Applications of Chi-Square Distribution

There are various applications of Chi-square distribution. These include

1. Test of independence between two categorical variables
2. Test of homogeneity between two categorical variables
3. Test for multiple proportions (to see whether proportions are equal or not)
4. Test for goodness of fit (to check that a given data follows some particular probability disturbing or not).

Test for Independence (Categorical Data)

The chi-squared test can also be used to test the hypothesis of independence of two variables of classification.

Suppose that we wish to determine whether the opinions of the voting residents of the state of Illinois concerning a new tax reform are independent of their levels of income.

A random sample of 1000 registered voters from the state of Illinois are classified as to whether they are in a low, medium, or high income bracket and whether or not they favor a new tax reform. The observed frequencies are presented below, which is known as a contingency table.

A contingency table with r rows and c columns is referred to as an *r* x c table ("r x c" is read "r by c").

The row and column totals in Table above are called **marginal** frequencies.

Our decision to accept or reject the null hypothesis, *Ho,* of independence between a voter's opinion concerning the new tax reform and his or her level of income is based upon how good a fit we have between the observed frequencies in each of the 6 cells of Table above and the frequencies that we would expect for each cell under the assumption that *Ho* is true.

To find these expected frequencies, let us define the following events:

L: A person selected is in the low-income level.

M: A person selected is in the medium-income level.

H: A person selected is in the high-income level.

F: *A* person selected is for the new tax reform.

A: A person selected is against the new tax reform.

The expected frequencies are obtained by multiplying each cell probability by the total number of observations. As before, we round these frequencies to one decimal. Thus the expected number of low-income voters in our sample who favor the new tax reform is estimated to be

Expected frequency = (Column total x row total) / grant total

The expected frequency for each cell is recorded in parentheses beside the actual observed value in Table 2. Note that the expected frequencies in any row or column add up to the appropriate marginal total. In our example we need to compute only the two expected frequencies in the top row of Table 2 and then find the others by subtraction. The number of degrees of freedom associated with the chi-squared test used here is equal to the number of cell frequencies that may be filled in freely when we are given the marginal totals and the grand total, and in this illustration that number is 2. A simple formula providing the correct number of degrees of freedom is

u = (r - l) (c - l) .

Table 2: Observed and Expected Frequencies

	Income Level			
Tax Reform	Low	Medium	High	Total
For	182 (200.9)	213 (209.9)	203 (187.2)	598
Against	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

Hence, for our example, v =(2-1)(3-1)= 2 degrees of freedom. To test the null hypothesis of independence, we use the following decision criterion:

Calculate

$$\chi^2 = \sum_{i=1}^{rc} \frac{(o_i - e_i)^2}{e_i}$$

where the summation extends over all re cells in the r x c contingency table. If $\chi^2 > \chi_\alpha^2$ with v = (r-1)(c-1) degrees of freedom, reject the null hypothesis of independence at the a level of significance; otherwise, fail to reject the null hypothesis.

Applying this criterion to our example, we find that

$$\chi^2 = \frac{(182 - 200.9)^2}{200.9} + \frac{(213 - 209.9)^2}{209.9} + \frac{(203 - 187.2)^2}{187.2} + \frac{(154 - 135.1)^2}{135.1} + \frac{(138 - 141.1)^2}{141.1} + \frac{(110 - 125.8)^2}{125.8} = 7.86,$$

The null hypothesis is rejected and we conclude that a voter's opinion concerning the new tax reform and his or her level of income are not independent.

It is important to remember that the statistic on which we base our decision has a distribution that is only approximated by the chi-squared distribution.

The computed χ^2-values depend on the cell frequencies and consequently are discrete. The continuous chi-squared distribution seems to approximate the discrete sampling distribution of χ^2 very well, provided that the number of degrees of freedom is greater than 1. In a 2 x 2 contingency table, where we have only 1 degree of freedom, a correction called **Yates' correction for continuity** is applied. The corrected formula then becomes

$$\chi^2_{(Corrected)} = \sum_{i=1}^{rc} \frac{\left(|o_i - e_i| - 0.5\right)^2}{e_i}$$

If the expected cell frequencies are large, the corrected and uncorrected results are almost the same.
When the expected frequencies are between 5 and 10, Yates' correction should be applied.

For expected frequencies less than 5, the Fisher-Irwin exact test should be used. The Fisher-Irwin test may be avoided, however, by choosing a larger sample.

A discussion of this test may be found in *Basic Concepts of Probability and Statistics* by Hodges and Lehmann.

Test for Several Proportions:

The chi-squared statistic for testing for homogeneity is also applicable when testing the hypothesis that *k* binomial parameters have the same value. This is, therefore, an extension of the test for determining differences between two proportions to a test for determining differences among *k* proportions. Hence we are interested in testing the null hypothesis

$$H_0 : P_1 = P_2 = ... = P_k$$

against the alternative hypothesis, *Hi,* that the population proportions are *not equal.* Tb perform this test, we first observe independent random samples of size *n1, n2,…, nk* from the k populations and arrange the data as in the 2 x *k* contingency table. Table.

