Today's Quiz - What is the significance of p value?

Company Name - Myntra

Role - Data Scientist

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true).

The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

However, if the p-value is below your threshold of significance (typically $p < 0.05$), you can reject the null hypothesis, but this does not mean that there is a 95% probability that the alternative hypothesis is true. The p-value is conditional upon the null hypothesis being true, but is unrelated to the truth or falsity of the alternative hypothesis.

A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

The term significance level (alpha) is used to refer to a pre-chosen probability and the term "P value" is used to indicate a probability that you calculate after a given study.

·The significance level (alpha) is the probability of type I error. Type I error is the false rejection of the null hypothesis and type II error is the false acceptance of the null hypothesis. The power of a test is one minus the probability of type II error (beta)

Today's Quiz - What are the types of random variables?

Company Name - Infosys

Role - Data Scientist

A random variable, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

1. Discrete Random Variables

·A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4, ... Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete.

Example: Imagine a coin toss where, depending on the side of the coin landing face up, a bet of a dollar has been placed. The possibility of winning a dollar corresponding to the outcome of a coin toss before tossing the coin defines the random variable. The outcome of the coin toss is either heads, or tails, creating an equal probability of either outcome. Because the value of the random variable is defined as a real-valued dollar, the probability distribution is discrete.

2. Continuous Random Variables

A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements.

A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the area under a curve (in advanced mathematics, this is known as an integral).

The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Example: Imagine wanting to study the effects of caffeine intake on height. One's height would be the continuous random variable as it is unknown before the completion of the experiment, and its value is taken from measuring within a range.

Today's Quiz - What is Hypothesis Testing?

Company Name - Deloitte

Role - Data Scientist

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

· There are two types of hypothesis – Null and Alternative.

Null Hypothesis: A null hypothesis is the one in which sample observations result purely from chance. This means that the observations are not influenced by some non-random cause.

Alternative Hypothesis: An alternative hypothesis is the one in which sample observations are influenced by some non-random cause.

A hypothesis test concludes whether to reject the null hypothesis and accept the alternative hypothesis or to fail to reject the null hypothesis.

The following steps are involved in hypothesis testing:

The first step is to state the null and alternative hypothesis clearly. The null and alternative hypothesis in hypothesis testing can be a one tailed or two tailed test.

The second step is to determine the test size. This means that the researcher decides whether a test should be one tailed or two tailed to get the right critical value and the rejection region.

The third step is to compute the test statistic and the probability value. This step of the hypothesis testing also involves the construction of the confidence interval depending upon the testing approach.

The fourth step involves the decision making step. This step of hypothesis testing helps the researcher reject or accept the null hypothesis by making comparisons between the subjective criterion from the second step and the objective test statistic or the probability value from the third step.

The fifth step is to draw a conclusion about the data and interpret the results obtained from the data.

A null hypothesis is accepted or rejected basis P value and the region of acceptance.

P value – it is a function of the observed sample results. A threshold value is chosen before the test is conducted and is called the significance level, which is represented as α. If the calculated value of $P \leq \alpha$, it suggests the inconsistency between the observed data and the assumption that the null hypothesis is true. This suggests that the null hypothesis must be rejected

Region of Acceptance – It is the range of values that leads you to accept the null hypothesis. When you collect and observe sample data, you compute a test static. If its value falls within the specific range, the null hypothesis is accepted.

Today's Quiz - What is bias-variance trade off in ML?

Company Name - Infosys

Role - Data Scientist

The goal of any supervised machine learning algorithm is to achieve low bias and low variance.

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand, if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

Thus, there is a trade-off between bias and variance; in order to achieve good prediction performance. To build a good model, we need to find a balance between bias and variance such that it minimizes the total error.

Total error = Bias^2 + Variance + Irreducible Error

where,

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict, and

Variance is the variability of model prediction for a given data point or a value which tells us the spread of our data.

Irreducible error is the error introduced from the chosen framing of the problem and may be caused by factors like unknown variables that influence the mapping of the input variables to the output variable.

For example: - The k-nearest neighbors algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute the prediction and in turn increases the bias of the model.

Today's Quiz - What is the difference between box plot and histogram?

Company Name - Philips

Role - Data Scientist

Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.

Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.

Histogram is preferable when there is very little variance among the observed frequencies. While box plot shows moderate variation among the observed frequencies

While histograms are better in displaying the distribution of data, a box plot is used to tell if the distribution is symmetric or skewed.

Box plots are less detailed than histograms and take up less space.

To conclude, both tools can be helpful to identify whether variability in data is within specification limits, and whether there is a shift in the process over time. Thus, the type of chart aid chosen depends on the type of data collected, rough analysis of data trends, and project goals.

Today's Quiz - What is the difference between descriptive, predictive and prescriptive analysis?

Company Name - General Mills

Role - Data Scientist

Most organizations emphasize data to drive business decisions. But data alone is not the goal. Facts and figures are meaningless if you can’t gain valuable insights that lead to more-informed actions. The different types of analytics is as follows:-

1. Descriptive Analytics: It looks at data statistically to tell you what happened in the past. Descriptive analytics helps a business understand how it is performing by providing context to help stakeholders interpret information. This can be in the form of data visualizations like graphs, charts, reports and dashboards.

For instance, say that an unusually high number of people are admitted to the emergency room in a short period of time. Descriptive analytics tells you that this is happening and provides real-time data with all the corresponding statistics (date of occurrence, volume, patient details, etc.).