Q3. Explain the significance of IQR.

The interquartile range is the measure of static depression and the spread of the data. Instead of data, there are two quartiles, namely the upper quartile and the lower quartile. The interquartile range is used to explain the difference between the upper and lower quartiles in the set of data. The interquartile range can be used to denote or indicate the variability of the set provided to you.

The interquartile range is a measure of variability based on splitting data into quartiles. Quartile divides the range of data into four equal parts. The values that split each part are known as the first, second, and third quartile. And they are represented by Q1, Q2, and Q3.

Q1 - It is the middle value in the first half of the rank-order data

Q2 - It is the median value in the set

Q3 - It is the middle value in the second half of the rank-ordered data set.

The interquartile range is equal to quartile 3 minus quartile 1.

Q12. Explain the difference between Correlation and Causation with an example.

Correlation describes an association between types of variables: when one variable changes, so does the other. A correlation is a statistical indicator of the relationship between variables. These variables change together: they covary. But this covariation isn’t necessarily due to a direct or indirect causal link.

Causation means that changes in one variable brings about changes in the other; there is a cause-and-effect relationship between variables. The two variables are correlated with each other and there is also a causal link between them.

A correlation does not imply causation, but causation always implies correlation.

Examples of correlation, NOT causation:

1. **Sunscreen and Sunburn**: There is a strong negative correlation between sunscreen usage and sunburn incidents. People who use sunscreen regularly are less likely to experience sunburn. However, this correlation does not mean that sunscreen directly prevents sunburn. Other factors, such as sun exposure and skin type, also play a role in determining the likelihood of sunburn.
2. **Exercise and Weight Loss**: Many studies have shown a positive correlation between regular exercise and weight loss. Individuals who exercise regularly tend to have lower body weights. While this correlation is well-documented, it does not imply that exercise directly causes weight loss. Diet, genetics, and metabolism also contribute to weight management.
3. **Education and Income**: Research has found a positive correlation between higher levels of education and higher income levels. Individuals with advanced degrees typically earn more money. However, this correlation does not prove that education directly causes increased income. Factors such as job skills, experience, and industry demand also influence income levels.

**Examples of causation:**

After I exercise, I feel physically exhausted.

This is cause-and-effect because I’m purposely pushing my body to physical exhaustion when doing exercise. The muscles I used to exercise are exhausted (effect) after I exercise (cause). This cause-and-effect **IS** confirmed.

Q13. Why do we need sampling? Provide a real-world example.

Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.

Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population’s characteristics by directly observing only a portion (or sample) of the population.

* Selecting a sample requires less time than selecting every item in a population
* Sample selection is a cost-efficient method
* Analysis of the sample is less cumbersome and more practical than an analysis of the entire population

Let us understand this at a more intuitive level through an example.

We want to find the average height of all adult males in Delhi. The population of Delhi is around 3 crore and males would be roughly around 1.5 crores (these are general assumptions for this example so don’t take them at face value!). As you can imagine, it is nearly impossible to find the average height of all males in Delhi.

It is also not possible to reach every male so we cannot really analyze the entire population. So, what *can* we do instead? We can take multiple samples and calculate the average height of individuals in the selected samples.

But then we arrive at another question — how can we take a sample? Should we take a random sample? Or do we have to ask the experts?

Let us say we go to a basketball court and take the average height of all the professional basketball players as our sample. This will not be considered a good sample because generally, a basketball player is taller than an average male and it will give us a bad estimate of the average male’s height.

Here is a potential solution — find random people in random situations where our sample would not be skewed based on heights.

Q14. Define Null Hypothesis, Alternate Hypothesis, Significance Level (α), and P-value.

The null and alternative hypotheses are two competing claims that researchers weigh evidence for and against using a statistical test:

* Null hypothesis (*H*0): There’s no effect in the population.
* Alternative hypothesis (*H*a or *H*1): There’s an effect in the population.

The effect is usually the effect of the independent variable on the dependent variable. If the sample provides enough evidence against the claim that there’s no effect in the population (*p* ≤ α), then we can reject the null hypothesis. Otherwise, we fail to reject the null hypothesis. The alternative hypothesis is the complement to the null hypothesis.

In hypothesis testing, the significance level, denoted as alpha (α), is a predetermined threshold that defines how much evidence we require to reject the null hypothesis. Commonly used significance levels include 0.05 (5%) and 0.01 (1%). Choosing the significance level depends on the desired balance between making correct decisions and avoiding errors.

In hypothesis testing, the p-value represents the probability of obtaining an observed result, or more extreme results, if the null hypothesis were true. A smaller p-value suggests stronger evidence against the null hypothesis.

The interpretation of p-values is as follows:

* If p-value ≤ alpha (α): You have enough evidence to reject the null hypothesis. It suggests that the observed data is unlikely to have occurred under the assumption that the null hypothesis is true.
* If p-value > alpha (α): You do not have enough evidence to reject the null hypothesis. It suggests that the observed data is consistent with the null hypothesis.

So, the significance level alpha helps determine the threshold for accepting or rejecting the null hypothesis. P-values provide a measure of the strength of evidence against the null hypothesis.

Q20. Summarize the key takeaways from the analysis performed above and describe how descriptive and inferential statistics can be used in real-world data analysis.

Descriptive and inferential statistics are two fields of statistics. Descriptive statistics is used to describe data and inferential statistics is used to make predictions. Descriptive and inferential statistics have different tools that can be used to draw conclusions about the data.

In descriptive and inferential statistics, the former uses tools such as central tendency, and dispersion while the latter makes use of hypothesis testing, regression analysis, and confidence intervals.

Descriptive and inferential statistics need to be used hand in hand to analyze the data in the best possible way. Some examples of descriptive and inferential statistics are given below:

* Suppose the scores of 100 students belonging to a specific country are available. The performance of these students needs to be examined. This data by itself will not yield any valuable results. However, by using descriptive statistics, the spread of the marks can be obtained thus, giving a clear idea regarding the performance of each student.
* Now suppose the scores of the students of an entire country need to be examined. Using a sample of, say 100 students, inferential statistics is used to generalize about the population.

If you do choose to use one of these methods, keep in mind that [your sample needs to be representative of your population](https://www.statology.org/representative-sample/), or the conclusions you draw will be unreliable.