Depending on whether the sizes of the random samples were predetermined or occurred at random, the test procedure is identical to the test for homogeneity or the test for independence. Therefore, the expected cell frequencies are calculated as before and substituted together with the observed frequencies into the chi-squared statistic

$$\chi^2 = \sum_{i=1}^{rc} \frac{(o_i - e_i)^2}{e_i}$$

With v= (2-1).(k-1) degrees of freedom.

By selecting the appropriate upper-tail critical region of the form $\chi^2 > \chi^2_\alpha$ we can now reach a decision concerning H_0.

Example:1 In a shop study, a set of data was collected to determine whether or not the proportion of defectives produced by workers was the

same for the day, evening, or night shift worked. The data were collected and shown in Table.

Shift:	Day	Evening	Night
Defectives	45	55	70
Nondefectives	905	890	870

Use a 0.025 level of significance to determine if the proportion of defectives is the same for all three shifts.

Solution: Let P1, P2, and P3 represent the true proportion of defectives for the day, evening, and night shifts, respectively.

1. H_0: $P_1 = P_2 = P_3$
2. H_A: at least two of P_1 ,P_2, P_3 are not equal.
3. level of significance = 0.025.
4. Critical region: $\chi^2 > 7.378$ for $v = 2$ degrees of freedom.
5. Computations: Corresponding to the observed frequencies o1 = 45 and 02 = 55, we find

$$e_1 = \frac{(950)(170)}{2835} = 57.0 \quad \text{and} \quad e_2 = \frac{(945)(170)}{2835} = 56.7.$$

All other expected frequencies are found by subtraction and are displayed in Table below

Table 10.12: Observed and Expected Frequencies

Shift:	Day	Evening	Night	Total
Defectives	45 (57.0)	55 (56.7)	70 (56.3)	170
Nondefectives	905 (893.0)	890 (888.3)	870 (883.7)	2665
Total	950	945	940	2835

Now

$$\chi^2 = \frac{(45-57.0)^2}{57.0} + \frac{(55-56.7)^2}{56.7} + \frac{(70\text{-}56.3)^2}{56.3} + \frac{(905-893.0)^2}{893.0} + \frac{(890\text{-}888.3)^2}{888.3} + \frac{(870-883.7)^2}{883.7} = 6.29.$$

6. Decision: We do not reject H_0 at a= 0.025. Same damage.

Chi-Squared GOODNESS OF FIT TEST

A. For discrete distributions

In discrete probability distributions/ relative frequency distributions, each observation in a sample is classified as belonging to one of a finite number of categories

(e.g., blood type could be one of the four categories O, A, B, or AB). Let *pi* denoting the probability that any particular observation belongs in category *I* (or the proportion of the population belonging to category *i*). We then wish to test a null hypothesis that completely specifies the values of all the (such as, when there are four categories).

The test statistic is based on how different the observed numbers in the categories are from the corresponding expected numbers when H_0 is true. Because a decision will be reached by comparing the test statistic value to a critical value of the chi-squared distribution, the procedure is called a chi-squared goodness-of-fit test.

For a chi-squared test to be performed, the values of any unspecified parameters must be estimated from the sample data. The methods are then applied to test a null hypothesis that states that the sample comes from a particular family of distributions, such as the Poisson family (with m estimated from the sample) or the normal family (with parameters m and s estimated).

A test to determine if a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.

To illustrate, consider the tossing of a die. We hypothesize t h a t the die is honest, which is equivalent to testing the hypothesis that the distribution of outcomes is the discrete uniform distribution

Observed and Expected Frequencies of 120 Tosses of a Die

Face:	1	2	3	4	5	6
Observed	20	22	17	18	19	24
Expected	20	20	20	20	20	20

Suppose that the die is tossed 120 times and each outcome is recorded. Theoretically, if the die is balanced, we would expect, each face to occur 20 times. The results are given in Table above

By comparing the observed frequencies with the corresponding expected frequencies, we must decide whether these discrepancies are likely to occur as a result, of sampling fluctuations and the die is balanced, or the die is not honest and the distribution of outcomes is not uniform.

It is common practice to refer to each possible outcome of an experiment as a cell. Hence, in our illustration, we have 6 cells.

The appropriate statistic on which we base our decision criterion for an experiment involving *k* cells is defined by the following theorem.

A **goodness-of-fit** test between observed and expected frequencies is based on the quantity

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Where χ^2 is a value of a random variable whose sampling distribution is approximated very closely by the chi-squared distribution with v = k-1 degrees of freedom. The symbols o_i and e_i represent the observed and expected frequencies, respectively, for the i^{th} cell.

The number of degrees of freedom associated with the chi-squared distribution used here is equal to k-1, since there are only k-1 freely determined cell frequencies. That is, once k-1 cell frequencies are determined, so is the frequency for the k^{th} cell.

If the observed frequencies are close to the corresponding expected frequencies, the χ^2 -value will be small, indicating a good fit. If the observed frequencies differ considerably from the expected frequencies, the χ^2 value will be large and the fit is poor.

