ASSIGNMENT – STATISTICS ADVANCE PART 2

vk2011383@gmail.com

1. What is hypothesis testing in statistics?
   - Hypothesis testing is a statistical method used to make decisions or inferences about a population based on sample data. It helps determine whether there is enough evidence to reject a null hypothesis.

2. What is the null hypothesis, and how does it differ from the alternative hypothesis?
   - Null Hypothesis (H_0):
     - It represents the status quo, assumption of no effect, or no difference.
     - It is the hypothesis that we assume to be true unless there is strong evidence against it.
     - The goal of hypothesis testing is to determine whether there is enough evidence to reject H_0 .
     - Example: *"The average height of students in a school is 5.5 feet."*

       Mathematically: $H_0: \mu = 5.5$
   - Alternative Hypothesis (H_a *or* H_1):
     - It represents a statement that contradicts the null hypothesis.
     - It is what the researcher aims to prove.
     - If there is significant evidence against H_0, we reject H_0 in favor of H_a.
     - Example: "*The average height of students in a school is different from 5.5 feet*."

       Mathematically: $H_a: \mu \neq 5.5$ (Two-tailed test)

3. What is the significance level in hypothesis testing, and why is it important?
   – The **significance level** (α) is the probability of rejecting the **null hypothesis (H_0)** when it is actually true. It is the threshold for determining statistical significance in a test.
   - *Why is the Significance Level Important?*
     - **Controls Type I Error** – A lower α reduces the risk of incorrectly rejecting H_0 (false positive).
     - **Determines Decision Criteria** – If the p-value is less than α, we reject H_0; otherwise, we fail to reject it.
     - **Balances Risk vs. Confidence** – A smaller α increases confidence in results but may require larger sample sizes to detect an effect.

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

- A **p-value** (probability value) is the probability of obtaining test results at least as extreme as the observed data, assuming the **null hypothesis (H_0)** is true.
- Small p-value (≤α) → Strong evidence against H_0.
- Large p-value (>α) → Weak evidence against H_0.

5. How do you interpret the P-value in hypothesis testing?
   - Small p-value (≤α) → Strong evidence against H_0:
     - We **reject H_0** in favor of the alternative hypothesis (H_a).
     - This suggests the observed effect is statistically significant.
   - Large p-value (>α) → Weak evidence against H_0.
     - We **fail to reject H_0**.
     - This means there is not enough evidence to support H_a, but it does **not** prove H_0 is true.

6. What are Type 1 and Type 2 errors in hypothesis testing?
   - Type I Error (False Positive):
     - **Definition**: Rejecting H_0 when it is actually true.
     - **Consequence**: Concluding there is an effect when none exists.
     - **Probability**: Equal to the significance level (α), often set at 0.05 (5%).
     - **Example**: A drug trial incorrectly concludes a new medicine works when it actually does not.
   - Type II Error (False Negative):
     - **Definition**: Failing to reject H_0 when it is actually false.
     - **Consequence**: Missing a real effect.
     - **Probability**: Denoted by β, where 1−β is the statistical power of the test.
     - **Example**: A test fails to detect that a new medicine works when it actually does.

7. What is the difference between a one-tailed and a two-tailed test in hypothesis testing?
   - A **one-tailed test** is used when the research hypothesis predicts a specific direction of the effect (e.g., greater than or less than).
     - The alternative hypothesis (H_a) is directional:
       a. **Right-tailed test**: $H_a: \mu > \mu_0$ (testing if the mean is **greater**).
       b. **Left-tailed test**: $H_a: \mu < \mu_0$ (testing if the mean is **less**).
     - The rejection region is on one side of the distribution.
   - A **two-tailed test** is used when the research hypothesis does **not** predict a specific direction; it just tests for any difference (either greater or less).
     - The alternative hypothesis (H_a) is non-directional:

$H_a: \mu \neq \mu_0$ (testing if the mean is different, either higher or lower).

- The rejection region is split **between both tails** of the distribution.

