Day 12 - Hypothesis Testing, P-value, Types of Errors, Tailed Tests

1. What is the main purpose of hypothesis testing in data science and machine learning?

Hypothesis testing is used to determine whether a claim or assumption about a population is supported by sample data. In data science and machine learning, it helps validate assumptions, compare methods, and check if observed differences are statistically significant.

2. Define null hypothesis (H_0) and alternative hypothesis (H_1) with examples from a real-world ML scenario.

Null Hypothesis (H_0) assumes there is no effect or difference, while the Alternative Hypothesis (H_1) assumes there is an effect or difference.

Example (ML scenario):

- Suppose we test a new recommendation algorithm.
  - H_0: The new algorithm does not improve user engagement.
  - H_1: The new algorithm increases user engagement.

3. Explain the difference between parametric and non-parametric hypothesis tests.

Parametric vs Non-Parametric Hypothesis Tests:

- **Parametric tests** assume the data follows a specific distribution (usually normal) and often use parameters like mean and standard deviation.
  - **Example:** t-test, ANOVA.
- **Non-parametric tests** make **no assumptions** about the data's distribution and are used for ordinal or non-normal data.
  - **Example:** Mann-Whitney U test, Kruskal-Wallis test.

In short: Parametric tests rely on distribution assumptions; non-parametric tests do not.

4. Why is hypothesis testing important before building a predictive model?

Hypothesis testing is important before building a predictive model because it validates assumptions, identifies significant features or relationships, and ensures that patterns in the data are statistically meaningful rather than due to random chance. This leads to more accurate and reliable models.

5. **Give one example where hypothesis testing could be used to validate a feature's importance in a dataset.**

Example:

Suppose you have a dataset predicting house prices, and you want to check if the number of bedrooms significantly affects the price.

- H_0: The number of bedrooms has no effect on house prices.
- H_1: The number of bedrooms significantly affects house prices.

Hypothesis testing (e.g., t-test or ANOVA) can determine if this feature is statistically important for the model.

6. **What does a P-value represent in hypothesis testing?**

The **P-value** represents the **probability of observing the data (or something more extreme) assuming the null hypothesis is true**.

- A **small P-value** (typically < 0.05) indicates that the observed result is **unlikely under H_0**, so we may **reject the null hypothesis**.
- A **large P-value** suggests there is **not enough evidence** to reject H_0.

7. **If the P-value is 0.03 and the significance level (α) is 0.05, should we reject the null hypothesis? Explain.**

   Yes, we should reject the null hypothesis because the P-value (0.03) is less than the significance level $\alpha = 0.05$.
   This means the observed result is unlikely under the null hypothesis, providing enough evidence to support the alternative hypothesis.

8. **In ML, how could a low P-value influence feature selection during exploratory data analysis (EDA)?**

   In ML, a low P-value indicates that a feature has a statistically significant relationship with the target variable. During EDA, such features are considered important and likely useful for the predictive model, while features with high P-values may be less relevant and could be dropped to simplify the model.\

9. **Why can't we directly interpret the P-value as the probability that the null hypothesis is true?**

   We can't interpret the P-value as the probability that the null hypothesis is true because the P-value assumes H_0 is already true. It only measures the likelihood of observing the data (or more extreme results) under H_0, not the probability that H_0 itself is correct.

10. Give a real-world dataset example where the P-value might lead to a wrong conclusion if sample size is too large.

If the sample size is too large (e.g., millions of users in an A/B test for a website button color), even a tiny, practically meaningless difference (like a 0.1% click-through increase) can yield a very low P-value, leading to the wrong conclusion that the change is important when it's not practically significant.

11. Define Type I error and give a real-world example in a machine learning context.

Type I Error (False Positive): Rejecting the null hypothesis when it is actually true.
Real-world ML example: In a spam detection model, a Type I error occurs when the model predicts a normal email as spam (false alarm).

12. Define Type II error and explain its consequences in a fraud detection model.

Type II Error (False Negative): Failing to reject the null hypothesis when it is actually false (in ML terms → the model misses something that actually exists).

🔹 In a fraud detection model:
A Type II error occurs when the model predicts a fraudulent transaction as *normal (non-fraud)*.

Consequences:

- Fraudulent activities go undetected.
- The company may suffer financial losses.
- Customers may lose trust in the system.
- It can encourage fraudsters to exploit the loophole further.

In fraud detection, Type II errors are usually more dangerous than Type I errors.

13. How does changing the significance level (α) affect the probability of Type I and Type II errors?

Changing the significance level (α) directly impacts Type I and Type II errors:

- Type I error (False Positive): Probability of Type I error = α.
  - If α increases (e.g., from 0.01 → 0.05), the chance of rejecting a true null hypothesis (false alarm) increases.
  - If α decreases, the chance of Type I error decreases.
- Type II error (False Negative): There is a trade-off with Type II error (β).
  - Lower α → reduces Type I error, but increases Type II error (harder to detect true effects).

- Higher α → increases Type I error, but reduces Type II error (easier to detect effects).

14. Why might a data scientist prefer a lower Type II error in medical diagnosis models?

A data scientist might prefer a lower Type II error in medical diagnosis models because a Type II error means failing to detect a disease when it is actually present (false negative).
Consequences: The patient would not receive necessary treatment, which could worsen their condition or even be life-threatening.
Therefore, minimizing Type II error is crucial to ensure that sick patients are correctly diagnosed and treated in time.

15. Explain the difference between one-tailed and two-tailed tests with examples relevant to model evaluation metrics.

- **One-tailed test:** Tests if a model's performance metric (e.g., accuracy) is significantly **greater (or less)** than a benchmark.
  Example: Testing if a new model's accuracy is **greater than 85%**.
- **Two-tailed test:** Tests if a model's performance metric is **different** (could be higher or lower) from a benchmark.
  Example: Testing if a model's accuracy is **different from 85%** (could be better or worse). ✅