Contents

Sujan prodhan CSE-29 fb.com/prodhan24 github.com/prodhan2

Theory of Statistics -2021

Part-A

1. a) Define with example
   (i) Population and Sample,
   (ii) Parameter and Statistic.

A population is the collection of all outcomes,responses,measurement or counts that are interest

A sample is a subset of a population

Statistics is the science of collecting ,organizing ,analyzing and interpreting data in order to make decisipons

In statistics, **a "parameter"** is a numerical value that describes a characteristic of a population. Parameters are key aspects of the population's distribution, and they provide important information about the population as a whole. Parameters are typically denoted using Greek letters.

Some common examples of parameters include:

1. Population Mean (μ or mu):

2. Population Variance (σ^2 or sigma squared):
3. Population Standard Deviation (σ or sigma):
4. Population Proportion (π or pi):.

b) Define Chi-square sampling distribution. Mention some important properties and uses of Chi- square distribution.

A chi-square distribution is a continuous probability distribution. The shape of a chi-square distribution depends on its degrees of freedom, k. The mean of a chi-square distribution is equal to its degrees of freedom (k) and the variance is 2k. The range is 0 to ∞.

Here are some important properties and uses of the chi-square distribution:

- The chi-square distribution is a continuous probability distribution. This means that it can take on any value between 0 and infinity.
- The chi-square distribution is non-negative. This means that the value of the chi-square statistic can never be negative.
- The shape of the chi-square distribution depends on the number of degrees of freedom. The degrees of freedom is a parameter of the chi-square distribution that determines its shape.
- The mean of the chi-square distribution is equal to its degrees of freedom.
- The variance of the chi-square distribution is equal to 2 times its degrees of freedom.
- The chi-square distribution approaches the normal distribution as the number of degrees of freedom increases.

c) Find MGF and CGF of Chi-square distribution. Hence find mean variance, B, and B. KHATAY ACHE

2**. a) Define with examples (i) Unbiased estimate (ii) Consistent estimate.**

Unbiased estimate

An unbiased estimate is an estimator whose **expected value** is equal to the true population parameter. For example, the **sample mean** is an unbiased estimator of the **population mean**.

Example:

The sample mean is calculated by taking the sum of all the observations in a sample and dividing by the number of observations. The expected value of the sample mean is equal to the population mean, so the sample mean is an unbiased estimator of the population mean.

- **Biased estimator** is an estimator that systematically overestimates or underestimates the true value of the parameter being estimated. **The sample mean is a biased estimator of the population mean if the population is not normally distributed.**

Consistent estimate

An estimator $T_1 = t(x1, x2, ...,xn)$ based on a random sample of size n is said to be consistent estimator of y, if T converges to y in probability.

Example:

As the sample size increases, the sample mean will get closer and closer to the population mean. This is because the sample mean is based on more and more observations, which reduces the amount of random variation in the estimate.

b) What do you mean by sufficient estimator? What is Factorization theorem?

A sufficient statistic is a statistic that contains all the information about the parameter being estimated. In other words, knowing the sufficient statistic is as good as knowing the entire sample.

Factorization Theorem:

The Factorization Theorem states that the joint probability distribution of a set of random variables $X_1, X_2, ..., X_n$, which depends on a parameter θ, can be factored into two functions:

$P(X_1, X_2, ..., X_n; \theta) = g(T(X_1, X_2, ..., X_n), \theta) * h(X_1, X_2, ..., X_n)$

In this equation:

- $X_1, X_2, ..., X_n$ are random variables representing the data.
- θ is the parameter you're trying to estimate.
- $T(X_1, X_2, ..., X_n)$ is a statistic derived from the data.
- g() is a function that depends only on the statistic $T(X_1, X_2, ..., X_n)$ and the parameter θ.
- h() is a function that depends only on the data $X_1, X_2, ..., X_n$.

c) Let X1, X2, X, be a random sample from N (μ, σ^2.) population. Find sufficient estimator μ and σ^2.

Sufficient Estimator for μ (Population Mean): $T_1(X_1, X_2, ..., X_n) = \bar{x}$

Sufficient Estimator for σ^2 (Population Variance): $T_2(X_1, X_2, ..., X_n) = (\bar{x}, s^2)$

These estimators are sufficient statistics for estimating the population mean (μ) and population variance (σ^2) based on the sample data $X_1, X_2, ..., X_n$

3. a**) What is point estimation? What are the properties of a good estimator?**

point estimation, in statistics, the process of finding an approximate value of some parameter—such as the mean (average)—of a population from random samples of the population

1. **Unbiasedness**: The estimator's average estimate equals the true value.
2. **Efficiency**: It gives specific estimates with small variability.
3. **Consistency**: As more data is collected, it gets closer to the true value.
4. **Sufficiency**: Contains all necessary information in the data for estimation.

b) **Mention some properties of maximum likelihood function.**

Maximum Likelihood Estimation (MLE) is a widely used statistical estimation method. In this lecture, we will study its properties: **efficiency, consistency and asymptotic normality.**

c) Suppose X1, X2, Xn be a random sample of size n from polsson distribution with parameter 0. Obtain the MLE of @ and show that the estimator is unbiased.

Part-B

4. a) What do you mean by contingency table?

In statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research.

What are the uses of such table?

They are heavily used in survey research, business intelligence, engineering, and scientific research.

b) For a 2x2 contingency table prove that Chi-square test of independence gives khatay ache

$N(ad-bc)2\ x^2$

$(a+c)(a+b)(c+d)(b+\alpha)$

N=a+b+c+d.

c)

In an experiment on immunization of cattle from tuberculosis, the following results are obtained

	Affected	Not affected
Inoculated	10	35
Not inoculated	13	6

Examine whether vaccination control diseases using the critical value 3.29 at $\alpha = 0.05$.

Khatay ache..

a)Explain Type-I and Type-II errors.

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population;

a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

b)What is power of a test? Explain the different steps to construct statistical test of hypothesis.

The power of a test is the probability of correctly rejecting the null hypothesis when it is false. It is denoted by 1 – β, where β is the probability of making a type II error. The power of a test depends on the sample size, the effect size, and the significance level.

c) Jeffrey, as an eight-year old, established an average time of 16.43 seconds for swimming the 25- yard freestyle, with a standard deviation of 0.8 seconds. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster by using goggles. Frank bought Jeffrey a new pair of expensive 25-yard freestyle are normal.

It seems like you provided some information about Jeffrey's swimming time and the standard deviation, but there is no specific hypothesis stated to perform a statistical test. In order to test whether Frank's claim is true or not, we need to set up a specific hypothesis to compare Jeffrey's swimming time with and without goggles.

Let's formulate a hypothesis to test Frank's claim:

Null hypothesis (H0): The average swimming time of Jeffrey with goggles (μ_goggles) is equal to 16.43 seconds. Alternative hypothesis (H1): The average swimming time of Jeffrey with goggles (μ_goggles) is faster than 16.43 seconds.

Now, to test the hypothesis, we can perform a one-sample z-test for the population mean. The formula for the z-test statistic is:

$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

where: $\bar{x}$ = Sample mean (average swimming time with goggles) μ = Population mean under the null hypothesis (16.43 seconds) σ = Population standard deviation (0.8 seconds) n = Sample size (number of trials)

Since we don't have information on the sample mean with goggles or the sample size, we cannot perform the actual calculation. However, the procedure would involve collecting a sample of Jeffrey's swimming times with the goggles and calculating the sample mean and sample standard deviation. Then, we would calculate the z-score using the formula above.

Next, we would compare the calculated z-score with the critical value from the standard normal distribution corresponding to the chosen significance level (e.g., 0.05 for a 5% level of significance) for a one-tailed test. If the calculated z-score is greater than the critical value, we would reject the null hypothesis in favor of the alternative hypothesis, indicating that Jeffrey swims faster with goggles. Otherwise, we would fail to reject the null hypothesis, suggesting no significant difference in swimming times with goggles.

Keep in mind that this is a general outline of the procedure, and specific data would be required to perform the actual test.

5.a) What is Non-parametric test of hypothesis? How does it differ from parametric test .

In statistics, a non-parametric test is a hypothesis test that **does not make any assumptions** about the **distribution of the population** from which the sample was drawn. This means that the test can be used even if the data is not normally distributed.

Parametric tests are make on assumptions **about the distribution of the population**, such as the assumption that the data is normally distributed. If these assumptions are not met, the parametric test may not be accurate.

Non-parametric tests are more robust to violations of assumptions than parametric tests. This means that they are less likely to be inaccurate if the data is not normally distributed.

b) Explain the advantages and disadvantages of Non-parametric test over parametric test.

c) Discuss about median test stating usual assumption.

goggles and timed Jeffrey for 15 25-yard freestyle swims. For the 15 swims, Jeffrey's average time was 16 seconds. Frank thought that the swim goggles helped Jeffrey to swim faster than the 16.43 seconds. Conduct a hypothesis test using a preset a = 0.05. Assume that the swim times for the

Theory of Statistics -2020

Section-A

1 a)**What do you mean by simple random sampling? Illustrate with an example.**

1. **Definition**: Simple random sampling is a method of selecting a subset (sample) from a larger population in such a way that every individual or item in the population has an equal chance of being included in the sample.
2. **Process**:
    - Assign a unique identifier (like a number) to each member of the population.
    - Use a random method (e.g., drawing lots, using a random number generator) to select the desired sample size from the population.
3. **Example**:
    - Population: All students in a school (1000 students).
    - Assign each student a number from 1 to 1000.
    - Use a random number generator to select 100 students.
    - The chosen 100 students represent a simple random sample.

b) Suppose we select a random sample of size n from a population and have independent random variables X1, X2,..., X. Do those random variables have the same distribution? Explain.

Explanation:

1. Each random variable X_i corresponds to an individual in the sample and takes on a value from the population.

2. While these random variables are related (as they come from the same sample), they are distinct because they correspond to different individuals in the sample, and each individual may have different characteristics.
3. Therefore, the distribution of each X_i depends on the underlying population distribution and the specific sampling process. As a result, these random variables typically have different distributions, unless the population distribution is such that all individuals are identical (in which case they would have the same distribution).

c) Suppose that of six students, the first student has $1, the second student has $2, the third student has $3, the fourth student has $4, the fifth student has $5, and the sixth student has $6. Consider the 1, 2, 3, 4, 5, 6 dollars as the population. A sample of two students is selected. What is the probability that the sample mean is equal to $3.5? Write down every step of your calculation.

