UNIT – V

Student's t Distribution: - A family of probability distributions distinguished by their individual degrees of freedom, similar in form to the normal distribution, and used when the population standard deviation is unknown and sample size is relatively small (n<=30).

Degree of freedom: The number of values in a sample we can specify freely once we know something about the sample.

Characteristics of the t-distribution similar to the normal distribution

1. It is bell-shaped.
2. It is symmetric about the mean.
3. The mean, median, and mode are equal to 0 and are located at the center of the distribution.
4. The curve never touches the x axis.

Characteristics of the t-distribution that differ from the normal distribution

1. The variance is greater than 1.
2. The t-distribution is a family of curves based on the concept of degrees of freedom, which is related to sample size.
3. As the sample size increases, the t distribution approaches the standard normal distribution.

The *t*-distribution is symmetric about zero and its general shape is like the bell-shape of a normal distribution. However, the tails of the *t*-distribution can approach zero much more slowly than those of the normal distribution- i.e. the *t*-distribution is heavier tailed than the normal. The degrees of freedom define how heavy-tailed the *t*-distribution is.

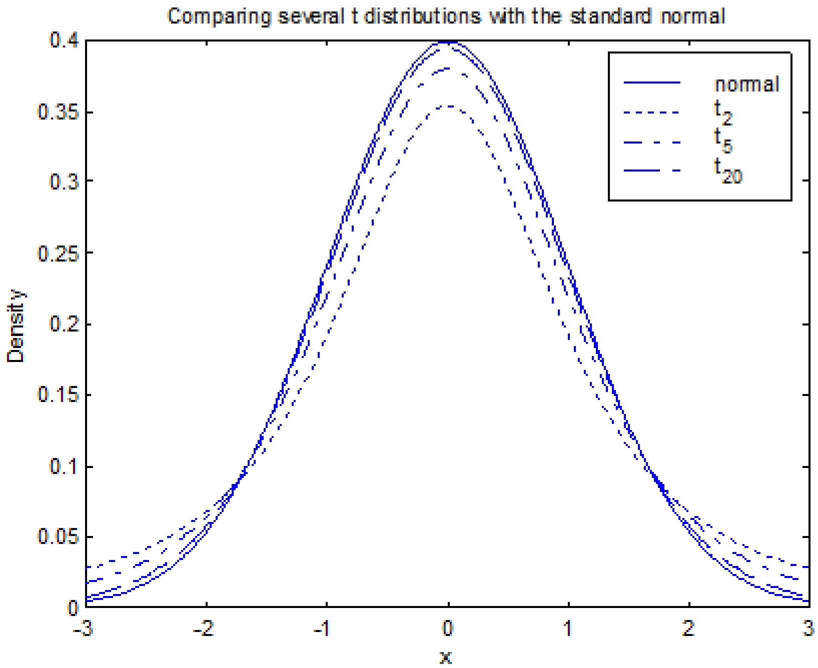
Note:

The *t*-distribution with $n = 1$ is sometimes referred to as the Cauchy distribution. This is so heavy tailed that its mean and variance do not exist! (This is because the integrals specifying the mean and variance are not absolutely convergent.)

Important note:

The density of a t-distribution converges to that of the standard normal as $n \to \infty$.

The diagram below shows how the t-distribution varies for different degrees of freedom.



Uses of t-test are:-

1. To test the population means when the sample is small and the population S.D.is unknown.
2. To test the equality of two sample means when the samples are small and population S.D. is unknown.
3. To test the difference in values of two dependent samples.
4. To test the significance of correlation coefficients.

The following are the important assumptions in t-test:-
1. The population from which the sample drawn is normal.
2. The sample observations are independent.
3. The population S.D.is known.
4. When the equality of two population means is tested, the samples are assumed to be independent and the population variance are assumed to be equal and unknown.

F Distribution: A family of distributions differentiated by two parameters (df-numerator, df-denominator), used primarily to test hypothesis regarding variances. The F-distribution is a *continuous* probability distribution, which means that it is defined for an *infinite* number of different values. The F-distribution can be used for several types of applications, including testing hypotheses about the equality of two population variances and testing the validity of a multiple regression equation.

