Introduction to statistics

Assignment no. 03

Rahee sutrave.

Statistical significance:

- 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.
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

Normal distribution is used to represent how data is distributed and is primarily defined by:
•The mean (μ): The mean represents the location of the center of your data (or the average).
•The standard deviation (σ): The standard deviation is a measure of the amount of variation or dispersion of a set of values and represents the spread in your data.

In statistics, the distance between a data point and the mean of the data set is assessed as a Z-score. The Z-score (also known as the standard score) is the number of standard deviations by which a data point is distanced from the mean.
The final concept we need to use the Z-test is that of P-values. A P-value is the probability of finding results at least as extreme as those measured when the null hypothesis is true.

This significance value varies by situation and field of study, but the most commonly used value is 0.05, corresponding to a 5% chance of the results occurring randomly.

For a Z-test, the normal distribution curve is used as an approximation for the distribution of the test statistic. To carry out a Z-test, find a Z-score for your test or study and convert it to a P-value. If your P-value is lower than the significance level, you can conclude that your observation is statistically significant.

statistical significance testing use for:
- Landing page conversions
- Notification / email response rates and conversion rates
- User reactions to product launches
- User reactions to pricing
- User reactions to a new design
- User reactions to newly launched features

Mean:
Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers. It is denoted by $\bar{X}$
Mean = (Sum of all the observations/Total number of observations)
$\bar{x}=\sum x/n$

In the case of a discrete probability distribution of a random variable X, the mean is equal to the sum over every possible value weighted by the probability of that value; that is, it is computed by taking the product of each possible value x of X and its probability P(x) and then adding all these products together.

There are majorly three different types of mean value that you will be studying in statistics.
1.Arithmetic Mean:

$$A.M. = \Sigma f_i x_i / \Sigma f_i.$$

2.Geometric Mean:

$$G.M. = \sqrt[n]{(x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n)}$$

3.Harmonic Mean:

$$H.M.= 1/[\Sigma(1/x_i)]/N = N/\Sigma(1/x_i)$$

Mean, median, and mode are the three statistical measures of the central tendency of data.

Standard deviation:

- Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a “typical” deviation from the mean.
- It is a popular measure of variability because it returns to the original units of measure of the data set.
- Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values.
- It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.
- The formula for standard deviation (SD) is

$$SD = N\sum|x-\mu|2$$

where ∑sum means sum of

x is a value in the data set,

μ is the mean of the data set,

N is the number of data points in the population.

If sample is there then standard deviation(SD) is

$$SD\ sample = n-1\sum|x-\bar{x}|2$$

where s = Sample standard deviation

n = Number of observations in sample

x = ith observation in the sample

$\bar{x}$ = Sample mean

Correlation:

- Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).
- It's a common tool for describing simple relationships without making a statement about cause and effect.

Correlation coefficient:

- The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables(x and y) and it ranges from -1 to +1.
- A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables.
- A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.
- A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable.

Scatter diagram:

A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis.

Type of correlation:
The scatter plot explains the correlation between the two attributes or variables.
•Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
•Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
•No Correlation – when there is no linear dependence or no relation between the two variables.

covariance:

- Covariance is a measure of the relationship between two random variables and to what extent, they change together.
- it defines the changes between the two variables, such that change in one variable is equal to change in another variable.
- This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.
- Formula for covariance:

$Cov(X,Y) = \Sigma E((X – \mu) E(Y – v)) / n-1$

- where:

•X is a random variable
•E(X) = μ is the expected value (the mean) of the random variable X and
•E(Y) = ν is the expected value (the mean) of the random variable Y
•n = the number of items in the data set.
•Σ summation notation.

Types of covariance:

- Positive covariance:

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behavior. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

- Negative covariance:

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

Uses of inferential statistics:

- Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population.