Step 1: Calculate the total number of possible samples of two students from a population of six.

- You can use combinations (n choose k) to calculate this. In this case, it's 6 choose 2.
- Formula: C(n, k) = n! / (k!(n - k)!), where n is the total number of students (6) and k is the number of students you want to select (2).

C(6, 2) = 6! / (2!(6 - 2)!) = (6 * 5) / (2 * 1) = 15

So, there are 15 possible samples of two students.

Step 2: Determine the number of ways you can get a sample mean of $3.5.

- To get a sample mean of $3.5, you need to select one student with $3 and another student with $4.
- There are 2 ways to select a student with $3 (student 3 or student 4).
- Once you've chosen a student with $3, there's only 1 way to select a student with $4 (the remaining student).
- So, there are 2 * 1 = 2 ways to get a sample mean of $3.5.

Step 3: Calculate the probability.

- Probability = (Number of favorable outcomes) / (Total number of possible outcomes)
- Number of favorable outcomes = 2 (from Step 2)
- Total number of possible outcomes = 15 (from Step 1)
- Probability = 2 / 15

2.a) Write some names of statistics that can be used as estimator of the population.

1. **Sample Mean ($\bar{x}$)**: An estimator for the population mean (μ).
2. **Sample Proportion ($\hat{p}$)**: An estimator for the population proportion (p).
3. **Sample Variance (s^2)**: An estimator for the population variance (σ^2).

4. **Sample Standard Deviation (s)**: An estimator for the population standard deviation (σ).
5. **Sample Median**: An estimator for the population median.
6. **Sample Range**: An estimator for the population range.
7. **Sample Correlation Coefficient (r)**: An estimator for population correlation.
8. **Sample Regression Coefficients**: Estimators for population regression parameters (slope and intercept).
9. **Sample Percentile**: Estimators for population percentiles (e.g., 25th percentile, 75th percentile).

b) What are the criteria for a good estimator?

1. **Unbiasedness: On average, it provides estimates that are equal to the true population parameter.**
2. **Efficiency: It offers precise estimates with minimal variability.**
3. **Consistency: As sample size increases, it approaches the true parameter value.**
4. **Sufficiency: Contains all relevant information about the parameter in the data.**

c) Is 'sample mean' a good estimator in terms of unbiasedness? Explain.

Sample Mean's Unbiasedness:

1. Unbiasedness means an estimator's average estimate equals the true population parameter.
2. The sample mean ($\bar{x}$) is unbiased for the population mean (μ).
3. Mathematical proof shows $E(\bar{x}) = \mu$, meeting the unbiasedness criterion.
4. This property makes $\bar{x}$ a reliable estimator for μ in data analysis.

d) Is 'median' a consistent estimator? Explain with example.

Consistency of the Median:

1. Consistency means an estimator becomes more accurate with larger samples.
2. The sample median ($\hat{M}$) is consistent for the population median (M).
3. As you collect more data (larger sample size), $\hat{M}$ gets closer to the true median (M).
4. Example: Estimating the median income in a city—larger samples give more accurate estimates.
5. The sample median converges to the true median with increasing sample size.

3.a) Define F-distribution with degrees of freedom n1 and n2, and point out its uses?

The F-

Distribution	Definition	Relationship
Chi-square	X^2 ~ chi-square(df)	The square of a standard normal distribution N(0,1) follows a chi-square distribution with one degree of freedom, i.e., X^2 ~ chi-square(1). The sum of the squares of k independent standard normal distributions follows a chi-square distribution with k degrees of freedom, i.e., $X1^2 + X2^2 + ... + Xk^2$ ~ chi-square(k).
t-distribution	t ~ t(df)	A t-distribution is defined as the ratio of a standard normal distribution N(0,1) divided by the square root of a chi-square distribution with df degrees of freedom, i.e., t = N(0,1) / sqrt(X^2 / df). When the sample size is small, we use the t-distribution instead of the standard normal distribution to test hypotheses about the mean of a population.
F-distribution	F ~ F(df1, df2)	The F-distribution is defined as the ratio of two independent chi-square distributions, each divided by their degrees of freedom, i.e., F = (X1 / df1) / (X2 / df2). The F-distribution is commonly used in ANOVA and regression analysis to test hypotheses about the equality of variances and the significance of regression models.

b) Discuss the relationship among F, t and x^2 distribution.

F-Distribution:

1. Used for comparing variances in hypothesis testing (ANOVA).
2. Involves two degrees of freedom: one for the numerator and one for the denominator.
3. Right-skewed and starts from zero.

T-Distribution:

1. Used when the population standard deviation is unknown.
2. Defined by a single parameter: degrees of freedom.
3. Bell-shaped and has heavier tails than the standard normal distribution.

Chi-Squared (χ^2) Distribution:

1. Used for testing population variances.
2. Defined by a single parameter: degrees of freedom.
3. Right-skewed and starts from zero.

Relationships:

- F-distribution involves ratios of chi-squared variables.
- Taking the square root of an F-variable leads to a t-distribution.

Section-B

4. a) What do you mean by tests of good-ness of fit?

Tests of Goodness of Fit:

1. **Definition**: Tests of goodness of fit are statistical procedures used to assess how well an observed data set fits a theoretical probability distribution or model.
2. **Purpose**: These tests determine if the observed data significantly deviates from the expected distribution.
3. **Common Scenarios**: Used when you want to check if your data follows a specific probability distribution, such as normal, exponential, or Poisson.
4. **Procedure**: Compare the observed data to the expected distribution using statistical tests like the chi-squared test or the Kolmogorov-Smirnov test.
5. **Outcome**: Results help you decide whether your data fits the chosen theoretical distribution or if there are significant discrepancies.
6. **Applications**: Used in various fields, including biology, finance, and quality control, to validate models and assumptions.

b). Suppose a coin is tested 60 times with the results given in the following Table

Event ni

Heads 32

Tails 28

Is this a fair coin? The null hypothesis is H_1: ni= npi

where ni,is the number of observed frequencies and npi is the expected frequencies in which pi, are the theoretical probabilities. Given that, a = 5%.

The expected frequencies can be calculated as (n/2) = 30 for both heads and tails, where n is the total number of trials. We can then calculate the test statistic using the formula: $\chi^2 = \Sigma [(ni - np)^2 / np]$

Plugging in the values from the table, we get: $\chi^2 = [(32-30)^2/30] + [(28-30)^2/30] = 0.8$

To determine whether the test statistic is significant at the 5% levb el, we can compare it to the critical value of the chi-square distribution with one degree of freedom (since there are two categories: heads and tails). Using a chi-square distribution table or calculator, we find that the critical value at the 5% level is 3.84.

Since our calculated test statistic (0.8) is less than the critical value (3.84), we fail to reject the null hypothesis. Therefore, we do not have sufficient evidence to conclude that the coin is unfair

c) Which statistical test is suitable to test equality of variances? Explain.

The statistical test that is commonly used to test the equality of variances between two populations is the F-test. This test compares the ratio of the variances of the two populations and tests whether it is significantly different from 1.

The null hypothesis for the F-test is that the two populations have equal variances, while the alternative hypothesis is that the variances are not equal. The test statistic is calculated as the ratio of the sample variances, where the larger sample variance is in the numerator.

If the calculated F-statistic is greater than the critical value from an F-distribution table, then we reject the null hypothesis and conclude that the variances are not equal. On the other hand, if the calculated F-statistic is less than the critical value, then we fail to reject the null hypothesis and conclude that there is not enough evidence to support the claim that the variances are different.

It is important to note that the F-test assumes that the populations are normally distributed and have equal means. If these assumptions are violated, then other tests such as the Welch's t-test or the Levene's test may be used instead.

5. a) Define (1) null hypothesis (ii) critical region.

(i) Null hypothesis: The null hypothesis is a statement that represents that there is no difference or relationship between two or more variables of interest. It is typically denoted by H0 and is the opposite of the alternative hypothesis. The null hypothesis is tested using statistical methods, and its rejection or failure to reject leads to conclusions about the underlying population.

(ii) Critical region: The critical region is a range of values or a set of conditions that are used to reject the null hypothesis. It is determined based on the level of significance or the probability of making a Type I error (rejecting a true null hypothesis)

If the calculated test statistic falls within the **critical region**, then the null hypothesis is rejected, and the alternative hypothesis is accepted.

b) Describe how you would test the hypothesis Ho: $\sigma 1^2 = \sigma 2^2$,

To test the hypothesis Ho: $\sigma 1^2 = \sigma 2^2$, where $\sigma 1^2$ and $\sigma 2^2$ are the population variances of two independent populations, we can use the F-test for equality of variances. **The F-test is a statistical test** that compares the **ratio of the variances of two samples** with the expected ratio, assuming the null hypothesis is true. If the difference between the observed and expected ratios is significant, we reject the null hypothesis and conclude that the population variances are not equal.

To perform the F-test, we calculate the ratio of the sample variances as $F = s1^2 / s2^2$, where $s1^2$ and $s2^2$ are the sample variances of the two populations. We then compare this ratio to the critical value from the F-distribution

with (n1-1) and (n2-1) degrees of freedom, where n1 and n2 are the sample sizes of the two populations. If the calculated F-value is greater than the critical value, we reject the null hypothesis and conclude that the variances are not equal.

It is important to note that the F-test assumes that the populations are normally distributed and have equal means. If these assumptions are not met, other tests such as the Welch's t-test or the Mann-Whitney U test may be more appropriate.

c) The estimates of variances from two random samples of sizes 7 and 16 are 5.6 and 7.0 respectively Test whether the samples could have arisen from populations with the same sample variance.

To test whether the samples could have arisen from populations with the same variance, we will conduct an F-test using the given sample variances. The F-test will compare the variances of the two samples and determine if they are significantly different. Here are the steps to perform the test:

Step 1: State the hypotheses:

- Null Hypothesis (Ho): The population variances are equal ($\sigma1^2 = \sigma2^2$).
- Alternative Hypothesis (Ha): The population variances are not equal ($\sigma1^2 \neq \sigma2^2$).

Step 2: Select the significance level (α): Let's assume a significance level of $\alpha = 0.05$, which is a common choice.

