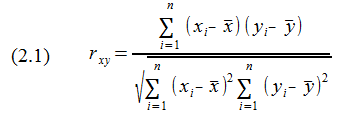
1. ***On addition of a constant to both the random variables does the covariance change?***

No, the covariance doesn’t change.

1. ***Why the denominator is and not while calculating sample variance?***

The whole purpose of calculating the sample variance is to get an estimate of population Dividing by would result in a biased estimator for population variance. To get an unbiased estimator, division by is recommended.

1. ***Does outlier effect correlation coefficients (Pearson and Spearman)?***
2. Pearson correlation: The formula is given as



A large outlier in either x or y will have different impacts on the numerator and on the denominator in (2.1). The Pearson correlation coefficient is therefore sensitive to outliers in the data, and it is therefore not robust against them.

1. Spearman correlation: Outliers do not effect spearman correlation coefficient because we are taking only the ranks of every instance of random variables and not its actual value.
2. ***Which coefficient is better to use when:***
3. ***Variables are normal distributed***
4. ***Variables are non-normally distributed but they have a monotonic relationship***
5. ***Variables are non-normally distributed and they do not have a monotonic relationship***
6. Pearson coefficient
7. Spearman coefficient
8. Neither Pearson nor Spearman coefficient. We should look towards other measures