Assignment

1. **What is the difference between inferential statistics and descriptive statistics?**
   Descriptive statistics summarize the characteristics of a data set. Inferential statistics allows you to test a hypothesis or assess whether your data is generalizable to the broader population.

2. **What is the difference between population and sample in inferential statistics?**
   Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from.

3. **Most common characteristics used in descriptive statistics?**
   A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information
   • Descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.
   • Some measures that are commonly used to describe a data set are:
   a) Central tendency - It includes the mean, median and mode

   The mean indicates the region where most values in a distribution fall to a central data point. The mean summarizes an entire dataset with a single number representing the data's centre point or typical value. It is also known as the arithmetic average, and it is one of several measures of central tendency.

   The median is the middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the average

   In statistics, the mode is the value that is repeatedly occurring in a given set.

   b) Variability or dispersion - It includes standard deviation, variance, the range and interquartile range (IQR) values of the variables.
   c) Skewness-The measure tells you whether the distribution of values is symmetric or skewed.
   The 3 main types of descriptive statistics concern the **frequency distribution, central tendency, and variability of a dataset**.

4. **How to calculate range and interquartile range?**
   The range of a dataset is the difference between the largest and smallest values in that dataset.

   The interquartile range is the middle half of the data.
   • To visualize it, think about the median value that splits the dataset in half. Similarly, you can divide the
   datainto quarters.
   • Statisticians refer to these quarters as quartiles and denote them from low to high as Q1, Q2, Q3, and
   Q4.
   • The lowest quartile (Q1) contains the quarter of the dataset with the smallest values.

   The formula for calculating the interquartile range takes the third quartile value and subtracts the first quartile value.
   To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1. (Distance)
   While the range gives you the spread of the whole data set, the interquartile range gives you the spread of the middle half of a data set.
   A smaller width means you have less dispersion, while a larger width means you have more dispersion.

5. **How is the statistical significance of an insight assessed?**

Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.

Here, a “hypothesis” is an assumption or belief about the relationship between your datasets. The result of a hypothesis test allows us to see whether this assumption holds under scrutiny or not.

A standard hypothesis test relies on two hypotheses.

Null hypothesis: The default assumption of a statistical test that you’re attempting to disprove (e.g., an increase in cost won’t affect the number of purchases).

Alternative hypothesis: An alternate theory that contradicts your null hypothesis (e.g., an increase in cost will reduce the number of purchases). This is the hypothesis you hope to prove.

The testing part of hypothesis tests allows us to determine which theory, the null or alternative, is better supported by data. There are many hypothesis testing methodologies, and one of the most common ones is the Z-test

Assignment

1.**What does symmetric distribution mean?**

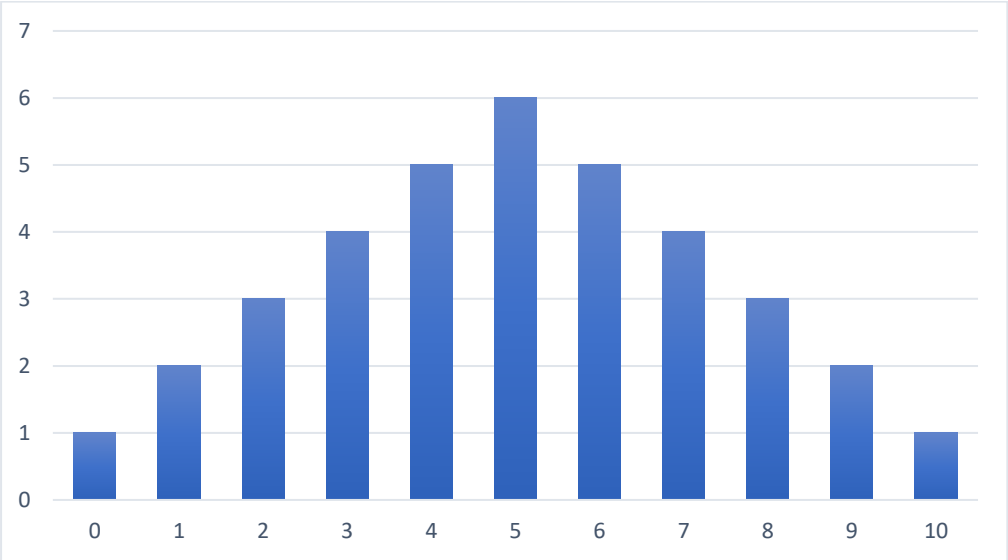
A symmetrical distribution occurs when the values of variables appear at regular frequencies and often the mean, median, and mode all occur at the same point. If a line were drawn dissecting the middle of the graph, it would reveal two sides that mirror one other.