Step 3: Given data: Sample 1 (n1 = 7): Sample variance (s1^2) = 5.6

Sample 2 (n2 = 16): Sample variance (s2^2) = 7.0

Step 4: Formulate the test statistic: The F-test statistic is calculated using the sample variances:

F = s1^2 / s2^2

Step 5: Determine the critical region: Find the critical values from the F-distribution table for $\alpha = 0.05$, degrees of freedom for the numerator (df1 = n1 - 1), and degrees of freedom for the denominator (df2 = n2 - 1).

For df1 = 6 and df2 = 15, the critical values are approximately F(**0.025**, 6, 15) ≈ 3.14 and F(0.975, **6, 15**) ≈ 0.23.

Step 6: Calculate the test statistic: Substitute the given values into the F-test formula:

F = 5.6 / 7.0 ≈ 0.8

Step 7: Make a decision: Compare the calculated F-test statistic with the critical values:

- **Since 0.23 ≤ 0.8 ≤ 3.14, the calculated F falls within the range of the critical values.**

- If the calculated F-test statistic falls within this range, we fail to reject the null hypothesis (Ho). This means that there is not enough evidence to conclude that the population variances are significantly different.

In conclusion, based on the F-test with a significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, we cannot conclude that the population variances are significantly different, and the samples could have arisen from populations with the same sample variance.

6.a) In what circumstances we need to perform non-parametric test?

Non-parametric tests are typically used in situations where the assumptions of parametric tests are violated, such as:

1. The data does not follow a normal distribution.
2. The sample size is small.
3. The data has outliers or extreme values.
4. The data is measured on an ordinal or nominal scale.
5. The variance of the population is unknown or unequal across groups.
6. The relationship between variables is not linear.

Non-parametric tests do not require normality, equal variances or linear relationships in the data, and are often more robust to outliers and extreme values. They are commonly used in social sciences, psychology, and healthcare research, where the sample size is often small and the data is often non-normal.

b) A particular shoe store believes that the median foot size of teenage boys is 10.25 inches. To test this hypothesis, the foot size of each of a random sample of 50 boys was determined. Suppose that 36 boys had sizes in excess of 10.25 inches. Does this disprove the hypothesis that the median size is 10.25? Use sign test to solve this problem.

The sign test is a non-parametric test used to test the median of a population. In this case, we can use the sign test to test the hypothesis that the median foot size of teenage boys is 10.25 inches.

To conduct the sign test, we need to determine the signs of the differences between the observed foot sizes and the hypothesized median. If the observed foot size is greater than the hypothesized median, the sign is positive, and if it is less, the sign is negative. If the observed foot size is equal to the hypothesized median, we can discard it.

In this problem, we have 36 boys with foot sizes greater than the hypothesized median, and 14 boys with foot sizes less than the hypothesized median. We can discard the 50th boy because his foot size is exactly the hypothesized median.

The null hypothesis is that the median foot size is 10.25 inches, and the alternative hypothesis is that the median foot size is greater than 10.25 inches.

We can now use the sign test to test this hypothesis. The test statistic is the number of positive signs. In this case, we have 36 positive signs and 14 negative signs. The probability of getting 36 or more positive signs in a sample of 49 is:

$P(X \geq 36) = 1 - P(X \leq 35)$

where X is a binomial random variable with n = 49 and p = 0.5. Using a binomial table or a calculator, we can find that:

$P(X \leq 35) = 0.0025$

Therefore, the probability of getting 36 or more positive signs is:

$P(X \geq 36) = 1 - P(X \leq 35) = 1 - 0.0025 = 0.9975$

This means that the probability of obtaining a result as extreme as 36 or more positive signs, assuming that the null hypothesis is true, is very low. Therefore, we can reject the null hypothesis and conclude that the median foot size of teenage boys is greater than 10.25 inches.

c) Discuss the purposes of run test and rank-sum test.

Certainly! Here are the purposes of the run test and rank-sum test explained in an easy-to-understand list:

Purposes of the Run Test:

1. To assess the randomness or lack of pattern in a sequence of data.
2. To determine if there are any systematic
3. shifts or trends in the data.
4. To test the hypothesis that the data is randomly distributed.
5. To identify any potential outliers or unusual patterns in the data.

Purposes of the Rank-Sum Test:

1. To compare two independent groups or samples and determine if there is a significant difference between their distributions.
2. To assess whether one group tends to have higher or lower values compared to the other group.
3. To test the hypothesis that the medians of the two groups are equal.
4. To analyze non-parametric data or data that does not meet the assumptions of parametric tests.

These tests serve different purposes but are both valuable tools in statistical analysis. The run test helps evaluate the randomness of data, while the rank-sum test compares two groups and assesses differences between them.

Theory of Statistics -2019

Section-A

1. (a) **What do you mean by sampling distribution? Write down the probability density function of Chi-square distribution with n degree of freedom (d.f).**

In statistics, a sampling distribution is a probability distribution that describes the frequencies of different possible outcomes of a random variable based on a sample of a population. It is essentially the distribution of statistics computed from multiple samples of the same size drawn from the same population.

The chi-square distribution is a continuous probability distribution that is widely used in statistics, particularly in hypothesis testing and confidence interval estimation. It is a special case of the gamma distribution, and its probability density function (pdf) is given by:

f(x) = (1 / (2^(n/2) * Γ(n/2))) * (x^(n/2-1) * e^(-x/2))

where x is the random variable, n is the degrees of freedom, Γ is the gamma function, and e is the mathematical constant e (approx. 2.71828).

The parameter n determines the shape of the chi-square distribution. As n increases, the distribution becomes more symmetric and concentrated around its mean. The mean and variance of the chi-square distribution are both equal to n.

(b) **If X has a Chi-square distribution with nd.f, find the MGF of Chi-square distribution.**

Sure, here is a table summarizing some important uses of the chi-square distribution in statistics:

Sure! To find the moment generating function (MGF) of a chi-square distribution with nd.f, we can use the definition of the MGF:

M(t) = E[e^(tX)]

where X is a random variable with the chi-square distribution with nd.f.

To find the MGF, we need to find the expected value E[e^(tX)]. We can do this by using the probability density function (pdf) of the chi-square distribution, which is:

f(x) = (1 / (2^(n/2) * Γ(n/2))) * (x^(n/2-1) * e^(-x/2))

where x is the random variable, n is the degrees of freedom, Γ is the gamma function, and e is the mathematical constant e (approx. 2.71828).

We can use this pdf to calculate the expected value of e^(tX):

E[e^(tX)] = $\int_0^{\infty}$ e^(tx) f(x) dx

Substituting the pdf into this equation, we get:

E[e^(tX)] = $\int_0^{\infty}$ e^(tx) * (1 / (2^(n/2) * Γ(n/2))) * (x^(n/2-1) * e^(-x/2)) dx

Simplifying this expression, we get:

$E[e^{(tX)}] = (1 / (2^{(n/2)} * \Gamma(n/2))) * \int_0^{\infty} x^{(n/2-1)} * e^{(-x/2 + tx)} dx$

Now, we can use the identity for the gamma function:

$\Gamma(k) = \int_0^{\infty} x^{(k-1)} * e^{(-x)} dx$

to simplify the integral. We can rewrite the integral in terms of the gamma function as follows:

$E[e^{(tX)}] = (1 / (2^{(n/2)} * \Gamma(n/2))) * \Gamma(n/2) * \int_0^{\infty} (2t)^{(-n/2)} * x^{(n/2-1)} * e^{(-(x/2 - tx))} dx$

Simplifying further, we get:

$E[e^{(tX)}] = (1 / (1 - 2t)^{n/2})$

This is the MGF of the chi-square distribution with nd.f.

I hope this helps!

(c) **Write down important uses of Chi-square distribution.**

Use of Chi-square Distribution	Description
Hypothesis Testing	Used to test hypotheses about the population variance or the independence of categorical variables in a contingency table
Goodness-of-Fit Tests	Used to determine whether a sample of data comes from a population with a specific distribution
Confidence Intervals	Used to construct confidence intervals for the population variance
Regression Analysis	Used to test the overall significance of a regression model and to test for the significance of individual predictors in the model
Multivariate Analysis	Used in multivariate analysis, such as principal component analysis and factor analysis, to test the significance of the components or factors
Genetics	Used in genetics to test for the goodness of fit of observed genetic ratios to expected ratios based on Mendelian genetics

I hope this table helps to summarize the important uses of the chi-square distribution in statistics.

2. (a) **What is likelihood function?**

The likelihood function is a function of the parameters of a statistical model that measures how likely the observed data is under those parameters. It is used in statistical inference to estimate the parameters of the model.

Let's say we have a statistical model with parameters θ. We have some data, x, that we want to use to estimate the parameters of the model. The likelihood function, L(θ|x), is defined as the probability of the data, x, given the parameters, θ. In other words, it is the probability of observing the data that we did, under the assumption that the parameters of the model are θ.

The likelihood function is a very important concept in statistical inference. It is used in a variety of methods, including maximum likelihood estimation, Bayesian inference, and hypothesis testing.

Here is an example of a likelihood function. Let's say we have a model for the height of a person, with parameters μ and σ. μ is the mean height and σ is the standard deviation. We have some data, x, which is a list of heights of people. The likelihood function is then:

$$L(\mu, \sigma|x) = (1/\sqrt{(2\pi\sigma^2)}) * \exp(-(x - \mu)^2 / (2\sigma^2))$$

(b) What are the principles of maximum likelihood estimation method?

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model by finding the values of the parameters that maximize the likelihood function. The likelihood function is a function of the parameters of the model that measures how likely the observed data is under those parameters.

The principles of maximum likelihood estimation are as follows:

The likelihood function is maximized by finding the values of the parameters that make the data as likely as possible.

The maximum likelihood estimates are often consistent, meaning that they converge to the true values of the parameters as the sample size increases.

The maximum likelihood estimates are often efficient, meaning that they have the smallest possible variance among all unbiased estimators.

The maximum likelihood estimates are often asymptotically normal, meaning that their distribution approaches a normal distribution as the sample size increases.

(c) Find maximum likelihood estimator of parameters when n samples are drawn from Normal distribution?

3. **(a) Define statistic, estimator and estimate with example.**

(b) Describe three properties of a good estimator, in short.

Sure, here is a table summarizing the three properties of a good estimator:

Property	Description
Unbiasedness	The expected value of the estimator is equal to the true value of the parameter being estimated. The estimator does not systematically over- or under-estimate the true value.
Efficiency	The estimator has a small variance compared to other estimators. The estimator produces estimates that are less variable than those produced by other estimators.
Consistency	The estimator's estimates converge to the true value of the parameter as the sample size increases. The estimator becomes more and more accurate as more data is collected.

