CHI-SQUARE DISTRIBUTION

$$\chi^2$$

INTRODUCTION

⩥ The Chi-Square distribution is a continuous probability distribution that represents the sum of squares of independent standard normal random variables.

⩥ The Chi-Square distribution is denoted by χ (Chi-Square) and is characterized by a single parameter, k (degrees of freedom).

⩥ The Chi-Square distribution is used to model the distribution of sums of squares of normal random variables.

⩥ The Chi-Square distribution is widely used in statistical testing, particularly in hypothesis testing and confidence intervals for population variances.

⩥The Chi-Square distribution is a fundamental distribution in statistics and is widely used in various applications, including hypothesis testing, confidence intervals, and regression analysis.

DEFINITION

The Chi-Square distribution is a continuous probability distribution defined as the sum of the squares of k independent standard normal random variables. It is denoted by χ^2 (Chi-Square) and is characterized by a single parameter, k (degrees of freedom)

[the Chi-Square distribution can be defined as:$\chi^2 = X_1^2 + X_2^2 + \ldots + X_k^2$where:- χ^2 is the Chi-Square random variable- $X_1, X_2, \ldots, X_k$ are independent standard normal random variables (mean 0 and variance 1)- k is the number of degrees of freedom ($k \geq 1$)]

APPLICATIONS

- Hypothesis Testing : The Chi-Square test is used to determine whether there is a significant association between two categorical variables.
- Goodness of Fit : The Chi-Square test is used to determine whether a distribution fits a theoretical distribution.
- Contingency Tables : The Chi-Square test is used to analyze the relationship between two categorical variables in a contingency table.
- Regression Analysis : The Chi-Square distribution is used to test the significance of regression coefficients.
- Quality Control : The Chi-Square distribution is used to monitor and control the quality of manufacturing processes

DERIVATION OF CHI-SQUARE DISTRIBUTION

$$
\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(t)\,dt \\
&= \int_{-\infty}^{x} c t^{n/2-1} \exp\left(-\frac{1}{2}t\right) dt \\
&= c\int_{-\infty}^{x/2} (2s)^{n/2-1} \exp(-s)\,2\,ds && \text{(by a change of variable: } s = t/2\text{)} \\
&= c2^{n/2}\int_{-\infty}^{x/2} s^{n/2-1} \exp(-s)\,ds \\
&= \frac{1}{2^{n/2}\Gamma(n/2)} 2^{n/2}\int_{-\infty}^{x/2} s^{n/2-1} \exp(-s)\,ds && \text{(by the definition of } c\text{)} \\
&= \frac{1}{\Gamma(n/2)}\int_{-\infty}^{x/2} s^{n/2-1} \exp(-s)\,ds \\
&= \frac{\gamma(n/2, x/2)}{\Gamma(n/2)}
\end{aligned}
$$

PROPERTIES

Mean :

$$
\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x)\,dx \\
&= \int_0^\infty x c x^{n/2-1} \exp\left(-\tfrac{1}{2}x\right) dx \\
&= c\int_0^\infty x^{n/2} \exp\left(-\tfrac{1}{2}x\right) dx \\
&= c\left\{\left[-x^{n/2} 2\exp\left(-\tfrac{1}{2}x\right)\right]_0^\infty\right. \\
&\quad \left. + \int_0^\infty \tfrac{n}{2} x^{n/2-1} 2\exp\left(-\tfrac{1}{2}x\right) dx\right\} \qquad \text{(integrating by parts)} \\
&= c\left\{(0-0) + n\int_0^\infty x^{n/2-1} \exp\left(-\tfrac{1}{2}x\right) dx\right\} \\
&= n\int_0^\infty c x^{n/2-1} \exp\left(-\tfrac{1}{2}x\right) dx \\
&= n\int_0^\infty f_X(x)\,dx \\
&= n \qquad \text{(integral of a pdf over its support equals 1)}
\end{aligned}
$$

Variance :

V

It can be derived thanks to the usual variance formula ($\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$):