1. making estimates about populations (for example, the mean SAT score of all 11th graders in the US).
2. testing hypotheses to draw conclusions about populations (for example, the relationship between SAT scores and family income).
3. To study a sample by applying the desired tool.
4. To make generalizations about the population from which the sample has been drawn.
5. To predict the behavior of the population with accuracy.

One sample t test:

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

You can use the test for continuous data. Your data should be a random sample from a normal population.

For a valid test, data should be:

- Independent (values are not related to one another).
- Continuous.
- Obtained via a simple random sample from the population.
- Population normally distributed.

The One Sample *t* Test is commonly used to test the following:

- Statistical difference between a mean and a known or hypothesized value of the mean in the population.
- Statistical difference between a change score and zero.

This approach involves creating a change score from two variables, and then comparing the mean change score to zero, which will indicate whether any change occurred between the two time points for the original measures. If the mean change score is not significantly different from zero, no significant change occurred.

- The null hypothesis (H_0) and (two-tailed) alternative hypothesis (H_1) of the one sample *T* test can be expressed as:

H_0: $\mu = \mu_0$ ("the population mean is equal to the [proposed] population mean")
H_1: $\mu \neq \mu_0$ ("the population mean is not equal to the [proposed] population mean")
where μ is the "true" population mean and μ_0 is the proposed value of the population mean.

formula of one sample t test:

$$t=(x-\mu 0)/sx$$
$$Sx= s/\sqrt{n}$$

where
$\mu 0$ = test value proposed constant for population mean
X= sample mean
sx= Estimated standard error of the mean (s/sqrt(n))
s= sample standard deviation
n=sample size

The calculated *t* value is then compared to the critical *t* value from the *t* distribution table with degrees of freedom *df* = *n* - 1 and chosen confidence level. If the calculated *t* value > critical *t* value, then we reject the null hypothesis.

Relation between standard deviation and standard variance

Standard deviation:

The degree of dispersion is computed by the method of estimating the deviation of data points. It is denoted by the symbol, 'σ'.

formula: $SD=N\sum|x-\mu|2$

where ∑sum means sum of

x is a value in the data set,

μ is the mean of the data set,

N is the number of data points in the population.

Variance:

the variance is a measure of how far a set of data are dispersed out from their mean or average value. It is denoted as 'σ^2'.

Formula: $\sigma 2=N\sum i=1n(xi-x)2$

σ^2 = Population variance

N = Number of observations in population

X_i = ith observation in the population

μ = Population mean

The relationship between the variance and the standard deviation for a sample data set is given below:

- Variance represents the average squared deviations from the mean value of data, while standard deviation represents the square root of that number.
- Both, the variance and the standard deviation measures variability in a distribution.
- Both have different units like the standard deviation has the same units as the original values like minutes or meters while the variance has much larger units like meters squared.
- The variance is equal to the square of standard deviation or the standard deviation is the square root of the variance.

According to formula
Variance = Square of Standard Deviation or
$V = \sigma^2$
Where, V = Variance and σ = Standard Deviation
Hence, the variance is equal to the square of standard deviation.

One way anova test:

- One way anova is a technique that can be used to compare whether two sample's means are significantly different or not (using the F distribution). This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way”

These estimates rely on various assumptions

•Response variable residuals are normally distributed (or approximately normally distributed).
•Variances of populations are equal.
•Responses for a given group are independent and identically distributed normal random variables (not a simple random sample (SRS)).

The test formulates a null hypothesis and an alternative hypothesis. The null hypothesis states that all population means are equal, whereas the alternative hypothesis states that at least one population mean will vary from others.
Comparing the F statistics calculated to the critical F value obtained from the F table will give prolific information. The result is significant if the F statistic is larger than the F value. A significant result points to the difference in mean. If the F-statistics is 1 or close to 1, then the two variances are equal and state that the null hypothesis is true.
Uses:

- To find solution of problems regarding population or sample
- To accept or reject null hypothesis
- test can reveal whether at least two groups were different from each other