These properties are important because they help us evaluate and compare different estimators, and choose the one that is best suited for a particular problem.

(c) Show that sample average x is an unbiased estimator of population mean, but sample variance $s^2 = (x-x)^2$ is a biased estimator of population variance.

Sure! Let's first define some notation:

- μ: population mean
- $\bar{x}$: sample mean
- σ^2: population variance
- s^2: sample variance
- n: sample size

We can show that $\bar{x}$ is an unbiased estimator of μ as follows:

- The expected value of $\bar{x}$ is given by $E(\bar{x}) = \mu$, by the definition of the sample mean.
- Therefore, $\bar{x}$ is an unbiased estimator of μ, since its expected value is equal to the true population mean.

However, s^2 is a biased estimator of σ^2. To show this, we need to calculate the expected value of s^2 and compare it to the true population variance σ^2:

- The expected value of s^2 is given by $E(s^2) = (n-1)/n * \sigma^2$, which is a well-known result in statistics. This means that the expected value of s^2 is slightly smaller than the true population variance σ^2.
- Therefore, s^2 is a biased estimator of σ^2, since its expected value is not equal to the true population variance.

The reason why s^2 is a biased estimator of σ^2 is because the formula for s^2 involves dividing by (n-1) instead of n, which introduces a small amount of bias. This is done to account for the fact that the sample mean $\bar{x}$ is also being estimated from the same data, and therefore introduces some additional uncertainty into the estimation of the sample variance. However, this correction does not completely remove the bias, and so s^2 is still a biased estimator of the population variance σ^2.

Section-B

4. (a) Distinguish between: i) Null hypothesis and alternative hypothesis

Sure! Here is a table summarizing the main differences between null hypothesis and alternative hypothesis:

Null Hypothesis	Alternative Hypothesis
The default assumption in a hypothesis test	The hypothesis that we are trying to prove or support
Assumes no effect or relationship between variables	Assumes there is an effect or relationship between variables
Denoted as H0	Denoted as Ha or H1
Tested against the alternative hypothesis	Tested against the null hypothesis
Usually a statement of **equality or no difference**	Can be a statement of **inequality, difference, or association**
Rejection of the null hypothesis supports the alternative hypothesis	Failure to reject the null hypothesis does not necessarily support the alternative hypothesis

I hope this table helps to distinguish between null hypothesis and alternative hypothesis in hypothesis testing!

ii)) Distinguish between: Type I error and Type II error

Sure! Here is a table summarizing the main differences between Type I error and Type II error:

Type I Error	**Type II Error**
Also known as a "false positive"	Also known as a "false negative"
Occurs when we reject a true null hypothesis	Occurs when we fail to reject a false null hypothesis
Probability denoted as alpha (α)	Probability denoted as beta (β)

Type I Error	Type II Error
Can be controlled by setting the level of significance (α)	Can be controlled by increasing sample size or decreasing the level of significance (α)
Involves the risk of drawing an incorrect conclusion	Involves the risk of not detecting a true effect or relationship
Can lead to incorrect decisions and waste of resources	Can lead to missed opportunities and ineffective interventions
Usually more serious than Type II error	Can be equally serious depending on the context

I hope this table helps to distinguish between Type I error and Type II error in hypothesis testing!

(b) **How does a t-test differ from a normal test?**

Sure! Here is a table summarizing the main differences between a t-test and a normal test, along with an easy explanation:

t-test	Normal Test
Used for small sample sizes (usually n < 30)	Used for large sample sizes (usually n > 30)
Assumes unknown population variance	Assumes known population variance
Test statistic follows a t-distribution	Test statistic follows a normal distribution
Calculates the estimated standard error of the mean	Uses the population standard deviation as the standard error
Tests hypotheses about the sample mean	Tests hypotheses about the population mean
More appropriate when the sample size is small	More appropriate when the sample size is large

In an easy explanation, a t-test is used when the sample size is small and the population variance is unknown, while a normal test is used when the sample size is large and the population variance is known. The test statistic used in a t-test follows a t-distribution, which is slightly different from the normal distribution used in a normal test. Additionally, a t-test calculates the estimated standard error of the mean to take into account the uncertainty introduced by having a small sample size. Overall, the choice of which test to use depends on the sample size and the assumptions made about the population variance.

(c) A sample of 900 items taken from population with standard deviation 2.61 cms. The mean of the sample is 3.4 cms. Test whether the sample come from the population with mean 3.25 cms. To be noted Pr{Z<-1.73)-Pr{Z>1.73)=0.089

- Null Hypothesis (Ho): The sample comes from a population with a mean of 3.25 cms ($\mu = 3.25$).

- Alternative Hypothesis (Ha): The sample does not come from a population with a mean of 3.25 cms ($\mu \neq 3.25$).

Step 2: Select the significance level (α): Let's assume a significance level of $\alpha = 0.05$.

Step 3: Given data: Sample size (n) = 900 Sample mean ($\bar{x}$) = 3.4 cms Population standard deviation (σ) = 2.61 cms Population mean (μ) = 3.25 cms

Step 4: Calculate the test statistic (Z-score): The Z-score is calculated using the formula:

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

where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size.

$Z = (3.4 - 3.25) / (2.61 / \sqrt{900}) \approx 1.73$

Step 5: Determine the critical region: Since the alternative hypothesis is two-tailed ($\mu \neq 3.25$), we need to find the critical Z-values that correspond to the significance level of $\alpha = 0.05$. The critical Z-values for a two-tailed test at $\alpha = 0.05$ are approximately -1.96 and 1.96.

Step 6: Make a decision: Compare the calculated Z-score with the critical values:

- If the calculated Z-score falls outside the critical region (-1.96 to 1.96), we reject the null hypothesis (Ho).
- If the calculated Z-score falls inside the critical region, we fail to reject the null hypothesis (Ho).

In this case, the calculated Z-score is approximately 1.6266, which falls outside the critical region. Therefore, we reject the null hypothesis.

Step 7: Conclusion: Based on the Z-test at a significance level of 0.05, there is enough evidence to conclude that the sample does not come from a population with a mean of 3.25 cms.

(a) What is critical region? How the alternative hypothesis affects in computing critical region?

(a) In hypothesis testing, the critical region is a range of values of the test statistic that would lead us to reject the null hypothesis. The critical region is determined by the significance level of the test, which is the probability of making a Type I error (rejecting the null hypothesis when it is true). The alternative hypothesis affects the computation of the critical region because it determines whether the test is one-tailed or two-tailed. In a one-tailed test, the critical region is either in the upper or lower tail of the distribution, depending on the direction of the alternative hypothesis. In a two-tailed test, the critical region is split between the upper and lower tails of the distribution.

For the given problem, the null hypothesis is that the sample mean is equal to 3.25 cms, and the alternative hypothesis is that the sample mean is greater than 3.25 cms. Since we are given that Pr{Z<-1.73)-Pr{Z>1.73}=0.089, we can use this information to find the critical value of the test statistic at the 5% level of significance.

Pr{Z<-1.73) = 0.0425 (from standard normal distribution table) Pr{Z>1.73} = 0.0425 Therefore, the critical value of the test statistic is 1.73.

The test statistic for this problem is:

z = (x - mu) / (sigma / sqrt(n)) = (3.4 - 3.25) / (2.61 / sqrt(900)) = 4.49

Since the test statistic is greater than the critical value, we reject the null hypothesis and conclude that the sample comes from a population with a mean greater than 3.25 cms.

(b) Describe different steps for testing a hypothesis of equality of two proportions.

Here are the different steps for testing a hypothesis of equality of two proportions:

1. **State the null and alternative hypotheses**: The null hypothesis is that the two proportions are equal, and the alternative hypothesis is that they are not equal.
2. **Choose the level of significance**: Determine the level of significance for the test, which will determine the critical value of the test statistic.
3. **Collect the data**: Collect random samples from both populations and calculate the sample proportions.
4. **Calculate the test statistic**: Calculate the test statistic for the difference in proportions between the two samples.
5. **Find the critical value and/or p-value**: Find the critical value of the test statistic at the chosen level of significance, or calculate the p-value for the test.
6. **Make a decision**: If the test statistic falls in the critical region or the p-value is less than the level of significance, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

For the given problem, the null hypothesis is that the proportion of men in favor of the proposal is equal to the proportion of women in favor of the proposal. The alternative hypothesis is that the proportions are not equal. The test statistic for this problem is a standard normal variable, and we are given that Pr{Z<-1.269}=Pr(Z>1.269}=1.02. Since the p-value is greater than the level of significance (5%), we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the proportions of men and women in favor of the proposal are different.

(c) Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. Among the respondents, 200 men and 325 women were in favor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are same. To be noted Pr{Z<- 1.269}=Pr(Z>1.269}=1.02.

To test the hypothesis that the proportions of men and women in favor of the flyover proposal are the same, we can use the two-sample proportion test.

Let p1 be the proportion of men in favor of the proposal, and p2 be the proportion of women in favor of the proposal. Then, the null hypothesis is that the two proportions are equal, i.e.

H0: p1 = p2. The alternative hypothesis is that the two proportions are not equal, i.e.

Ha: p1 ≠ p2.

We can use the Z-test for the difference between two proportions to test this hypothesis. The test statistic is:

z = (p1 - p2) / sqrt(p*(1-p)*(1/n1 + 1/n2))

where p = (x1 + x2) / (n1 + n2), x1 and x2 are the number of men and women in favor of the proposal, n1 and n2 are the sample sizes of men and women, respectively.

Using the given probabilities, we can find the critical value for the test statistic as z = ±1.96 at a significance level of 0.05.

Calculating the test statistic, we have:

p1 = 200/400 = 0.5

p2 = 325/600 = 0.5417

n1 = 400 n2 = 600

p = (200 + 325) / (400 + 600) = 0.4417

z = (0.5 - 0.5417) / sqrt(0.4417*(1-0.4417)*(1/400 + 1/600)) = -2.13

Since the test statistic (-2.13) falls within the critical region (z < -1.96), we reject the null hypothesis and conclude that the proportions of men and women in favor of the flyover proposal are not the same.

6.

(a) What is a contingency table? How does it differ from a correction table?

