Pabna University of Science and Technology

Faculty of Engineering and Technology

Department of ICE

Assignment

Course name: Engineering Statistics

Course Code: STAT- 2201

Submitted By :

Name: Saklam Sakib
Roll No: 220609
Session: 2021-2022
2nd Year 2nd Semester
Dept. of Information and Communication Engineering

Submitted To :
Dr. Md. Sarwar Hosain

Associate Professor

Department of ICE,, PUST

Date of submission: 23.04.2025

Distribution of the Sample Correlation Coefficient in the Null Case

Introduction

The sample correlation coefficient is a widely used statistic to measure the strength and direction of the linear relationship between two variables. In statistical hypothesis testing, understanding the distribution of the sample correlation coefficient under the null hypothesis is crucial for assessing whether an observed correlation is statistically significant. This assignment explores the distribution of the sample correlation coefficient when the true population correlation coefficient is zero (the null case), focusing on its theoretical foundation, derivation, and practical applications.

Sample Correlation Coefficient

Consider a sample of n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from a bivariate population. The sample correlation coefficient, denoted r, measures the linear relationship between x and y. It is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means of x and y, respectively.

Alternatively, in terms of sample covariance and standard deviations:

$$r = \frac{s_{xy}}{s_x s_y}$$

where:

- $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is the sample covariance,

- $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ are the sample standard deviations.

The value of r ranges from -1 to 1, with $r = 0$ indicating no linear relationship, $r > 0$ indicating a positive linear relationship, and $r < 0$ indicating a negative linear relationship.

Null Hypothesis and the Null Case

In hypothesis testing for correlation, the null hypothesis typically states that the population correlation coefficient ρ is zero:

$$H_0 : \rho = 0$$

against the alternative hypothesis $H_a : \rho \neq 0$ (two-tailed test) or $H_a : \rho > 0$ or $H_a : \rho < 0$ (one-tailed tests).

The "null case" refers to the scenario where H_0 is true, i.e., $\rho = 0$, meaning there is no linear correlation between the two variables in the population. Understanding the distribution of r under this null hypothesis is essential for determining critical values and p-values in hypothesis testing.

Distribution of the Sample Correlation Coefficient in the Null Case

To derive the distribution of r when $\rho = 0$, we assume the following:

- The paired observations (x_i, y_i) are drawn from a bivariate normal distribution.
- The population correlation coefficient $\rho = 0$, implying that x and y are independent.
- Both x and y are normally distributed with means μ_x, μ_y and variances σ_x^2, σ_y^2, respectively.

Under these assumptions, the exact distribution of the sample correlation coefficient r can be derived. When $\rho = 0$, the probability density function (PDF) of r for a sample of size n is:

$$f(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\sqrt{\pi}}(1 - r^2)^{\frac{n-4}{2}}, \quad -1 < r < 1$$

where Γ is the gamma function, and $n \geq 3$ (since the correlation coefficient requires at least three observations to be well-defined).

This density function is symmetric about $r = 0$, reflecting that positive and negative correlations are equally likely when $\rho = 0$. The distribution depends on the sample size n, and as n increases, the distribution of r becomes more concentrated around zero.

Transformation to a t-Distribution

For hypothesis testing, it is often more convenient to use a transformation of r that follows a known distribution. When $\rho = 0$, the statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

follows a Student's t-distribution with $n - 2$ degrees of freedom. This transformation is pivotal because it allows us to perform hypothesis tests and construct confidence intervals using the t-distribution, which is well-tabulated and widely implemented in statistical software.

The t-statistic arises because, under the null hypothesis, r is a function of the sample covariance and variances, and the bivariate normality assumption ensures that the transformed statistic follows a t-distribution.

Derivation Outline

To understand why r leads to a t-distribution, consider the following:

Rewrite the sample correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Standardize the variables by letting $z_{xi} = \frac{x_i - \bar{x}}{s_x}$ and $z_{yi} = \frac{y_i - \bar{y}}{s_y}$. Then:

$$r = \frac{\sum_{i=1}^{n} z_{xi} z_{yi}}{\sqrt{\sum_{i=1}^{n} z_{xi}^2 \sum_{i=1}^{n} z_{yi}^2}}$$

When $\rho = 0$, x and y are independent, and the numerator $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is a sum of products of independent normal variables, scaled by the sample variances. The t-statistic is derived by normalizing r to account for the variability in the sample, leading to the t-distribution with $n - 2$ degrees of freedom.