BASIC PROPERTIES

A few of the more important features of this distribution are listed below:

- The F-distribution is a family of distributions. This means that there are an infinite number of different F-distributions. The particular F-distribution that we use for an application depends upon the number of degrees of freedom that our sample has. This feature of the F-distribution is similar to both the *t*-distribution and the chi-square distribution.
- The F-distribution is either zero or positive, so there are no negative values for *F*. This feature of the F-distribution is similar to the chi-square distribution.
- The F-distribution is skewed to the right. Thus this probability distribution is nonsymmetrical. This feature of the F-distribution is similar to the chi-square distribution.

These are some of the more important and easily identified features. We will look more closely at the degrees of freedom.

- For a *t* distribution the number of degrees of freedom is one less than our sample size. The number of degrees of freedom for an F-distribution is determined in a different manner than for a t-distribution or even chi-square distribution.
- The F-distribution is derived from a ratio involving two populations. There is a sample from each of these populations and thus there are degrees of freedom for both of these samples. In fact, we subtract one from both of the sample sizes to determine our two numbers of degrees of freedom.
- Statistics from these populations combine in a fraction for the F-statistic. Both the numerator and denominator have degrees of freedom. Rather than combining these two numbers into another number, we retain both of them. Therefore any use of an F-distribution table requires us to look up two different degrees of freedom.

USES OF THE F-DISTRIBUTION

- The F-distribution arises from inferential statistics concerning population variances. More specifically, we use an F-distribution when we are studying the ratio of the variances of two normally distributed populations.
- This type of distribution is also used in one factor analysis of variance (ANOVA). ANOVA is concerned with comparing the variation between several groups and variation within each group. To accomplish this we utilize a ratio of variances. This ratio of variances has the F-distribution.

- The F-distribution is a *continuous* probability distribution, which means that it is defined for an *infinite* number of different values. The F-distribution can be used for several types of applications, including testing hypotheses about the equality of two population variances and testing the validity of a multiple regression equation.

The F distribution is an asymmetric distribution that has a minimum value of 0, but no maximum value. The curve reaches a peak not far to the right of 0, and then gradually approaches the horizontal axis the larger the F value is. The F distribution approaches, but never quite touches the horizontal axis. The F distribution has two degrees of freedom, d1 for the numerator, d2 for the denominator. For each combination of these degrees of freedom there is a different F distribution. The F distribution is most spread out when the degrees of freedom are small. As the degrees of freedom increase, the F distribution the F distribution is less dispersed. Figure 1.1 shows the shape of the distribution. The F value is on the horizontal axis, with the probability for each F value being represented by the vertical axis. The shaded area in the diagram represents the level of significance α shown in the table. There is a different F distribution for each combination of the degrees of freedom of the numerator and denominator. Since there are so many F distributions, the F tables are organized somewhat differently than the tables for the other distributions. The three tables which follow are organized by the level of significance. The first table gives F values for that are associated with α = 0.10 of the area in the right tail of the distribution. The second table gives the F values for α = 0.05 of the area in the right tail, and the third table gives F values for the α = 0.01 level of significance. In each of these tables, the F values are given for various combinations of degrees of freedom. In order to use the F table, first select the significance level to be used, and then determine the appropriate combination of degrees of freedom. For example, if the α = 0.10 level of significance is selected, use the first F table. If there are 5 degrees of freedom in the numerator, and 7 degrees of freedom in the denominator, the F value from the table is 2.88. This means that there is exactly 0.10 of the area under the F curve that lies to the right of F = 2.88. When the significance level is α = 0.05, use the second F table. If there are 20 degrees of freedom in the numerator, and 5 degrees of freedom in the denominator, then the critical F value is 4.56. This could be written $F_{20,5;0.05} = 4.56$ That is, for 20 and 5 degrees of freedom, the F value that leaves exactly 0.05 of the area under the F curve in the right tail of the distribution is 4.56. For the α = 0.01 level of significance, the third F table is used. Suppose that there is 1 degree of freedom in the numerator and 12 degrees of freedom in the denominator. Then $F_{1,12;0.01} = 9.33$. An F value of 9.33 leaves exactly 0.01 of area under the curve in the right tail of the distribution when there are 1 and 12 degrees of freedom.