A contingency table is a table used to display the frequencies and proportions of the joint occurrence of two or more categorical variables. It allows for the visual inspection of the relationship between the variables, making it useful for analyzing and summarizing data in fields such as statistics, social sciences, and market research.

A contingency table is used to display the relationship between categorical variables, while a correction table is used to adjust for bias or confounding factors when making statistical inferences.

(b) What is the null hypothesis in a test of independence? Define the expression for the value of Chi-square in arxe contingency table.

,the null hypothesis is that there is no association between two categorical variables. In other words, the null hypothesis is that the distribution of one variable is independent of the distribution of the other variable. The alternative hypothesis is that there is an association between the two variables.

The expression for the value of Chi-square in a rxc contingency table is given by:

Chi-square = $\sum\sum [(Oij - Eij)^2 / Eij]$

where Oij is the observed frequency in the i-th row and j-th column, Eij is the expected frequency in the i-th row and j-th column under the null hypothesis, and the summation is taken over all the cells in the contingency table.

The expected frequency in each cell is calculated by:

Eij = (ri * cj) / n

where ri is the total frequency in the i-th row, cj is the total frequency in the j-th column, and n is the total sample size.

(c) Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the adjoining table. Examine whether the nature of the area is related to voting preference in the election.

The null hypothesis (H0) assumes that the nature of the area and voting preference are independent, while the alternative hypothesis (H1) assumes that they are dependent.

Let's set up the observed and expected frequency tables:

Observed Frequency Table:

Area	**Votes for A**	**Votes for B**	**Total**
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Expected Frequency Table (assuming independence):

To calculate the expected frequencies, we use the formula:

Expected Frequency (E) = (Row Total * Column Total) / Grand Total

Area	**Votes for A**	**Votes for B**	**Total**
Rural	585	415	1000
Urban	585	415	1000
Total	1170	830	2000

Now, we can perform the chi-square test using the following formula for each cell:

Chi-Square = Σ [(Observed Frequency - Expected Frequency)^2 / Expected Frequency]

Calculating the chi-square values for each cell:

For Rural, Votes for A: Chi-Square = [(620 - 585)^2 / 585] ≈ 1.191

For Rural, Votes for B: Chi-Square = [(380 - 415)^2 / 415] ≈ 2.115

For Urban, Votes for A: Chi-Square = [(550 - 585)^2 / 585] ≈ 1.191

For Urban, Votes for B: Chi-Square = [(450 - 415)^2 / 415] ≈ 2.115

Next, we sum up all the individual chi-square values:

Total Chi-Square = 1.191 + 2.115 + 1.191 + 2.115 ≈ 6.612

Now, we need to find the critical value of the chi-square test for a given significance level and degrees of freedom. Assuming a significance level of 0.05 and 1 degree of freedom (df = (Number of rows - 1) * (Number of columns - 1) = (2 - 1) * (2 - 1) = 1), the critical value is approximately 3.841.

Since the total chi-square value (6.612) is greater than the critical value (3.841**), we reject the null hypothesis.**

Conclusion: Based on the chi-square test of independence, we have sufficient evidence to conclude that the nature of the area is related to voting preference in the election.

Theory of Statistics -2018

Section-A

1. a) Define chi-square (7) variate and write down its pdf.

b) Show that distribution tends to normal distribution for large degrees of freedom.

e) Write down some important properties of distribution

2. a) Define F-statistic. Write down the p.d.f. of F-statistic by If F follows F-statistic on (ni, na) degrees of freedom and show that the statistics

(1+F) has beta distribution,

c) State 2 (two) main applications of F-statistic.

Two Main Applications of F-Statistic:

1. Analysis of Variance (ANOVA): The F-statistic is used to assess whether there are significant differences in means among multiple groups or treatments. It helps determine if a categorical independent variable has an impact on a continuous dependent variable.
2. Regression Analysis: In regression analysis, the F-statistic is employed to evaluate the overall significance of a regression model by comparing the fit of the full model to a reduced model (e.g., testing if all predictors are jointly significant).

3. a) What are the methods of point estimation?

Methods of Point Estimation:

1. **Method of Moments (MoM):** Match sample moments to population moments.

2. **Maximum Likelihood Estimation (MLE):** Find parameter that maximizes likelihood of observed data.
3. **Bayesian Estimation:** Combine prior beliefs with data to get parameter estimate.

b) Define estimate and estimator with examples.

b) **Estimate and Estimator:**

- **Estimate:** Single value representing a population parameter. Example: Mean height of sampled students (165 cm).
- **Estimator:** Formula or rule to calculate an estimate. Example: Sample mean ($\bar{x}$) estimates population mean (μ).

c) Let X1, X2,...,x, be a random sample from a poisson distribution with pdf given by **KHatay ace**

Find the maximum likelihood estimator of A. Show that the estimator is unbiased.

Section-B

4. a) What do you mean by statistical test of hypothesis? Define with examples (i) null hypothesis (ii) critical region and

(iii) level of significance..

a) **Statistical Test of Hypothesis:**

(i) **Null Hypothesis (H0):** The starting assumption with no effect or difference. Example: New drug is as effective as a placebo.

(ii) **Critical Region (Rejection Region):** Range of extreme values leading to null hypothesis rejection. Example: Values in the tails of a distribution beyond a set threshold.

(iii) **Level of Significance (α):** Predefined probability threshold for making Type I errors. Example: Accepting a 5% chance of wrongly rejecting the null hypothesis ($\alpha = 0.05$).

b) What is BCR? When does a test become MP test?

b) **BCR (Base Conversion Rate) and MP Test (Most Powerful Test):**

1. **BCR (Base Conversion Rate):** The conversion rate of the control group, used as a baseline in A/B testing.
2. **MP Test (Most Powerful Test):** A statistical test that maximizes the ability to detect real effects while minimizing the risk of missing them, making it the best choice for sensitivity in an analysis.

c) A bulb manufacturing company claims that the average longevity of their bulb is 3.65 years with a standard deviation of 0.16 years. A random sample of 36 bulbs gave a mean longevity of 4.45 years. Does the sample mean justify the claim of the manufacturer? [Use 5% level of significance)

1. **a) What do you mean by the power of a test?**

the power of a hypothesis test is the probability of **rejecting the null hypothesis when the alternative hypothesis is the hypothesis that is true.** **P(rejecting H_0 | H_1 True)**

b) Describe how you will test the following hypothesis Ho: p= p2…….. Pk (k>2).

To test the hypothesis Ho: p = p2 = ... = Pk (k > 2), where p, p2, ..., Pk are population proportions, you can use a **chi-square goodness-of-fit test**. This test compares the observed frequencies of a categorical variable to the expected frequencies under the null hypothesis.

Here's a step-by-step guide on how to perform the chi-square goodness-of-fit test for this hypothesis:

Step 1: State the null and alternative hypotheses:

Null hypothesis (Ho): p = p2 = ... = Pk (All population proportions are equal).

Alternative hypothesis (Ha): At least one population proportion is different from the others.

Step 2: Collect and organize the data:

You need observed frequencies for each category and the total sample size (n). The data should be categorical, divided into k categories.

Step 3: Calculate the expected frequencies:

Under the null hypothesis, where all population proportions are equal, the expected frequency for each category is given by the formula:

Expected Frequency (E) = (Total Sample Size * Proportion assumed under Ho for that category)

In this case, since all proportions are assumed to be equal, you can set E = n / k for each category.

Step 4: Calculate the chi-square test statistic:

Chi-Square = Σ [(Observed Frequency - Expected Frequency)^2 / Expected Frequency]

Sum the values of [(Observed Frequency - Expected Frequency)^2 / Expected Frequency] for all k categories.

Step 5: Determine the degrees of freedom (df):

df = k - 1 (where k is the number of categories).

Step 6: Find the critical value:

At a given significance level (e.g., 0.05), find the critical value from the chi-square distribution table with df degrees of freedom.

Step 7: Compare the test statistic with the critical value:

If the test statistic is greater than the critical value, reject the null hypothesis. If it is smaller, fail to reject the null hypothesis.

Step 8: Draw the conclusion:

Based on the comparison, draw a conclusion about the null hypothesis. If the null hypothesis is rejected, it suggests that at least one population proportion is different from the others.

Note: The chi-square goodness-of-fit test assumes that the expected frequency for each category is at least 5. If this assumption is not met, you might need to consider combining categories or using an alternative test.

c) A set of 8 correlation coefficients and the corresponding sample sizes are given below. Test the homogeneity of these coefficients.

Sample	Correlation Coefficient (r)	Sample Size (N)
1	0.231	10
2	0.464	12
3	0.539	8
4	0.357	15
5	0.628	18
6	0.136	11
7	0.204	10
8	0.461	7

6. a) What is contingency table? For a 2 x 2 contingency table

ab Where N = a+b+c+d.

cd

N(ad-bc)2 (a+b)(c+d)(a+c)(b+4)*

show that x=

Theory of Statistics -2017

Part -1

1.(a) Define chi-square (x2) variate and its p.d.f.

(b) By using moment generating function (MGF) find B1 and B2

(c) Find the mode of x-distribution. Mention some important properties and application of x-distribution.

a) **Chi-Square (χ^2) Variate and its Probability Density Function (p.d.f.):**

- χ^2 Variate: A random variable following the chi-square distribution.
- p.d.f. of χ^2: $f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2}$

b) **Finding B1 and B2 using Moment Generating Function (MGF):**

- B1: First moment, $\left.\frac{dM(t)}{dt}\right|_{t=0}$.
- B2: Second moment, $\left.\frac{d^2M(t)}{dt^2}\right|_{t=0}$.

c) **Mode of χ^2-Distribution, Properties, and Applications:**

- Mode: $\nu - 2$ (if $\nu > 2$).
- Properties: Positively skewed, unimodal, mean = ν, variance = 2ν.
- Applications: Hypothesis testing, confidence intervals, sums of squares, statistical analysis.

2.

(a) Define Student's t statistics. Formulate its sampling distribution.

(b) Show that Student's t distribution reduces to the standard normal distribution for large degrees of freedom.

3.

(a) Define point estimator. What are the criteria of good estimator?

1. Unbiasedness

2. Efficiency
3. Consistency
4. Sufficiency

point estimator is a statistic that provides a single numerical value as an estimate of an unknown population parameter based on sample data. The main purpose of a point estimator is to make inferences about the population parameter using information from a subset of the population, known as the sample.