In graphical form, symmetrical distributions may appear as a normal distribution (i.e., bell curve). The normal distribution is symmetric. It is also a unimodal distribution (it has one peak).By definition, a symmetric distribution is never a skewed distribution.

Distributions don't have to be unimodal to be symmetric. They can be bimodal (two peaks) or multimodal (many peaks).

In a symmetric distribution, the mean, mode and median all fall at the same point. The mode is the most common number and it matches with the highest peak

Below is an example of a symmetric data distribution, shown as a symmetric bar graph. The graph shows the number of questions answered correctly by a fictional class of students on a pop quiz. The scores, in order from lowest to highest, are 0, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 9, 9, 10.



A z score can be defined as a measure of the number of standard deviations by which a score is below or above the mean of a distribution. In other words, it is used to determine the distance of a score from the mean. If the z score is positive it indicates that the score is above the mean. If it is negative then the score will be below the mean. However, if the z score is 0 it denotes that the data point is the same as the mean.

2**.What is left skewed distribution and right skewed distribution?**

For skewed distributions, it is quite common to have one tail of the distribution considerably longer or drawn out relative to the other tail. A "skewed right" distribution is one in which the tail is on the right side. A "skewed left" distribution is one in which the tail is on the left side.

A left-skewed distribution has a long left tail. Left-skewed distributions are also called negatively-skewed distributions. That's because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak.

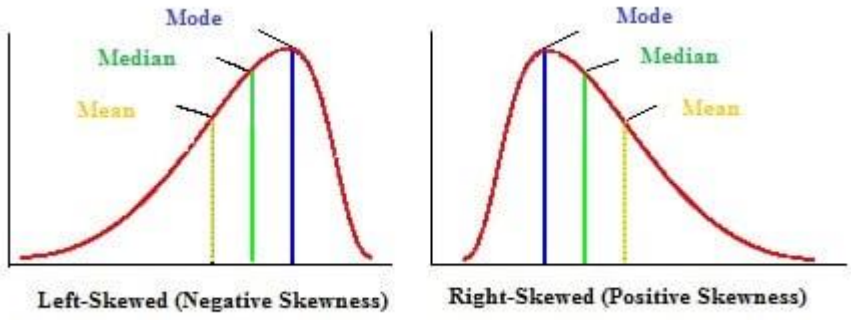
A left-skewed, negative distribution will have the mean to the left of the median.

Mean < Median < Mode

A right-skewed distribution has a long right tail. Right-skewed distributions are also called positive-skew distributions. That's because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak.

A right-skewed distribution will have the mean to the right of the median.

Mode < Median < Mean



3.**Where are long-tailed distributions used?**

A long tail distribution has tails that taper off gradually rather than drop off sharply. They are a subset of heavy-tailed distributions.

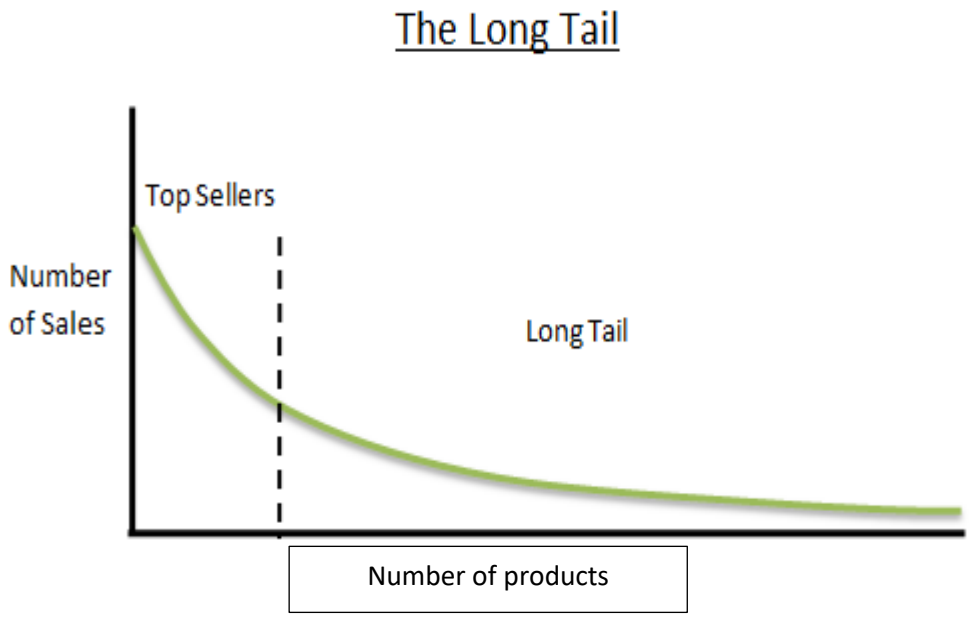
It is used to model many internet-era phenomena such as the frequency distribution of book titles sold at Amazon.com or the frequency of internet search terms.