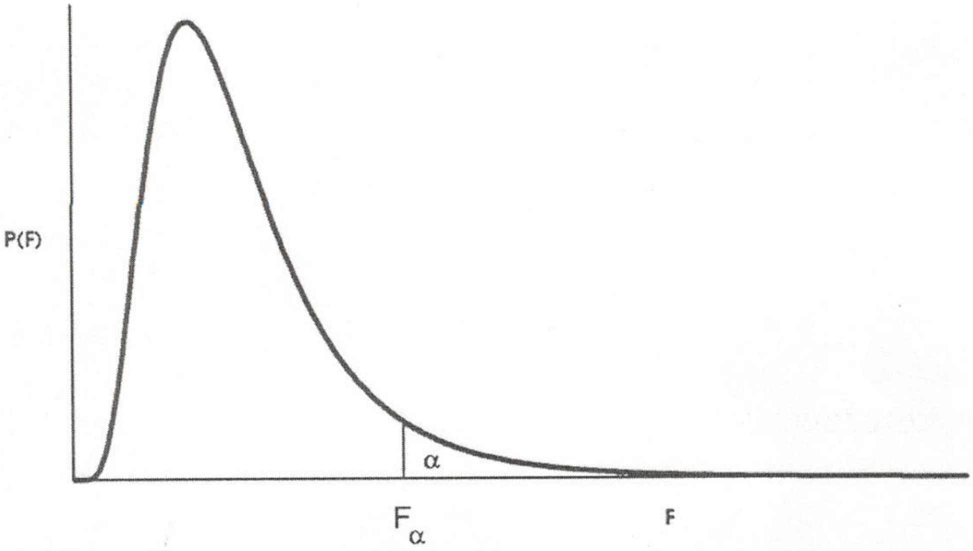
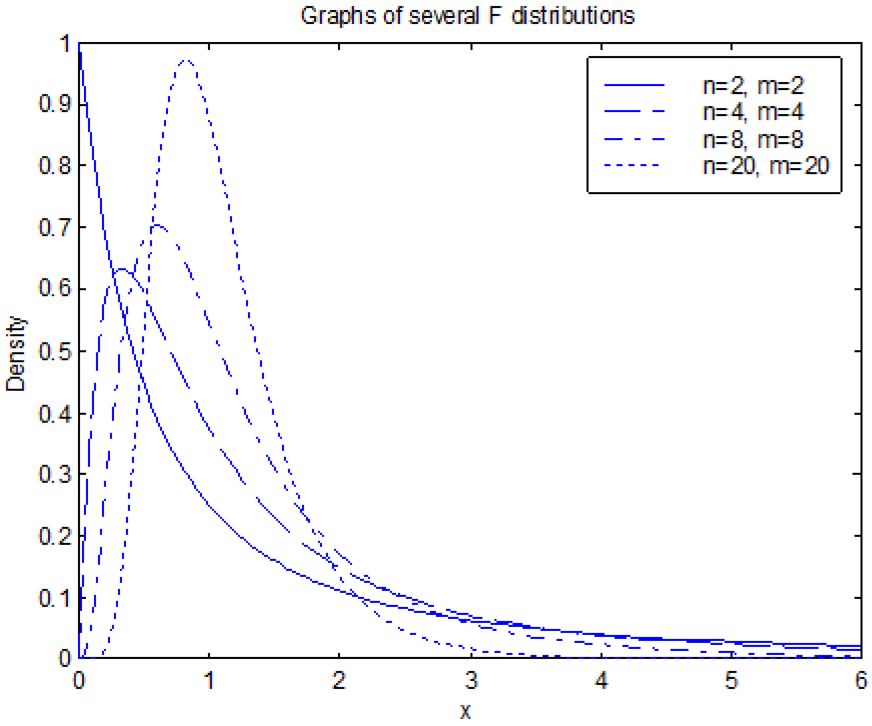
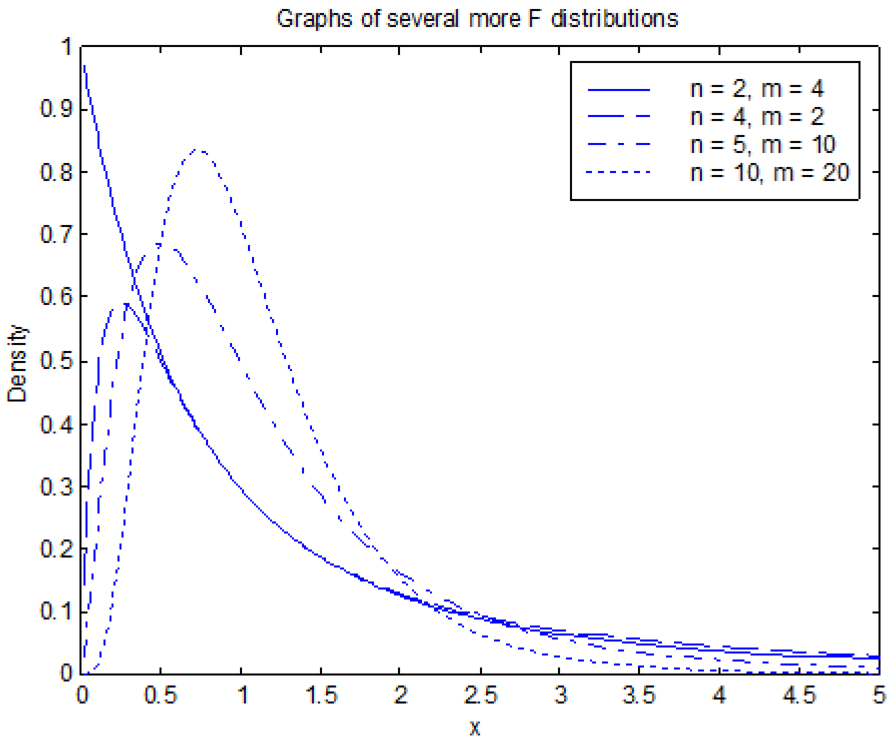