For example, in statistics, if we want to estimate the population mean (μ) based on a sample, we can use the sample mean ($\bar{x}$) as a point estimator. Similarly, if we want to estimate the population proportion (p) based on a sample, we can use the sample proportion ($\hat{p}$) as a point estimator.

(b) What do you mean by sufficient statistic? Lot X1, X2, Xn be a random sample from a poison variate with parameter ' μ' then show that x' is a sufficient static for ' μ'.

(c) Let x1, x2 be a random sample of size 'n' from normal distribution with p.d.f is Find the MLE of ''and o -a<x<α,-a<μ<α, o^2 >0

Part-B

4.)

(a) Distinguish between Type 1 and Type 2 errors. Define: (i) Power of a test, (ii) Level of significance and (iii) Degree of freedom. Describe the procedure for Testing of Hypothesis.

(b) The coefficient of correlation obtained from a random sample of 20 pairs is 0.50. Test the 03

population correlation coefficient (p-0) at 5% level of significance. [40.05.18 = 2.10]

5.

(a) When do you use independent samples t-test?

Researchers are interested in the mean level of some enzyme in a certain population. They take a sample of 10 individuals, determine the level of enzyme in each and compute a sample mean 22. It is known that the variable of interest is approximately normally distributed with a variance of 45. Can you conclude that the mean enzyme level in this population is different from 25 at the 5% level of significance? [Zoos = 1.96].

Step 1: State the null and alternative hypotheses:

Null hypothesis (H0): The mean enzyme level in the population is equal to 25, μ = 25. Alternative hypothesis (H1): The mean enzyme level in the population is different from 25, μ ≠ 25.

Step 2: Set the significance level (α):

Choose a significance level (α) to determine the probability of making a Type I error. In this case, α = 0.05 (5% level of significance).

Step 3: Collect and summarize the data:

Given the sample size (n = 10), sample mean ($\bar{x}$ = 22), and population variance (σ^2 = 45), we can proceed to the next step.

Step 4: Calculate the test statistic (z-score):

The z-score formula for a one-sample z-test is given by:

$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

where: $\bar{x}$ = Sample mean (22) μ = Assumed population mean under the null hypothesis (25) σ = Population standard deviation (since σ^2 = 45, $\sigma = \sqrt{45} \approx 6.71$) n = Sample size (10)

Calculating the z-score:

$z = (22 - 25) / (6.71 / \sqrt{10}) \approx -1.25$

Step 5: Find the critical value:

At a 5% level of significance (α = 0.05), the critical value for a two-tailed test is approximately ±1.96. Since this is a two-tailed test, we need to consider both the positive and negative critical values.

Step 6: Make a decision:

Compare the absolute value of the calculated z-score from Step 4 with the critical value from Step 5.

$|z| = |-1.25| = 1.25$ $1.25 < 1.96$

Step 7: Draw the conclusion:

Since the absolute value of the calculated z-score (1.25) is less than the critical value (1.96), we fail to reject the null hypothesis (H0).

Conclusion: At the 5% level of significance, we do not have sufficient evidence to conclude that the mean enzyme level in the population is different from 25. The data does not provide enough support to reject the null hypothesis.

(b) For a simple random sample of adults, IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. A simple random sample of 13 statistics professors yields a standard deviation of s-7.2. Assume that IQ scores of statistics professors are normally distributed and use a 0.05 significance level to test the claim that σ = 15. [The tabulated value x^2 with d.f. 12 at 5% level of significance are 4.404 and 23.337].

To test the claim that the population standard deviation (σ) of IQ scores for statistics professors is equal to 15, we can perform a chi-square test for the variance.

The chi-square test statistic for testing the population variance is given by:

$\chi^2 = (n - 1) * (s^2) / (\sigma_o^2)$

where: n = sample size (number of statistics professors) = 13 s^2 = sample variance = (standard deviation)^2 = (7.2)^2 = 51.84 σ_o^2 = hypothesized population variance = (15)^2 = 225

Step 1: State the null and alternative hypotheses:

Null hypothesis (H0): The population variance of IQ scores for statistics professors is equal to 225 ($\sigma^2 = 225$). Alternative hypothesis (H1): The population variance of IQ scores for statistics professors is not equal to 225 ($\sigma^2 \neq 225$).

Step 2: Set the significance level (α):

Given that the significance level is 0.05 (5% level of significance), $\alpha = 0.05$.

Step 3: Calculate the test statistic (χ^2):

$\chi^2 = (n - 1) * (s^2) / (\sigma_o^2)$ $\chi^2 = (13 - 1) * (51.84) / (225)$ $\chi^2 \approx 12.263$

Step 4: Find the critical values:

At a 5% level of significance ($\alpha = 0.05$) and degrees of freedom (df = n - 1 = 13 - 1 = 12), the critical values from the chi-square distribution table are approximately 4.404 and 23.337 for the lower-tail and upper-tail tests, respectively.

Step 5: Make a decision:

Compare the calculated chi-square test statistic (χ^2) with the critical values.

4.404 < 12.263 < 23.337

Step 6: Draw the conclusion:

Since the calculated chi-square test statistic (12.263) falls between the critical values, we fail to reject the null hypothesis (H0).

Conclusion: At the 5% level of significance, we do not have sufficient evidence to reject the claim that the population variance of IQ scores for statistics professors is equal to 225. The data does not provide enough support to conclude that the population standard deviation (σ) is different from 15.

6.

(a) Distinguish between parametric and non-parametric statistical tests. Discuss the advantages and disadvantages of non-parametric test.

■ **Advantages**

- ❖ There is no parametric alternative
- ❖ Nominal data or ordinal data are analyzed

- ❖ Less complicated computations for small sample size
- ❖ Exact method. Not approximation.

■ **Disadvantages**

- ❖ Less powerful if parametric tests are available.
- ❖ Not widely available and less well know
- ❖ For large samples, calculations can be tedious.

(b) Derive sign test, stating clearly the assumptions made for small sample case

(e) Use the sign test to see whether there is a difference between the number of days required to collect an account receivable before and after a new collection policy. Use the 0.05 significance level.

Before: 33 36 41 32 39 47 34 29 32 34 40 42

After: 35 29 38 34 37 47 36 32 30 34 41 38

To determine whether there is a difference between the number of days required to collect an account receivable before and after a new collection policy, we can use the sign test. The sign test is a non-parametric test that compares paired data to assess if there is a significant difference between the two conditions.

The steps for conducting the sign test are as follows:

Step 1: State the null and alternative hypotheses:

Null hypothesis (H0): There is no difference in the number of days required to collect an account receivable before and after the new collection policy. Alternative hypothesis

(H1): There is a difference in the number of days required to collect an account receivable before and after the new collection policy.

Step 2: Collect and summarize the data:

Given the before and after data: Before: 33 36 41 32 39 47 34 29 32 34 40 42 After: 35 29 38 34 37 47 36 32 30 34 41 38

Step 3: Calculate the differences:

Calculate the differences between the before and after values for each pair. The differences will help us determine whether there is an increase or decrease in the number of days after the new collection policy.

Differences: (35-33) (29-36) (38-41) (34-32) (37-39) (47-47) (36-34) (32-29) (30-32) (34-34) (41-40) (38-42) = 2 -7 -3 2 -2 0 2 3 -2 0 1 -4 = 2 -7 -3 2 -2 0 2 3 -2 0 1 -4

Step 4: Count the number of positive and negative differences:

Count the number of positive differences (number of cases where "After" value is greater than "Before" value) and the number of negative differences (number of cases where "After" value is less than "Before" value).

Positive differences: 9 Negative differences: 3

Step 5: Determine the test statistic:

The test statistic for the sign test is the smaller of the number of positive differences (p) and the number of negative differences (q).

Test statistic (T) = min(p, q)

In this case, T = min(9, 3) = 3

Step 6: Find the critical value:

At a 5% level of significance (α = 0.05) and for a two-tailed test, the critical value for T is 2.

Step 7: Make a decision:

Compare the test statistic (T) with the critical value.

3 > 2

Step 8: Draw the conclusion:

Since the test statistic (T = 3) is greater than the critical value (2), we reject the null hypothesis (H0).

Conclusion: At the 5% level of significance, we have sufficient evidence to conclude that there is a difference in the number of days required to collect an account receivable before and after the new collection policy.

Theory of Statistics -2016

Part-A

1. a) Define Chi-square (2) distribution, Show that distribution tends to normal distribution for large degrees of freedom.

b) State and prove additive property of y-distribution. If x has density function f(x) = ex>0. Then show that 2x follows x-distribution with 2-degrees of freedom

2. a) Define F variate. Find the mode of F distribution. If X has F distribution with m and n degrees of freedom, show that 1/X has also P distribution with n and m degrees of freedom. Mention some important properties of F distribution

3. a) Define point estimation with example. What are the methods of point estimation?

b) What is MLE? State and prove the invariance property of MLE.

c) Let x1, x2,.....Xn, be a random sample from f(x; n,p) = ()p(1-p -p)",x=0,1,2,...n. Find the MLE of p.

Part-B

4. a) What do you mean by statistical hypothesis? Distinguish between simple and composite hypothesis. Let a random sample of size n is drawn from a normal population with mean u and known variance o. How would you test the hypothesis that mean is equal to μ0 ?

b) The average IQ of university female students in Bangladesh is suspected to be more than the average 110 for all students. A random sample of 64 female students yielded a sample average IQ of 115.5 and standard deviation of 20. Can you conclude that the average score of the female students is really more than 1107 [Z0.05=1.64]

5. a) Define rxe contingency table. Show that in case of 2x2 contingency table, the test statistics becomes x2 = N(ad-bc)2 (a+b)(c+d)(a+c)(b+d) , also mention Yate's correction for continuity.

b) In a psychological test, 70 out of 100 boys came out successful while 60 out of 100 girls of the same age group as the boys passed the test. Do the data provide any evidence of difference in respect of abilities between the genders?

6. a) What do you mean by non-parametric test? Discuss it's importance. Describe the testing procedure of the run test.

b) The following sequence is purported to be a set of random integers from 0 to 99. Use the run's test to test the hypothesis of the randomness at a 0.05 significance level. The sequence is

28, 4, 23, 98, 44, 10, 6, 25, 54, 81, 12, 6, 4, 33, 67, 55, 71, 66, 22, 18, 49, 85

Theory of Statistics -2015

Part-A

1. (a) Define Sampling Distribution with examples. Mention the names of sampling

distributions which are frequently used.