One distribution is said to have a longer tail than another if its probability density (or mass) function is (asymptotically) larger than the other distribution's for very large values of the variable. Long tail distributions are the basis of many business models like Amazon, Netflix

Applications of a Long Tail Distribution

Commerce and marketing schemes often find that there sales can best be modelled by long tail distributions. For instance, an internet store may have certain items with very high sales (modelled by the centre of the distribution curve) and a large number of items with much lower sales (modelled by the long tail).

Although the sales volume for every individual item in the tail may be negligible, there are enough items that they play a significant role in the general profit taking. In fact, the profit from low-sale volume items can rival or even sometimes leave in the dust the profit made from best-sellers—provided only the tail is long enough.



4.**What is the central limit theorem?**

The central limit theorem in statistics states that, given a sufficiently large sample size, the sampling distribution of the mean will approximate a normal distribution regardless of that variable's distribution in the population. If you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

In a population, the values of a variable can follow different probability distributions. These distributions can range from normal, left-skewed, right-skewed, and uniform among others.

Assumptions Behind the Central Limit Theorem

It is important to understand the assumptions behind CLT:

The data must adhere to the randomization rule. It needs to be sampled at random.

The samples should be unrelated to one another. One sample should not impact the others.

When taking samples without replacement, the sample size should not exceed 10% of the population.

When the population is symmetric, a sample size of 30 is generally considered reasonable.

The shape of the sampling distribution of the means changes with the sample size. And, the definition of the central limit theorem states that when you have a sufficiently large sample size, the sampling distribution starts to approximate a normal distribution. How large does the sample size have to be for that approximation to occur?

It depends on the shape of the variable's distribution in the underlying population. The more the population distribution differs from being normal, the larger the sample size must be. Typically, statisticians say that a sample size of 30 is sufficient for most distributions. However, strongly skewed distributions can require larger sample sizes.

Consider there are 15 sections in class X, and each section has 50 students. Our task is to calculate the average marks of students in class X.

To begin, select groups of students from the class at random. This will be referred to as a sample. Create several samples, each with 30 students.

Calculate each sample's individual mean.

Calculate the average of these sample means.

The value will give us the approximate average marks of the students in Class X.

The histogram of the sample means marks of the students will resemble a bell curve or normal distribution.

Significance of Central Limit Theorem

The CLT has several applications. Look at the places where you can use it.

Political/election polling is a great example of how you can use CLT. These polls are used to estimate the number of people who support a specific candidate. You may have seen these results with confidence intervals on news channels. The CLT aids in this calculation.

You use the CLT in various census fields to calculate various population details, such as family income, electricity consumption, individual salaries, and so on.

(Example: Consider 10 samples of size=10, plot a graph for the sample mean distribution, it will approximate a normal distribution, More the sample size, more the sample mean distribution will approximate a normal distribution

5. **What are observational and experimental data in statistics?**

In an observational study, we measure or survey members of a sample without trying to affect them.

In a controlled experiment, we assign people or things to groups and apply some treatment to one of the groups, while the other group does not receive the treatment.

Example for observational study:

Problem 1: Drinking tea before bedtime

A study took random sample of adults and asked them about their bedtime habits. The data showed that people who drank a cup of tea before bedtime were more likely to go to sleep earlier than those who didn't drink tea.

This study was a survey that looked at if people drank tea or not and when they went to bed.

The people were not randomly assigned to groups.

This was an observational study.

Example for an Experiment:

Another study took a group of adults and randomly divided them into two groups. One group was told to drink tea every night for a week, while the other group was told not to drink tea that week. Researchers then compared when each group fell asleep.

This study randomly assigned people to groups.

One group was given a treatment and the other group was not.

This was an experiment.