Data Science

Hypothesis Testing

Hypothesis

- Hypotheses are always statements or assumptions about the population which we want to verify.
- For eg.
  - Will I improve my grades if I spend 4 hrs studying daily ?
  - If breakfast helps children perform better in schools ?
  - If the average time that college students spend studying each week is 20 hours per week ?

Hypothesis

- It is an idea made from limited evidence, and is a starting point of further investigation.
- This is where you can use the sample data to answer the research questions.
- A Famous saying -
  - "A fact is a simple statement that everyone believes. It is innocent, unless found guilty. A hypothesis is a novel suggestion that no one wants to believe. It is guilty, until found effective."

Hypothesis Testing

- Formal definition -
    - “Hypothesis testing is an inferential procedure that uses sample data to evaluate the credibility of a hypothesis about a population.”
- A hypothesis test is a rule that specifies whether to accept or reject a claim about a population depending on the evidence provided by a sample of data.
- Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed.

Why Hypothesis Testing ?

Hypothesis Testing

Hypothesis Testing

- For drawing some inferences, we have to make some assumptions that lead to two terms that are used in the hypothesis testing -
  - The null hypothesis (H_0)
  - The alternative hypothesis (H_1 or H_A)

Null & Alternate Hypothesis

- Null Hypothesis (H_0)
  - In hypothesis testing, we begin by making a tentative assumption about population parameter. This tentative assumption is called null hypothesis.
  - A statement about the population parameter i.e. the statement which we want to test
  - Can include : =, >= , <=
- Alternate Hypothesis (H_1 or H_A)
  - A statement that directly contradicts Null Hypothesis
  - Can include : ≠, >, <

Examples

- A school claims that on an average their students get at least 70% marks
  - Claim - Average marks >= 70%
  - Counterclaim - Average marks < 70%
- According to a survey, average number of hours spent by phd students in their research work is more than 10 hours per day
  - Claim - Average hours > 10 hrs
  - Counterclaim - Average hours <= 10 hrs
- A company states that average life of their car tyres is 36 months.
  - Claim - Average life = 36 months
  - Counterclaim - Average life ≠ 36 months

Level of Significance

Significance Level

- It is denoted by Alpha
- Refers to the degree of significance in which we accept or reject the null-hypothesis.
- It is the probability of rejecting the null hypothesis, if it is true.
- Typical values for Alpha are - 0.01, 0.05, 0.1

Significance Level

- The choice of Alpha is determined by the context you are operating in but 0.05 or 5% is the most commonly used value.
- Alpha = 0.05 means, your output should be 95% confident to give similar kind of result in each sample.
- Based on the level of significance, we make a decision to accept the Null or Alternate hypothesis.

Test Statistics

Decision

- To decide the rejection or acceptance of Null hypothesis, we can use -
  - Test statistic
  - p value

Test Statistics

- A test statistic is calculated from sample data and used in a hypothesis test
- It is used to determine whether to reject or accept the null hypothesis.

Test Statistics

- There are 4 test statistics which we can use in hypothesis testing -
    - Z-test : Z-score
    - T-test : T-score
    - ANOVA : F-statistic
    - Chi-square test : Chi-square statistic
- The calculated test statistic is compared to the respective critical statistic to decide the rejection or acceptance of null hypothesis.

Test Statistic : z-score

- Z-test is a statistical test which is used when -
  - Data is normally distributed
  - Sample size is large, i.e. n >= 30
- Expression for z-test :

Critical Value

Critical Value

- In statistical hypothesis testing, the critical values of a statistical test are the boundaries of the acceptance region of the test.
- If a test statistic on one side of the critical value results in accepting the null hypothesis, a test statistic on the other side will result in rejecting the null hypothesis.
- Steps for using critical value in hypothesis testing -
  - Calculate the test statistic
  - Calculate critical values based on significance level (alpha)
  - Compare test statistic with critical values.

Rejection Region or Critical Region

Rejection Region

- Rejection region represents a set of values for the test statistic, for which the null hypothesis is rejected.
- If the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis

Type of Tests

Type of Tests

- Depending on the nature of alternate hypothesis, nature of statistical test will be used
- Different type of tests
    - Left tailed test
    - Right tailed test
    - Two tailed test

Rules

- If there is less than sign in alternate hypothesis, then we use left tail test
- If there is greater than sign in alternate hypothesis then we use right tail test
- If there is not equal sign in alternate hypothesis then we use two tail test

Examples

- A school claims that on an average their students get at least 70% marks
  - Claim - Average marks >= 70%
  - Counterclaim - Average marks < 70%
- According to a survey, average number of hours spent by phd students in their research work is more than 10 hours per day
  - Claim - Average hours > 10 hrs
  - Counterclaim - Average hours <= 10 hrs
- A company states that average life of their car tyres is 36 months.
  - Claim - Average life = 36 months
  - Counterclaim - Average life ≠ 36 months

Type of Errors

Errors in Hypothesis Testing

- We have 2 claims to be tested - Null hypothesis and alternate hypothesis. Only one of them can be true.
- Ideally we should not reject the null hypothesis when it is true and we should reject it when it is false (or alternate hypothesis is true)
- There are 2 types of errors -
  - Type 1 error
  - Type 2 error

Type 1 Error

- When you reject a true null hypothesis
- It is also called a false positive
- The probability of making this error is alpha, the level of significance
- An α of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.

Type 2 Error

- When you accept a false null hypothesis.
- The probability of making this error is denoted by Beta
- We should also mention that the probability of rejecting a false null hypothesis is : 1 - Beta
- This is the researchers goal to reject a false null hypothesis.
- Therefore (1 - Beta) is called “The Power of the test”

Minimise Type 1 & Type 2 errors