(b) Define x^2-distribution. Find moment generating function of x2-distribution. Also

find mean and variance of x^2-distribution.

3. (a) State and prove t-distribution.
   Show that the odd ordered moments of t-

distribution is zero. Write down the properties of t-distribution.

(b) Show that t-distribution tends to normal distribution if the degrees of freedom tend to infinity.

3.

(a) What are the desirable criteria of a good estimator?

(b) Let (x1,x2,...,x,) be a random sample from a Poisson distribution with p.d.f. given by f(x12)= Find the maximum likelihood estimator of A. Show that the estimator is unbiased.

x = 0,1,2,.....

4.

Part-B

(a) What do you mean by a statistical a statistical hypothesis?

In statistics, a statistical hypothesis is a statement or assumption about a population parameter (or parameters) that we want to test using sample data. Hypothesis testing is a fundamental tool for making inferences about populations based on sample data. The process involves setting up two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (H1 or Ha), and then using sample data to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

- **Null Hypothesis (H0):** It states that there is no significant difference or relationship between variables, or no effect of a treatment, intervention, or manipulation. It is the hypothesis that we aim to test and reject if there is sufficient evidence.
- **Alternative Hypothesis (H1 or Ha**): It represents what we want to show or what we suspect is true. It contradicts the null hypothesis and states that there is a significant difference, relationship, or effect.

Describe different steps for testing statistical hypothesis.

the steps for testing a statistical hypothesis:

1. State the Hypotheses.
2. Set the Significance Level (α).
3. Collect Data.
4. Calculate the Test Statistic.
5. Determine the Critical Region.
6. Compare Test Statistic and Critical Region.
7. Draw Conclusion.

Write down the procedure to test the significance of regression coefficient.

(b) A random sample of 10 persons is selected as follows: 5, 2, 0, 4, 16, 14, 10, 11, 6, 8. Do you think that the average schooling year of the persons in population is 5? (Tabulated value at 5% with 9 d.f. is 2.26)

Null hypothesis (H0): The average schooling year of the persons in the population is 5.

Alternative hypothesis (H1): The average schooling year of the persons in the population is not 5.

Step 1: Calculate the sample mean ($\bar{x}$) and sample standard deviation (s).

Sample: 5, 2, 0, 4, 16, 14, 10, 11, 6, 8 Number of observations (n) = 10

Sample mean ($\bar{x}$) = (5 + 2 + 0 + 4 + 16 + 14 + 10 + 11 + 6 + 8) / 10 = 76 / 10 = 7.6

Sample standard deviation (s) = $\sqrt{[(5-7.6)^2 + (2-7.6)^2 + (0-7.6)^2 + (4-7.6)^2 + (16-7.6)^2 + (14-7.6)^2 + (10-7.6)^2 + (11-7.6)^2 + (6-7.6)^2 + (8-7.6)^2) / (10-1)]}$ = $\sqrt{[(42.4 + 33.6 + 57.6 + 11.6 + 68.4 + 41.6 + 5.6 + 8.4 + 2.4 + 0.16) / 9]}$ = $\sqrt{[271.4 / 9]}$ = $\sqrt{30.16}$ = 5.49 (approximately)

Step 2: Calculate the t-statistic:

$t = (\bar{x} - \mu) / (s / \sqrt{n})$

$t = (7.6 - 5) / (5.49 / \sqrt{10})$

$t = 2.6 / (5.49 / 3.162)$

$t = 2.6 / 1.739$ $t \approx 1.495$

Step 3: Compare t-statistic with critical value:

The critical value at 5% significance level ($\alpha = 0.05$) with 9 degrees of freedom is 2.26 (as given in the table).

Since the absolute value of the t-statistic (|1.495|) is less than the critical value of 2.26, we fail to reject the null hypothesis.

Conclusion: Based on the correct t-test, there is not enough evidence to suggest that the average schooling year of the persons in the population is significantly different from 5 at the 5% significance level.

6.(a) Define Type I error, Type II error, Level of significance and Most powerful test.

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population;

a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

(b) Suppose k random samples are drawn from normal population with mean HH2H and common variance a^2. Describe the procedure to test that the means are equal.

6. (a) What is contingency table?

In statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering, and scientific research.

What is form of x^2 test statistic in case of a 2x2 contingency table?

In the case of a 2x2 contingency table, where both variables have two categories each, the Chi-Square (χ^2) test statistic is used to test the independence of the two variables. The formula for the Chi-Square test statistic in a 2x2 table is as follows:

$$\chi^2 = (ad - bc)^2 * n / [(a + b) * (c + d) * (a + c) * (b + d)]$$

where: a = frequency count in cell (1,1) of the table b = frequency count in cell (1,2) of the table c = frequency count in cell (2,1) of the table d = frequency count in cell (2,2) of the table n = total number of observations (sum of all frequencies in the table)

The Chi-Square test statistic measures the difference between the observed frequencies in the cells and the expected frequencies under the assumption of independence. If the Chi-Square test statistic is large, it indicates that there is a significant association between the two categorical variables. On the other hand, if the test statistic is small, it suggests that the variables are independent.

By comparing the calculated Chi-Square test statistic with the critical value from the Chi-Square distribution table at a chosen significance level and degrees of freedom, we can determine whether to reject or fail to reject the null hypothesis, which states that there is no significant association between the two variables.

(b) For given information in the following table, test at level of significance 0.05 that whether level of education affects the job performance. [x20.054 = 13.3]

Sujan prodhan CSE-29 fb.com/prodhan24 github.com/prodhan2

Theory of Statistics -CT-1 question -2023

Answer any Two questions.

(a) What Chi-square statistic? Write down the pdf of Chi-square statistic. What are the uses of Chi-square statistic?

(b) What is sampling distribution? Define Fisher t statistic? Show that for t statistic $B_1=0$ and $B_2=3(n-2)/(n-4)$

(C) What is F statistic? Write down the uses of F-statistic. Establish the relationship between F and Chi-square statistic. Under usual notations prove that F = (p-1)

(d) Define with examples

(i) statistic,

(ii) Estimation and

(iii) Estimator.

What are the characteristics of a good estimator? A random sample of size 3 is drawn from a normal population with mean u. Consider the following estimator to estimate μ, t= X1+3X2+2X3/3

Find the value of ⌀ such that t is an unbiased estimator of u.

(a) Chi-square Statistic:

- The chi-square statistic is a measure used in statistical hypothesis tests, especially for testing the independence of categorical variables.
- PDF (Probability Density Function): The chi-square statistic follows a chi-square distribution. Its PDF depends on the degrees of freedom (df).
- Uses: Chi-square statistic is used for testing independence in contingency tables, goodness-of-fit tests, and in various statistical analyses involving categorical data.

(b) Sampling Distribution and Fisher t Statistic:

- Sampling Distribution: It's the distribution of a statistic (like the mean) calculated from multiple random samples drawn from the same population.
- Fisher t Statistic: It's a statistic used in t-tests to assess differences between sample means.
- B_1 and B_2 are correction factors in formulas for pooled and unpooled variance estimators, respectively.

(c) F Statistic:

- F Statistic: It's a statistic used in analysis of variance (ANOVA) and regression analysis to compare variances between two or more groups.
- Uses: F-statistic helps determine if group means are significantly different or if a regression model is statistically significant.
- Relationship with Chi-square: In some cases, the F statistic is equal to the square of a chi-square statistic. F = (Chi-square)^2.

(d) Definitions with Examples:

- Statistic: A statistic is a numerical value calculated from sample data (e.g., sample mean, sample standard deviation).

- Estimation: Estimation is the process of making educated guesses or predictions about population parameters based on sample data.
- Estimator: An estimator is a formula or rule used to calculate an estimate (e.g., sample mean as an estimator for the population mean).

Characteristics of a Good Estimator:

- Unbiasedness: An estimator is unbiased if its expected value equals the true population parameter.
- Efficiency: An estimator is efficient if it has a small variance, providing precise estimates.
- Consistency: An estimator is consistent if it converges to the true parameter as the sample size increases.

Estimator t for μ:

- $t = (X_1 + 3X_2 + 2X_3) / 3$
- To make it unbiased (E[t] = μ), set ⌀ = -2.

These concise explanations should help you understand these statistical concepts and definitions more easily.

Theory of Statistics -CT-2 question -2023

Answer any three of the following questions.

1.. Define with examples

(i) Parameter and Statistic,

(ii) Null and Alternative hypotheses,

(iii) Critical Value and Critical Region.

2.Explain clearly about type I and type II errors.

What do you mean by p value and level of significance?

Discuss about the test procedures of hypothesis testing.

3.What is power of a test? What is BC R stand for? Explain the test procedure for testing the

(i) Ho μ=μo vs. H_1: μ<μo (For large sample case) and

(ii) Ho: H=μo vs. HI: μ# Ho (for small and unknown variance).

4. What is dichotomous and manifold classification? State the uses Yate's correction.

5. Explain about Brandt and Snedecor 2xk contingency table for testing independence of attributes.

6. The mean breaking strength of cables supplied by a manufacturer is 1800 with a standard deviation 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cable has increased. In the order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 0.01 and 0.05 level of significances?

7. A random sample of 12 boys had the following I.Q.'s: 70, 120, 111, 98, 110, 88, 107, 100, 83, 90, 115, 81. Do these data support the assumption of a population mean I.Q. of 100?

1. Parameter and Statistic:

- Parameter: A population characteristic (e.g., population mean μ).
- Statistic: A sample characteristic (e.g., sample mean $\bar{x}$).

2. Type I and Type II Errors:

- Type I Error: Incorrectly saying there's an effect when there isn't (false positive).
- Type II Error: Incorrectly saying there's no effect when there is (false negative).

3. P-value and Level of Significance:

- P-value: Measures evidence against null hypothesis; smaller means stronger evidence against.
- Level of Significance (α): Threshold for rejecting the null hypothesis (e.g., 0.05).

4. Power of a Test:

- Power: Probability of correctly detecting an effect when it exists (higher is better).

