Ep 11 - Inferential Statistics 1

Dec 6, 2023

What does Inferential Statistics encompass?

Inferential statistics is a branch of statistics that involves using *sample data to make inferences or draw conclusions about a population*. It extends the information obtained from a sample to make predictions or generalizations about a larger group (population) from which the sample was drawn. This is crucial when it is impractical or impossible to study the entire population. Inferential statistics include the following measures and techniques:

1. Central Limit Theorem
2. Confidence Intervals
3. Hypothesis Testing
4. Tests and Errors

Sampling Distribution

Sampling Distribution is a probability distribution that describes the statistical properties of a sample statistic (sample mean or sample proportion) computed from multiple independent samples of the same size from a population.

What is the use of sampling distribution?

It is important in statistics and machine learning because it allows us to estimate the variability of a sample statistic, which is useful in making inferences about the population. By analysing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis testing, and make predictions about the population from the sample data.

Central Limit Theorem

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables.

The conditions required for the CLT to hold are:

1. The sample size is large enough, typically greater than or equal to 30.
2. The sample is drawn from a finite population or an infinite population with a finite variance.
3. The random variables in the sample are independent and identically distributed.

> **Important**
>
> The sampling mean and population mean are the same i.e. $\bar{x} = \mu$, but the sampling variance is equal to the population variance divided by the sampling size i.e.
>
> $$s^2 = \frac{\sigma^2}{n}$$

The CLT is important in statistics and machine learning because it allows us to make probabilistic inferences about a population based on a sample of data.

Short and Sweet Explanation of Central Limit Theorem
A Really Beautiful Graphical Explanation of Central Limit Theorem

Point Estimate

A point estimate is a single value, calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.

Confidence Interval

Confidence interval is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

Confidence Interval = **Point Estimate** + **Margin of Error**

There are two ways to calculate confidence intervals. They are:

1. **Z procedure**, for when σ is available
2. **t procedure**, for when σ isn't available

Nice Explanation of Confidence Intervals

> **Note**
>
> Confidence Intervals are for parameters and not statistics. Statistics help us calculate confidence intervals for parameters.

> **Important**
>
> The higher the Confidence Level, greater the Margin of Error.
> At 100% Confidence Level, the Margin of Error is ∞ and at 0% Confidence Level, the Margin of Error is $-\infty$

> **Important**
>
> The higher the Confidence Level, greater the range/confidence interval.

Z - procedure

The following assumptions must be true for Z - procedure to be applicable.

- Random sampling
- Known σ
- Normal distribution or large sample size

The formula to find confidence interval with confidence of $(1-\alpha)$, sample size n and mentioned standard deviation σ is given by

$$\text{C.I.} = \bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

t - procedure

The following assumptions must be true for t - procedure to be applicable.

- Random sampling
- Sample standard deviation s
- Approximate normal distribution, or a large sample size. If it is normally distributed, we can use a very small sample size as well
- Independent observations, particularly important when working with time series data

The formula to find confidence interval with confidence of $(1-\alpha)$, sample size n and sample standard deviation s is given by

$$\text{C.I.} = \bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

> ⚠ **Caution**
>
> If the population distribution is heavily skewed or has a lot of outliers, it is better to use non-parametric methods.

> ✎ **Note**
>
> For smaller sample sizes, the t value is comparatively greater than the corresponding z value. This can be explained due to the fact that with σ we are more confident as compared to s, therefore the greater value of t compensates that uncertainty.

Hypothesis Testing

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make *probabilistic statements about population parameters*.

It involves formulating a **null hypothesis** (often stating no effect or no difference) and an **alternative hypothesis**, then using sample data to assess the evidence against the null hypothesis. The goal is to *determine whether the observed results are statistically significant or if they could have occurred by chance*.

Applications of Hypothesis Testing

1. **Testing the effectiveness of interventions or treatments**: Hypothesis testing can be used to determine whether a new drug, therapy, or educational intervention has a significant effect compared to a control group or an existing treatment.
2. **Comparing means or proportions**: Hypothesis testing can be used to compare means or proportions between two or more groups to determine if there's a significant difference. This can be applied to compare average customer satisfaction scores, conversion rates, or employee performance across different groups.
3. **Analysing relationships between variables**: Hypothesis testing can be used to evaluate the association between variables, such as the correlation between age

and income or the relationship between advertising spend and sales.

4. **Evaluating the goodness of fit**: Hypothesis testing can help assess if a particular theoretical distribution (e.g., normal, binomial, or Poisson) is a good fit for the observed data.
5. **Testing the independence of categorical variables**: Hypothesis testing can be used to determine if two categorical variables are independent or if there's a significant association between them. For example, it can be used to test if there's a relationship between the type of product and the likelihood of it being returned by a customer.
6. **A/B testing**: In marketing, product development, and website design, hypothesis testing is often used to compare the performance of two different versions (A and B) to determine which one is more effective in terms of conversion rates, user engagement, or other metrics.

Hypothesis Testing in Machine Learning

1. **Model Evaluation:** Used to assess the significance of differences between observed and predicted values, helping determine whether a machine learning model's performance is statistically significant.
2. **Feature Selection:** When dealing with multiple features, it can be employed to identify the most relevant ones by testing whether the inclusion or exclusion of a feature significantly impacts model performance.
3. **Anomaly Detection:** It can be applied to identify unusual patterns or outliers in data, helping flag potential anomalies that might require further investigation.
4. **Comparing Models:** When working with different machine learning algorithms or variations of a model, hypothesis testing aids in determining whether one model significantly outperforms another.
5. **Hyperparameter Tuning**: It can be used to evaluate the performance of a model trained with different hyperparameter settings. By comparing the performance of models with different hyperparameters, you can determine if one set of hyperparameters leads to significantly better performance.
6. **Assessing Model Assumptions**: In some cases, machine learning models rely on certain statistical assumptions, such as linearity or normality of residuals in linear regression. Hypothesis testing can help assess whether these assumptions are met, allowing you to determine if the model is appropriate for the data.

Null Hypothesis (H_0)

Null Hypothesis is a statement that *assumes there is no significant effect or relationship between the variables* being studied. It serves as the starting point for hypothesis testing and represents the **status quo** or the *assumption of no effect until proven*

otherwise. The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of the alternative hypothesis, which claims there is a significant effect or relationship.

Alternate Hypothesis (H_1 or H_a)

The alternative hypothesis, is a statement that contradicts the null hypothesis and *claims there is a significant effect or relationship between the variables* being studied. It represents the **research hypothesis** or the claim that the researcher wants to support through statistical analysis.

Approaches to Hypothesis Testing

There are two approaches to hypothesis testing. They are:

1. Rejection Region Approach
2. p - value approach

> **How to decide what statement will be the null hypothesis?**
>
> Typically the null hypothesis claims nothing new is happening.

> **Important**
>
> Failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true. It just means that there isn't enough evidence to support the alternate hypothesis.
> Hypothesis testing is a lot like court trials, innocent until proven guilty.

Hypothesis Testing and The Null Hypothesis
Alternate Hypothesis