8. What is the Z-test, and when is it used in hypothesis testing?
   - A **Z-test** is a statistical test used to determine whether there is a significant difference between sample and population means (or between two sample means) when the **population variance is known** and the **sample size is large (typically n≥30)**.
   - *When is a Z-Test Used?*
     - The sample size is **large** (n≥30) or the population follows a normal distribution.
     - The **population variance (σ^2) is known**.
     - The data is **continuous** (e.g., heights, weights, test scores).

9. How do you calculate the Z-score, and what does it represent in hypothesis testing?
   - For a **sample mean** ($\bar{x}$):

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Where:

- $\bar{x}$ = Sample mean
- μ = Population mean (hypothesized mean)
- σ = Population standard deviation
- n = Sample size
- $\sigma/\sqrt{n}$ = **Standard error of the mean (SEM)**

   - For a **single data point** (X):

$$Z = \frac{\mathrm{X} - \mu}{\sigma}$$

Where:

- X = Individual data point
- μ = Population mean
- σ = Population standard deviation

   - *What Does the Z-Score Represent?*
     - Measures Distance from the Mean:
       a. A Z-score of **0** means the value is exactly at the population mean.
       b. A positive Z-score (+Z) means the value is **above** the mean.
       c. A negative Z-score (−Z) means the value is **below** the mean.
     - Used for Probability and Hypothesis Testing:
       a. The Z-score helps determine how extreme a sample mean is under H_0.

b. It is compared to critical values from the **standard normal distribution**.

10. What is the T-distribution, and when should it be used instead of the normal distribution?
    - The **T-distribution** (also called the **Student's t-distribution**) is a probability distribution used in statistical hypothesis testing when:
        - The sample size is **small** ($n < 30$).
        - The population **standard deviation (σ) is unknown** and is estimated from the sample.
    - It is similar in shape to the **standard normal distribution (Z-distribution)** but has **heavier tails**, meaning it accounts for more variability in small samples.

11. What is the difference between a Z-test and a T-test?

Feature	**Z-Test**	**T-Test**
When Used?	Large sample (n≥30)	Small sample ($n < 30$)
Population Standard Deviation (σ)	**Known**	**Unknown** (estimated from sample)
Distribution Used	**Standard Normal Distribution (Z-distribution)**	**T-Distribution** (heavier tails)
Degrees of Freedom (df)	Not required	$df = n - 1$
Application Examples	Testing population mean or proportion with large samples	Small sample mean comparison, paired tests, independent sample tests

12. What is the T-test, and how is it used in hypothesis testing?
    - The **T-test** is a statistical method used in hypothesis testing to determine whether there is a significant difference between the means of one or two groups. It is used when:
        - The sample size is **small** ($n < 30$).
        - The **population standard deviation (σ) is unknown** (so we estimate it from the sample).

- The data is **approximately normally distributed** (especially for small samples).

– **Steps to Conduct a T-Test in Hypothesis Testing**
  - **State the Hypotheses:**
    a. Null Hypothesis (H_0): No difference ($\mu 1 = \mu 2$)
    b. Alternative Hypothesis (H_a): There is a difference ($\mu 1 \neq \mu 2, \mu 1 > \mu 2, or\ \mu 1 < \mu 2$).
  - **Choose a Significance Level ($\boldsymbol{\alpha}$)**:
    Common values: **0.05 (5%)** or **0.01 (1%)**.
  - **Calculate the T-Statistic** using the appropriate formula.
  - **Find the Critical Value or P-Value** using the T-distribution table or software.
  - **Compare & Make a Decision**:
    a. If $|\boldsymbol{T}| > \boldsymbol{critical\ value}$, reject H_0.
    b. If $\boldsymbol{p - value} < \boldsymbol{\alpha}$, reject H_0.