5. Test Procedures for Hypothesis Testing:

- (1) Formulate null and alternative hypotheses.
- (2) Collect sample data.
- (3) Calculate a test statistic.
- (4) Find the p-value.
- (5) Compare p-value to α (level of significance).
- (6) If $p \leq \alpha$, reject null; otherwise, fail to reject.
- **6. Dichotomous and Manifold Classification:**
- Dichotomous Classification: Divides data into two categories or groups (e.g., pass/fail).
- Manifold Classification: Divides data into multiple categories or groups (e.g., A/B/C grades).
- **7. Yates's Correction:**
- Yates's Correction: A statistical adjustment used in 2x2 contingency tables to reduce the chance of Type I errors when conducting a chi-squared test for independence. It's applied by subtracting 0.5 from the absolute difference between observed and expected frequencies in each cell.
- **8. Brandt and Snedecor 2xk Contingency Table:**
- Brandt and Snedecor Table: A statistical table used to test the independence of attributes in a 2xk contingency table (cross-tabulation). It helps determine if variables are related or independent.
- **9. Testing Claim about Cable Strength:**
- To test the cable manufacturer's claim about increased strength, compare the sample mean (1850) to the population mean (1800) using hypothesis testing.
- At a 0.01 level of significance (α), calculate the p-value. If $p \leq 0.01$, support the claim.
- Repeat the process for a 0.05 level of significance (α). If $p \leq 0.05$, also support the claim.

- **10. Testing Population Mean IQ:**
- Use a hypothesis test to check if the sample mean IQ (data given) supports the assumption of a population mean IQ of 100.
- Formulate null (H0: μ = 100) and alternative (H1: μ ≠ 100) hypotheses.
- Calculate the test statistic and find the p-value.
- If p-value ≤ α (chosen level of significance), either reject or fail to reject the null hypothesis based on the given α (e.g., 0.05).
- These simplified explanations should help you understand these statistical concepts more easily.
-

4. **Dichotomous and Manifold Classification:**

- **Dichotomous Classification:** Classification into two exclusive and exhaustive categories or classes. Example: Classifying individuals into 'Yes' or 'No' based on a specific characteristic.
- **Manifold Classification:** Classification into multiple categories or classes, not limited to two. Example: Classifying fruits into 'Apples,' 'Oranges,' 'Bananas,' and 'Grapes.'

Uses of Yates' Correction: It is used in 2x2 contingency tables in statistics to correct for the potential overestimation of statistical significance in small sample sizes. It adjusts chi-square calculations to improve accuracy in hypothesis testing.

5. **Brandt and Snedecor 2xk Contingency Table for Independence Testing:**
    - A 2xk contingency table is used to analyze the association between two categorical variables, where one variable has 2 categories and the other has k categories.
    - It's employed to test the independence of attributes and determine if there's a significant relationship between the two variables.
6. **Testing the Claim on Cable Breaking Strength:**
    - Mean breaking strength of cables (μ) = 1800 (claim by manufacturer)
    - New mean breaking strength (�¯x¯) = 1850 (sample mean)
    - Standard deviation (�σ) = 100 (given)
    - Sample size (n) = 50

Hypotheses:

- Null Hypothesis (H0): ==1800μ=1800 (No increase in strength)
- Alternative Hypothesis (H1): >1800μ>1800 (Claim is supported)

Level of Significance:

- α=0.01α=0.01 and α=0.05α=0.05

Procedure:

- Calculate the z-score: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
- Compare the z-score with critical z-values for the given significance levels.
- If the z-score is greater than the critical z-value, we support the claim.

7. **Testing Population Mean IQ with Sample Data:**

- Given sample IQ data (12 boys).
- Population mean IQ assumption (μ) = 100

Hypotheses:

- Null Hypothesis (H0): μ=100μ=100 (Population mean IQ is 100)
- Alternative Hypothesis (H1): $\mu \neq 100$$\mu$≠100 (Population mean IQ is different)

Procedure:

- Calculate the sample mean ($\bar{x}$) and standard deviation (s) from the given data.
- Use the t-test or z-test to determine if the sample supports or rejects the population mean IQ assumption of 100.

1. DICHOTOMY CLASSIFICATION

When classified on the basis of one quality. Ex. Literate-illiterate, male- female, etc.

2. MANIFOLD CLASSIFICATION When classified on the basis of two or more than two qualities. Ex. Population classified on the basis of sex & religion.

BCR stands for Bayesian credible region. It is a region of the parameter space that contains the true parameter value with a specified probability. The BCR can be used to make inferences about the parameter value, even when the sample size is small.

Power of a Test: The power of a statistical hypothesis test is the probability of correctly rejecting a null hypothesis (H0) when the alternative hypothesis (H1) is true. In other words, it measures the test's

ability to detect a true effect or difference when it exists. A high-power test is desirable because it indicates a lower probability of making a Type II error (failing to reject a false null hypothesis).

The power of a test depends on several factors, including the significance level (alpha), the sample size (n), the effect size (the magnitude of the difference or effect you want to detect), and the variability (standard deviation) of the data.

Higher power can be achieved by increasing the sample size, using a more sensitive test statistic, or reducing the variability in the data.

Now, let's discuss the test procedures for two different cases:

(i) Testing H0: μ = μo vs. H1: μ < μo (For Large Sample Case):

In this case, you are testing whether the population mean (μ) is less than a specified value (μo) when dealing with a large sample size. The steps for this test are as follows:

1. **State Hypotheses:**
    - Null Hypothesis (H0): μ = μo (The population mean is equal to μo).
    - Alternative Hypothesis (H1): μ < μo (The population mean is less than μo).
2. **Choose Significance Level (α):**
    - Decide on a significance level, such as α = 0.05, which represents the probability of making a Type I error (incorrectly rejecting a true null hypothesis)
3. **Collect Data and Calculate the Test Statistic:**
    - Collect a large sample of data.
    - Calculate the sample mean ($\bar{x}$) and sample standard deviation (s).
    - Compute the test statistic, typically a z-score, using the formula: $z = (\bar{x} - \mu o) / (\sigma / \sqrt{n})$
4. **Determine the Critical Value:**
    - Based on your chosen significance level (α) and the one-tailed nature of the test (since it's a left-tailed test), find the critical value from the standard normal distribution (z-table or calculator).
5. **Make a Decision:**
    - If the test statistic (z) is less than the critical value, reject the null hypothesis (H0).
    - If the test statistic is greater than or equal to the critical value, fail to reject the null hypothesis.
6. **Draw a Conclusion:**

- Based on the decision, draw a conclusion about whether there is enough evidence to support the alternative hypothesis (H1).

(ii) Testing H0: μ = μo vs. H1: μ ≠ μo (For Small and Unknown Variance):

In this case, you are testing whether the population mean (μ) is not equal to a specified value (μo) when dealing with a small sample size and an unknown population variance. This typically involves using a t-test. The steps for this test are as follows:

1. **Formulate Hypotheses:**
   - Null Hypothesis (H0): μ = μo (The population mean is equal to μo).
   - Alternative Hypothesis (H1): μ ≠ μo (The population mean is not equal to μo).
2. **Choose Significance Level (α):**
   - Decide on a significance level, such as α = 0.05.
3. **Collect Data and Calculate the Test Statistic:**
   - Collect a small sample of data.
   - Calculate the sample mean ($\bar{x}$) and sample standard deviation (s).
   - Compute the t-statistic using the formula: $t = (\bar{x} - \mu o) / (s / \sqrt{n})$
4. **Determine the Degrees of Freedom:**
   - Degrees of freedom (df) depend on the sample size and are used to find the critical t-value from the t-distribution table.
5. **Find the Critical Values:**
   - Determine the critical t-values corresponding to the chosen significance level and degrees of freedom (df). This is a two-tailed test.
6. **Make a Decision:**
   - If the calculated t-statistic falls outside the range defined by the critical t-values, reject the null hypothesis (H0).
   - If the calculated t-statistic falls within the range, fail to reject the null hypothesis.
7. **Draw a Conclusion:**
   - Based on the decision, draw a conclusion about whether there is enough evidence to support the alternative hypothesis (H1).

Remember that the choice between a z-test (large sample, known variance) and a t-test (small sample, unknown variance) depends on the characteristics of your data and the assumptions you can make about the population.

Suggestions :::

what are the methods of non perametric test ??

1. **Mann-Whitney U Test (Wilcoxon Rank-Sum Test)**:
   - Compares two groups to see if one tends to have larger values.
2. **Kruskal-Wallis Test**:
   - Compares three or more groups to check if at least one group is different.
3. **Wilcoxon Signed-Rank Test**:
   - Checks if two related groups have similar distributions.
4. **Chi-Square Test**:
   - Checks if there's a relationship between categories.
5. **Kolmogorov-Smirnov Test**:
   - Checks if a sample differs significantly from a known distribution.
6. **Spearman's Rank Correlation**:
   - Measures how closely two variables follow each other.
7. **Runs Test**:
   - Checks for randomness in ordered data.

uses of chi square test in statistics??

1. **Goodness of Fit**:
   - Checks if our data fits a specific expected pattern or distribution.
2. **Test of Independence**:
   - Checks if two things are related or if changes in one affect the other.
3. **Comparison of Proportions**:
   - Compares proportions or percentages across different categories or groups.
4. **Homogeneity Testing**:
   - Determines if the distribution of a categorical variable is similar across multiple groups.
5. **Association Testing**:
   - Checks if there's a significant association or relationship between categorical variables.
6. **Hypothesis Testing**:
   - Helps in testing hypotheses related to categorical data.

7. **Predictive Modeling**:
    - Informs predictive models by assessing relationships between variables.

15-point list of uses for the F-statistic in statistics:

1. **Compare Group Means**: Determines if means of groups are significantly different.
2. **ANOVA (Analysis of Variance)**: Tests group mean differences in experiments.
3. **Regression Analysis**: Evaluates the overall model significance.
4. **Model Comparison**: Compares nested or non-nested models to see which fits better.
5. **Variable Selection**: Assists in selecting relevant predictors for a model.
6. **Hypothesis Testing**: Tests specific hypotheses about model parameters.
7. **Experiments**: Assesses treatment effects in designed experiments.
8. **Time Series Analysis**: Tests differences or patterns over time.
9. **Quality Control**: Compares variances to maintain product quality.
10. **Multivariate Analysis**: Evaluates relationships in multivariate data.
11. **Factor Analysis**: Assesses the number of factors to retain in the analysis.
12. **Structural Equation Modeling (SEM)**: Evaluates the goodness of fit of the model.
13. **Reliability Analysis**: Assesses the reliability of scales or measures.
14. **Multilevel Modeling**: Evaluates the significance of group-level effects.
15. **Survival Analysis**: Compares survival curves for different groups or treatments.