Hypothesis Testing

To test $H_0 : \rho = 0$, compute the sample correlation coefficient r and the t-statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Compare the t-statistic to critical values from the t-distribution with $n-2$ degrees of freedom, or compute the p-value. For a two-tailed test at significance level α, reject H_0 if:

$$|t| > t_{\alpha/2,n-2}$$

where $t_{\alpha/2,n-2}$ is the critical value from the t-distribution.

The p-value is:

$$\text{p-value} = 2 \cdot P(T > |t|)$$

where $T \sim t_{n-2}$.

Illustration with Examples

Example 1

Suppose a researcher collects data on 10 pairs of observations (x_i, y_i) and computes the sample correlation coefficient as $r = 0.45$. Test the null hypothesis $H_0 : \rho = 0$ at the 5% significance level.

Solution:

The sample size is $n = 10$. Compute the t-statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{10-2}}{\sqrt{1-(0.45)^2}} = \frac{0.45\sqrt{8}}{\sqrt{1-0.2025}} = \frac{0.45 \cdot 2.828}{\sqrt{0.7975}} \approx \frac{1.2726}{0.893} \approx 1.425$$

The degrees of freedom are $n-2 = 8$. For a two-tailed test at $\alpha = 0.05$, the critical value from the t-distribution is approximately $t_{0.025,8} \approx 2.306$.

Since $|t| = 1.425 < 2.306$, we fail to reject H_0. There is insufficient evidence to conclude that $\rho \neq 0$.

To compute the p-value:

$$\text{p-value} = 2 \cdot P(T > 1.425) \text{ for } T \sim t_8$$

Using a t-table or software, $P(T > 1.425) \approx 0.096$, so:

$$\text{p-value} \approx 2 \cdot 0.096 = 0.192$$

Since $0.192 > 0.05$, we fail to reject H_0.

Example 2

A study with 15 pairs of observations yields a sample correlation coefficient of $r = -0.62$. Test $H_0 : \rho = 0$ at the 1% significance level.

Solution:

The sample size is $n = 15$. Compute the t-statistic:

$$t = \frac{-0.62\sqrt{15-2}}{\sqrt{1-(-0.62)^2}} = \frac{-0.62\sqrt{13}}{\sqrt{1-0.3844}} = \frac{-0.62 \cdot 3.606}{\sqrt{0.6156}} \approx \frac{-2.2357}{0.7846} \approx -2.849$$

The degrees of freedom are $n-2 = 13$. For a two-tailed test at $\alpha = 0.01$, the critical value is approximately $t_{0.005,13} \approx 3.012$.

Since $|t| = 2.849 < 3.012$, we fail to reject H_0.

The p-value is:

$$\text{p-value} = 2 \cdot P(T > 2.849) \text{ for } T \sim t_{13}$$

Using a t-table or software, $P(T > 2.849) \approx 0.007$, so:

$$\text{p-value} \approx 2 \cdot 0.007 = 0.014$$

Since $0.014 > 0.01$, we fail to reject H_0.

Properties of the Distribution

- **Symmetry**: When $\rho = 0$, the distribution of r is symmetric about zero, as positive and negative correlations are equally likely.
- **Dependence on Sample Size**: The variance of r decreases as n increases, making r a more precise estimator of $\rho = 0$ for larger samples.
- **Asymptotic Behavior**: For large n, the distribution of r approaches normality, but the t-transformation is exact for any $n \geq 3$ under bivariate normality.

Practical Notes

Assumptions

The t-test for the correlation coefficient relies on the bivariate normality of (x, y). If this assumption is violated, the distribution of r may not follow the t-distribution, and alternative methods (e.g., nonparametric tests like Spearman's rank correlation) may be needed.

Confidence Intervals

To construct a confidence interval for ρ, Fisher's z-transformation is often used, as r itself is not normally distributed. However, in the null case, the t-test is sufficient for hypothesis testing.

Software Implementation

Statistical software like R or Python can compute r and perform the t-test. For example, in R, the function `cor.test(x, y)` returns the sample correlation, t-statistic, and p-value.

Conclusion

The distribution of the sample correlation coefficient in the null case ($\rho = 0$) is symmetric and depends on the sample size n. Under bivariate normality, the transformed statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ follows a t-distribution with $n-2$ degrees of freedom, enabling hypothesis testing. The examples illustrate how to test for zero correlation, highlighting the importance of the t-statistic in determining statistical significance.