13. What is the relationship between Z-test and T-test in hypothesis testing?
  – A Z-test is used when the population variance is known and the sample size is large ($n > 30$). It assumes normality in the population distribution.
  – A T-test is used when the population variance is unknown, and the sample size is small ($n \leq 30$). It uses the sample standard deviation as an estimate of the population standard deviation.
  – Both tests determine whether a sample mean significantly differs from a known value or another sample mean. The choice depends on data characteristics and sample size.
  – The T-test is more flexible for smaller samples, while the Z-test is preferred for larger samples with known variance.

14. What is a confidence interval, and how is it used to interpret statistical results?
  – A confidence interval (CI) is a range of values that likely contain a population parameter with a certain level of confidence (e.g., 95%). It is calculated as:
$$CI = Sample\ Mean \pm (Critical\ Value) \times Standard\ Error)$$
  – A wider interval indicates greater uncertainty, while a narrower one suggests more precise estimation. CI helps assess statistical results' reliability; if a CI does not contain a hypothesized value (e.g., zero in regression), the result is statistically significant.

15. What is the margin of error, and how does it affect the confidence interval?
  – The margin of error (MoE) measures the maximum expected difference between a sample estimate and the true population parameter. It is given by:
$$MoE = Critical\ Value \times Standard\ Error$$

- A larger MoE results in a wider confidence interval, increasing uncertainty. It is influenced by sample size, confidence level, and data variability. A higher confidence level or more variability increases MoE, while a larger sample size reduces it, leading to a more precise estimate.

16. How is Bayes' Theorem used in statistics, and what is its significance?
    - Bayes' Theorem calculates the probability of an event based on prior knowledge and new evidence. It is given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

      - where $P(A \mid B)$ is the probability of A given B.
    - It is significant in fields like machine learning, medical diagnosis, and spam detection. Unlike frequentist methods, Bayesian statistics incorporate prior beliefs, allowing dynamic probability updates based on new data.

17. What is the Chi-square distribution, and when is it used?
    - The Chi-square distribution is a right-skewed distribution used for hypothesis testing with categorical data. It is applied in:
      - **Goodness-of-fit tests** (checking if observed data follows an expected distribution).
      - **Independence tests** (analyzing relationships between categorical variables).

18. What is the Chi-square goodness-of-fit test, and how is it applied?
    - The Chi-square goodness-of-fit test determines if an observed categorical frequency distribution matches an expected distribution.
    - The formula is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

      where O and E are observed and expected frequencies. A high Chi-square value indicates a significant difference. It is used in genetics, marketing, and survey analysis to validate theoretical distributions.

19. What is the F-distribution, and when is it used in hypothesis testing?
    - The F-distribution is a right-skewed probability distribution used to compare variances between two populations. It appears in:
      - **ANOVA (Analysis of Variance):** Testing differences among group means.
      - **Regression analysis:** Assessing model significance.

      The F-test statistic is calculated as:

$$F = \frac{\text{Variance of group 1}}{\text{Variance of group 2}}$$

- A higher F-value suggests significant variance differences, leading to the rejection of the null hypothesis.

20. What is an ANOVA test, and what are its assumptions?
    - ANOVA (Analysis of Variance) determines whether means of three or more groups are significantly different. Assumptions include:
        - **Normality:** Data should be normally distributed.
        - **Independence:** Observations must be independent.
        - **Homogeneity of variances:** All groups should have similar variances.
    - ANOVA helps identify significant mean differences but requires post-hoc tests (e.g., Tukey's test) to determine which groups differ.

21. What are the different types of ANOVA tests?
    - **One-way ANOVA:** Compares means of one categorical factor across multiple groups.
    - **Two-way ANOVA:** Examines two independent factors' effects on a dependent variable.
    - **Repeated Measures ANOVA:** Compares means of the same subjects over time or under different conditions.

22. What is the F-test, and how does it relate to hypothesis testing?
    - The F-test is used to compare variances or test multiple group differences in ANOVA. It is calculated as the ratio of variances:
    $$F = \frac{\text{Variance of one sample}}{\text{Variance of another sample}}$$
    - A high F-value indicates significant differences, leading to rejection of the null hypothesis. It is crucial in validating assumptions of equal variances in regression and ANOVA.