Figure K.1: The F distribution





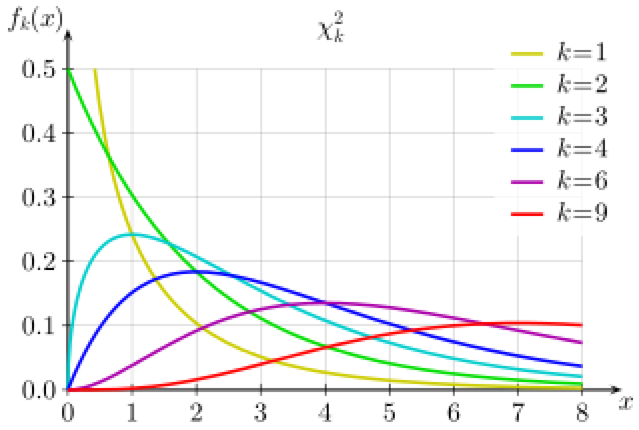
CHI-SQUARE DISTRIBUTION

In probability theory and statistics, the chi-squared distribution (also chi-square or χ2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-squared distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing or in construction of confidence intervals. When it is being distinguished from the more general non central chi-squared distribution, this distribution is sometimes called the central chi-squared distribution.

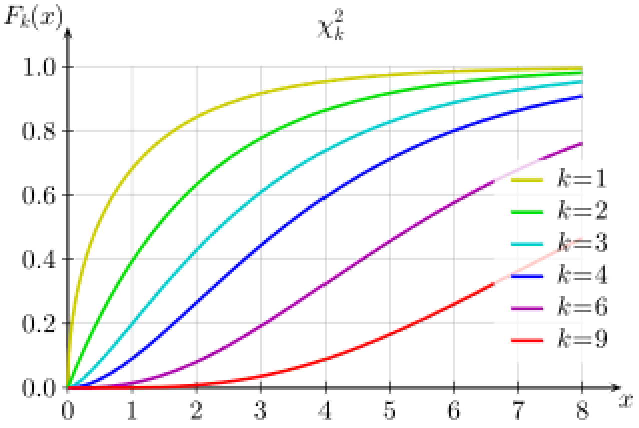
The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution, such as Friedman's analysis of variance by ranks.

chi-squared

Probability density function



Cumulative distribution function



A random variable X has a *chi-square distribution with n degrees of freedom* if it is a gamma random variable with parameters $m = n/2$ and $\beta = 2$, i.e $X \sim \Gamma(n/2,2)$. Therefore, its probability density function (pdf) has the form

$$f(t) \quad = \quad f(t;\, n) \quad = \quad \begin{cases} \dfrac{t^{(n/2)-1}e^{-t/2}}{2^{n/2}\Gamma(n/2)} & \text{if } t > 0 \\ 0 & \text{if } t < 0 \end{cases}$$

In this case we shall say X is a *chi-square random variable* with *n degrees of freedom* and write $X \sim \chi^{\square}(n)$. Usually n is assumed to be an integer, but we only assume $n > 0$.