A good fit leads to the acceptance of H_o whereas a poor fit leads to its rejection. The critical region will, therefore, fall in the right tail of the chi-squared distribution. For a level of significance equal to α, we find the critical value χ^2_α from Table, and then $\chi^2 > \chi^2_\alpha$ constitutes the critical region.

** NOTE: **The decision criterion described here should not be used unless each of the expected frequencies is at least equal to 5.** This restriction may require the combining of adjacent cells resulting in a reduction in the number of degrees of freedom.

B. For Continuous distributions:

Consider the following table showing frequency distribution for battery life. Test for normality of the battery life dataset. (That's is to test either the battery life follow normal distribution or any other distribution).

Table 1.7: Relative Frequency Distribution of Battery Life

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

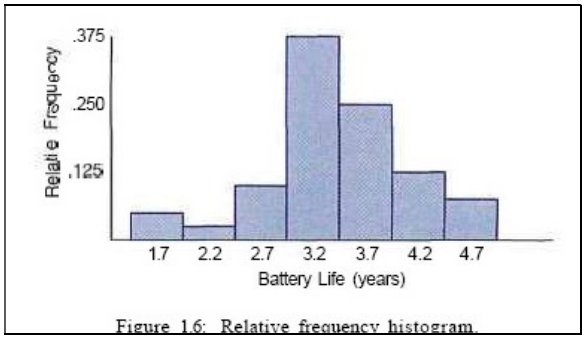

Figure 1.6: Relative frequency histogram.

Class Interval	Freq.	Area	Expected freq. (ei= N*Area)
1.45 - 1.95	2	**P(1.45 < x <1.95)=**	0.5
1.95 - 2.45	1	**P(1.95 < x <2.45)=**	2.1
2.45 - 2.95	4	**P(2.45 < x <2.95)=**	5.9
2.95 - 3.45	16	**P(2.95 < x <3.45)=**	10.3
3.45 - 3.95	10	**P(3.45 < x <3.95)=**	10.7
3.95 - 4.45	5	**P(3.95 < x <4.45)=**	7
4.45 - 4.95	3	**P(4.45 < x <+inf)= 1- P(x < 4.45)**	3.5

To test the hypothesis that the frequency distribution of battery lives given in Table 1.7 on page 23 may be approximated by a normal distribution with mean= 3.5 and standard deviation= 0.7. The expected frequencies for the 7 classes (cells), listed in Table, are obtained by computing the areas under the hypothesized normal curve that fall between the various class boundaries.

For example, the z-values corresponding to the boundaries of the fourth class (2.95- 3.95) are

$$z_1 = \frac{2.95-3.5}{0.7} = -0.79 \quad \text{and} \quad z_2 = \frac{3.45-3.5}{0.7} = -0.07,$$

Area= P(-0.79 < z < -0.07) = P(Z < -0.07) – P(Z< -0.79) = 0.4721 – 0.2148 = 0.2573

Hence the expected frequency for the fourth class is
e_4 = (0.2573)(N=40) = 10.3.
It is customary to round these frequencies to one decimal.

**** Procedure:** The expected frequency for the first class interval is obtained by using the total area under the normal curve to the left of the boundary 1.95. For the last class interval, we use the total area to the right of the boundary 4.45. All other expected frequencies are determined by the method described for the fourth class.

Note that we have combined adjacent classes, where the expected frequencies are less than 5. Consequently, the total number of intervals is reduced from 7 to 4, resulting in v = 3 degrees of freedom.

$$\chi^2 = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{103} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05.$$

Since the computed χ^2-value is less than $\chi^2_{0.05} = 7.815$ for 3 degrees of freedom, we have no reason to reject the null hypothesis and conclude that the normal distribution with $p. = 3.5$ and $a - 0.7$ provides a good fit. for the distribution of battery lives.

Many statistical procedures in practice depend, in a theoretical sense, on the assumption that the data gathered come from a specific distribution type. As we have already seen, the normality assumption is often made. Now we shall continue to make normality assumptions in order to provide a theoretical basis for certain tests and confidence intervals.(eg ANOVA, regression etc).

Test for Normal distribution

There are tests in the literature that are more powerful than the chi-squared test for testing normality. One such test is called **Geary's test**. This test is based on a very simple statistic which is a ratio of two estimators of the population standard deviation a. Suppose that a random sample $X_1, X_2, \ldots, X_n$, is taken from a normal distribution, N(μ, σ^2). Consider the ratio

$$U = \frac{\sqrt{\pi/2}\sum_{i=1}^{n}|X_i - \bar{X}|/n}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}}.$$

It's easily recognizable that the denominator is a reasonable estimator of a whether the distribution is normal or not. The numerator is a good estimator

of σ^2 if the distribution is normal but may overestimate or underestimate a when there are departures from normality. Thus, values of U differing considerably from 1.0 represent the signal that the hypothesis of normality should be rejected.

For large samples a reasonable test is based on approximate normality of U. The test statistic is then a standardization of U, given by

$$z = \frac{U - 1}{0.2661/\sqrt{n}}.$$

Of course, the test procedure involves the two-sided critical region. We compute a value of z from the data and do not reject the hypothesis of normality when $-Z_{\alpha/2} < Z < +Z_{\alpha/2}$.