$$
\begin{aligned}
&\mathrm{E}[X^2] \\
&= \int_0^\infty x^2 f_X(x)\,dx \\
&= \int_0^\infty x^2 c x^{n/2-1} \exp\left(-\tfrac{1}{2}x\right) dx \\
&= c\int_0^\infty x^{n/2+1} \exp\left(-\tfrac{1}{2}x\right) dx \\
&= c\left\{\left[-x^{n/2+1} 2\exp\left(-\tfrac{1}{2}x\right)\right]_0^\infty\right. \\
&\quad \left.+\int_0^\infty \left(\tfrac{n}{2}+1\right) x^{n/2} 2\exp\left(-\tfrac{1}{2}x\right) dx\right\} \qquad \text{(integrating by parts)} \\
&= c\left\{(0-0) + (n+2)\int_0^\infty x^{n/2} \exp\left(-\tfrac{1}{2}x\right) dx\right\} \\
&= c(n+2)\left\{\int_0^\infty x^{n/2} \exp\left(-\tfrac{1}{2}x\right) dx\right\} \\
&= c(n+2)\left\{\left[-x^{n/2} 2\exp\left(-\tfrac{1}{2}x\right)\right]_0^\infty\right. \\
&\quad \left.+\int_0^\infty \tfrac{n}{2} x^{n/2-1} 2\exp\left(-\tfrac{1}{2}x\right) dx\right\} \qquad \text{(integrating by parts)} \\
&= c(n+2)\left\{(0-0) + n\int_0^\infty x^{n/2-1} \exp\left(-\tfrac{1}{2}x\right) dx\right\} \\
&= (n+2)n\int_0^\infty c x^{n/2-1} \exp\left(-\tfrac{1}{2}x\right) dx \\
&= (n+2)n\int_0^\infty f_X(x)\,dx \\
&= (n+2)n \qquad \text{(integral of a pdf over its support equals 1)} \\
&\mathrm{E}[X]^2 = n^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\
&= (n+2)n - n^2 = n(n+2-n) = 2n
\end{aligned}
$$

CHI-SQUARE TEST (GOODNESS OF FIT)

- The Chi-Square Goodness of Fit Test is a statistical test used to determine whether a distribution fits a set of data. Here's how to perform the test:
- 1. Formulate a null hypothesis (H0) that the data follows a specific distribution (e.g. normal, Poisson, etc.).
- 2. Formulate an alternative hypothesis (H1) that the data does not follow the specified distribution.
- 3. Collect a random sample of data from the population.
- 4. Calculate the observed frequencies (O) for each category or interval.
- 5. Calculate the expected frequencies (E) for each category or interval based on the specified distribution.
- 6. Calculate the Chi-Square statistic (χ^2) using the

Note: The Chi-Square Goodness of Fit Test assumes that the sample size is sufficiently large and that the observations are independent.

- Determine the degrees of freedom (k) for the test, which is usually the number of categories or intervals minus 1.
- Compare the calculated Chi-Square statistic (χ^2) to a critical value from the Chi-Square distribution with k degrees of freedom.
- If the calculated Chi-Square statistic (χ^2) is greater than the critical value, reject the null hypothesis (H0) and conclude that the data does not fit the specified distribution.
- If the calculated Chi-Square statistic (χ^2) is less than or equal to the critical value, fail to reject the null hypothesis (H0) and conclude that the data fits the specified distribution.

FORMULA

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

T-distribution

INTRODUCTION

- The t-distribution was first introduced by William Sealy Gosset in 1908.
- Gosset published his work under the pseudonym "Student".
- The t-distribution is a special case of the F-distribution.
- The t-distribution is widely used in statistical analysis due to its flexibility and robustness.
- The t-distribution can be used with small sample sizes, making it a useful tool for researchers.
- The t-distribution has been extensively studied and is well-documented in statistical literature.

DEFINITION

The t-distribution is a hypothetical probability distribution. It is also known as the student's t-distribution and used to make presumptions about a mean when the standard deviation is not known to us. It is symmetrical, bell-shaped distribution, similar to the standard normal curve. As high as the degrees of freedom (d.f.), the closer this distribution will approximate a standard normal distribution with a mean of 0 and a standard deviation of 1.

FORMULA

- Let x have a normal distribution with mean 'μ' for the sample of size 'n' with sample mean $\bar{x}$ and the sample standard deviation 's', then the t variable has student's t-distribution with a degree of freedom, d.f = n – 1. The formula for t-distribution is given by;

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

Where, $\bar{x}$ is the mean of first sample.

μ is the mean of second sample.

$\frac{s}{\sqrt{N}}$ = the estimate of the standard error of difference between the means.

ASSUMPTIONS OF T-DISTRIBUTION

- It ranges from -∞ to +∞.
- It has a bell-shaped curve and symmetry similar to normal distribution.
- The shape of the t-distribution varies with the change in degrees of freedom.
- The variance of the t-distribution is always greater than '1' and is limited only to 3 or more degrees of freedom. It means this distribution has a higher dispersion than the standard normal distribution.

MERITS

- Flexibility : The t-distribution can be used with small sample sizes, making it a useful tool for researchers who often have limited data.
- Robustness : The t-distribution is robust to outliers and non-normality, meaning that it can still provide reliable results even when the data doesn't meet the assumptions of normality.
- Wide applicability : The t-distribution can be used in a variety of contexts, including hypothesis testing, confidence intervals, and regression analysis.
- Well-established : The t-distribution has been extensively studied and is well-documented in statistical literature, making it a trusted and widely accepted tool.