There are also a few properties of the Chi-square distribution that you might find useful. The expected value of a Chi-square distribution is its degrees of freedom (mean = $\mu = df$), and its variance is 2 times its degrees of freedom. Thus, its standard deviation is the square root of 2 times the degrees of freedom ($\sigma^2 = 2 * df$ so $\sigma = \sqrt{2 * df}$). This frame of reference can help us assess if our observed statistic is unusual under the null hypothesis or somewhat consistent with the null hypothesis.

CHI-SQUARE TEST (χ^2 - test)

Chi-square tests: (the test of goodness of fit, the test of independence, and the test of homogeneity) are used to analyze categorical responses.
For all three tests the data are generally presented in the form of a contingency table (a rectangular array of numbers in cells). All three tests are based on the Chi-Square statistic:
The value of chi-square describes the magnitude of difference between observed frequencies and expected frequencies under certain assumptions. χ^2 value (χ^2 quantity) ranges from zero to infinity. It is zero when the expected frequencies and observed frequencies completely coincide. So greater the value of χ^2, greater is the discrepancy between observed and expected frequencies.

χ^2-test is a statistical test which tests the significance of difference between observed frequencies and corresponding theoretical frequencies of a distribution without any assumption about the distribution of the population. This is one of the simplest and most widely used nonparametric test in statistical work. This test was developed by Prof. Karl Pearson in 1990.

Uses of χ^2 - test

The uses of chi-square test are:-

1. Useful for the test of goodness of fit:- χ^2 - test can be used to test whether there is goodness of fit between the observed frequencies and expected frequencies.

2. Useful for the test of independence of attributes:- χ^2 test can be used to test whether two attributes are associated or not.

3. Useful for the test of homogeneity:- χ^2 -test is very useful t5o test whether two attributes are homogeneous or not.

4. Useful for testing given population variance:- χ^2-test can be used for testing whether the given population variance is acceptable on the basis of samples drawn from that population.

χ^2 -test as a test of goodness of fit:

As a non-parametric test, χ^2-test is mainly used to test the goodness of fit between the observed frequencies and expected frequencies.
Procedure:-
1. Set up mull hypothesis that there is goodness of fit between observed and expected frequencies.
2. Find the χ⍰ value using the following formula:-

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O = Observed frequencies
E = Expected frequencies
3. Compute the degree of freedom.

d. f. = n – r – 1

Where ‘r’ is the number of independent constraints to be satisfied by the frequencies
4. Obtain the table value corresponding to the lord of significance and degrees of freedom.
5. Decide whether to accept or reject the null hypothesis. If the calculated value is less than the table value, we accept the null hypothesis and conclude that there is goodness of fit. If the calculated value is more than the table value we reject the null hypothesis and conclude that there is no goodness of fit.

The **goodness of fit test** answers the question, “Do the data fit well compared to a specified distribution?” This test considers one categorical variable and assesses whether the proportion of sampled observations falling into each category matches well enough to the null distribution for the given problem. For instance, the null distribution might be specified by a manufacturer, a product label, or the results of a previous study. The null hypothesis for the goodness of fit test specifies this null distribution which describes the population proportion of observations in each category.

The **test of homogeneity** answers the question, “Do two or more populations have the same distribution for one categorical variable?” This test considers one categorical variable and assesses whether this variable is distributed the same in two (or more) different populations. The null hypothesis for the test of homogeneity is that the distribution of the categorical variable is the same for the two (or more) populations.

The **test of independence** answers the question, “Are two factors (or variables) independent for a population under study?” This test considers two categorical variables and assesses whether there is a relationship between these two variables for a single population. The null hypothesis for the test of independence is that the two categorical variables are independent (that is, they are not related) for the population of interest.