DEMERITS

- Assumes normality : The t-distribution assumes that the data is normally distributed, which may not always be the case. Non-normality can lead to inaccurate results.
- Sensitive to outliers : While the t-distribution is robust to some extent, it can still be affected by extreme outliers, which can skew the results.
- Not suitable for large samples : As the sample size increases, the t-distribution approaches the normal distribution, and other tests (such as the z-test) may be more appropriate.
- Degrees of freedom : The t-distribution requires the degrees of freedom to be specified, which can be a limitation if the sample size is small or the data is complex.

PROPERTIES

- **Mean :** The mean of the Student t-distribution is always 0.

- **Variance :** The variance of the Student t-distribution

$$Var[X] = \frac{v}{v-2}$$

T Test vs Chi Square

The difference between T Test and Chi Square Test is tabulated below.

T Test	Chi Square Test
•It is used to compare two group means.	•It is used for raw counts.
•The given data must be measured.	•The sample size must be large, maybe more than 50.
•It can only be used for two groups and not more.	•It compares an experimental result with a theoretical outcome.
•It is a parametric test.	•It is a non-parametric test.

DEGREE OF FREEDOM

- The degree of freedom (df) is refers to the number of independent pieces of information used to estimate a population parameter. It represents the number of data points that are free to vary when estimating a population parameter.
- In other words, df is the number of independent observations minus the number of parameters estimated.
- For example:- In a sample of size n, there are n-1 degrees of freedom when estimating the mean (since the mean is calculated from n observations, but one observation is used to estimate the population mean).

FISHER'S Z-DISTRIBUTION

ASSUMPTIONS

- **Normality** : The data should follow a normal distribution
- **Independence** : The observations should be independent of each other.
- **Identical Distribution** : The observations should come from the same distribution.
- **Equal Variances** : The variances of the two populations should be equal.
- **F-distribution** : The data should follow an F-distribution with m1 and m2 degrees of freedom.
- **No Outliers** : The data should not contain outliers or extreme values that can affect the analysis.

MERITS

- Fisher's Z-distribution can be used in various statistical tests and procedures, such as the F-test, process capability analysis, and quality control.
- The distribution is robust to minor deviations from normality and equal variances, making it a reliable choice in many practical applications.
- The probability density function (PDF) and cumulative distribution function (CDF) of Fisher's Z-distribution can be easily calculated using statistical software or programming languages.
- Fisher's Z-distribution can be applied to a wide range of fields, including engineering, economics, biology, and social sciences.
- The distribution has a well-established theoretical framework, making it easy to understand and work with.

DEMERITS

- The PDF and CDF of Fisher's Z-distribution involve complex mathematical functions, such as the gamma function, which can be challenging to work with.
- The parameters of Fisher's Z-distribution (m1 and m2) may not have a clear practical interpretation, making it difficult to understand the results in some cases.
- While the distribution is robust to minor deviations from normality, it can be sensitive to outliers or extreme values, which can affect the analysis.
- Fisher's Z-distribution is not suitable for data that is not normally distributed or has unequal variances, which can limit its applicability.
- Some methods involving Fisher's Z-distribution, such as simulation and resampling, can be computer-intensive and require significant computational resources.

F-DISTRIBUTION

INTRODUCTION

- The F-distribution is a continuous probability distribution that represents the ratio of two chi-squared distributions, each divided by its degrees of freedom.
- The F-distribution has two parameters, m1 and m2, which are the degrees of freedom of the two chi-squared distributions.
- The F-distribution is used to compare the variability of two populations, test hypotheses, and determine the significance of statistical models.
- The F-distribution is commonly used in statistical tests such as the F-test, Analysis of Variance (ANOVA), and regression analysis, as well as in quality control and finance.
- The F-distribution is asymmetric and skewed to the right, with a shape that depends on the degrees of freedom m1 and m2.

PDF OF F-DISTRIBUTION

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1}\ d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\, x\right)^{-\frac{d_1+d_2}{2}}$$

Which is the pdf of F-distribution with d_1 & d_2 degree of freedom

mean :

$$\mathrm{E}[X] = \frac{n_2}{n_2 - 2}$$

Variance :

$$\mathrm{Var}[X] = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$$

With n_1 & n_2 degree of freedom

MERITS

- The F-distribution is used in various statistical tests, such as ANOVA, regression, and hypothesis testing.
- The F-distribution is robust to minor deviations from normality and equal variances.
- The F-distribution can be used to compare the variability of two populations with different sample sizes.
- The F-distribution has a well-established theoretical framework, making it easy to understand and work with.
- The F-distribution can be easily calculated using statistical software or programming languages.

DEMERITS

- The F-distribution involves complex mathematical calculations, which can be challenging for non-statisticians.
- The F-distribution is sensitive to outliers and extreme values, which can affect the analysis.
- The F-distribution assumes normality and equal variances, which may not always be met in practice.
- The F-distribution can be difficult to interpret, especially for non-statisticians.
- The F-distribution is not easily visualized, making it difficult to understand the results.