Chapter- 6

**The Chi-Square Statistic**

**Objectives:**

* To introduce the chi- square and F distributions and learn how to use them in statistical inferences;
* To use the chi- square distribution to see whether two classifications of the same data are independent of each other;
* To use a chi- square test to check whether a particular collection of data is well described by a specified distribution;
* To use the chi- square distribution for confidence intervals and testing hypotheses about a single population variance;
* To compare more than two population means using analysis of variance (ANOVA); and
* To use the F distribution to test hypotheses about two population variances.

**Contents:**

* Basic Terminology Used in this Chapter;
* Chi- Square as a Test of Independence;
* Chi- Square as a Test of Goodness of Fit: Testing the Appropriateness of Distribution;
* Analysis of Variance (ANOVA);
* Inferences about a Population Variance; and
* Inferences about Two Population Variances.

**Basic Terminology**

***Chi- Square Distribution:***

A family of probability distributions, differentiated by their degrees of freedom, used to test a number of hypotheses about variances, proportions and distributional goodness of fit.

***Goodness- of- Fit Test:***

A statistical test for determining whether there is a significant difference between an observed frequency distribution and a theoretical probability distribution hypothesized to describe the observed distribution.

***Test of Independence:***

A statistical test of proportions of frequencies to determine whether membership in categories of one variable is different as a function of membership in the categories of a second variable.

***Expected Frequencies:***

The frequencies we would expect to see in a contingency table or frequency distribution if the null hypothesis is true.

***Analysis of Variance (ANOVA):***

A statistical technique used to test the equality of three or more sample means and thus make inferences as to whether the samples come from populations having the same mean.

***F- Distribution:***

A family of distributions differentiated by two parameters (df- numerator, df- denominator), used primarily to test hypotheses regarding variances.

***R- Ratio:***

A ratio used in the analysis of variance, among other tests, to compare the magnitude of two estimates of the population variance to determine whether the two estimates are approximately equal; in ANOVA, the ratio of between – column variance to within- column variance is used.

***Between- Column Variance:***

An estimate of the population variance derived from the variance among the sample means.

***Within- Column Variance:***

An estimate of the population variance based on the variances within the k samples, using a weighted average of k sample variances.

***Contingency Table:***

A table having R rows and C columns. Each row corresponds to a level of one variable, each column to a level of another variable. Entries in the body of the table are the frequencies with which each variable combination occurred.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Introduction:**

This chapter introduces two *non-parametric* hypothesis tests using the chi-square statistic: *the chi-square test for goodness of fit* and *the chi-square test for independence*.

The term "non-parametric" refers to the fact that the chi‑square tests do not require assumptions about population parameters nor do they test hypotheses about population parameters.

The t- tests and analysis of variance are **parametric tests** and they do include assumptions about parameters and hypotheses about parameters.

The most obvious difference between the chi‑square tests and the other hypothesis tests we have considered (t and ANOVA) is the nature of the data.

Chi-square (χ2) procedures measures the differences between observed (O) and expected (E) frequencies of nominal variables, in which subjects are grouped in categories or cells. There are three basic uses of chi-square analysis, *the Goodness of Fit Test* (used with a single nominal variable), *the Test of Independence (*used with two nominal variables) and the test of homogeneity. These types of chi-square use the same formula.

The chi-square formula is as follows:

χ2 = ∑ (O-E)2 ÷ E

Where O = observed frequency (the actual count -- in a given cell);

E = expected frequency (a theoretical count -- for that cell). Its value must be computed.

For chi‑square, the data are frequencies rather than numerical scores.

**Conditions or Assumptions for Applying χ2 Test:**

1. Large number (generally not less than 50) of observations or frequencies;
2. Expected frequency should not be small (less than 5). If it is less than 5, then frequencies taken from adjacent items or cells are pooled in order to make it 5 or more than 5.Yate’s correction may also be applied in such case;
3. Data should be in original units such as percentage or proportion;
4. Random sampling; and
5. Events should be mutually exclusive.

**The Chi-Square Test for Goodness-of-Fit**

The Goodness of Fit Test is applied to a single nominal variable and determines whether the frequencies we observe in k categories fit what we might expect. Some textbooks call this procedure *the Badness of Fit Test* because a significant χ2 value means that Observed counts do not fit what we expect. The Goodness of Fit Test can be applied with *equal or proportional expected frequencies* (EE, PE).

***Equal Expected (EE) Frequencies:***

Equal expected frequencies are computed by dividing the number of subjects (N) by the number of categories (k) in the variable. A classic example of equal expected frequencies is testing the fairness of a die. If a die is fair, we would expect equal tallies of faces over a series of rolls.

**The Example of a Die:**

Let’s say I roll a real die 120 times (N) and count the number of times each face (k = 6) comes up. The number “1” comes up 17 times, the number “2” 21 times, “3” 22 times, “4” 19 times, “5” 16 times, and “6” 25 times. Results are listed under the “O” column below. We would Expect a count of 20 (E=N/k) for each of the six faces (1-6). This E value of 20 is listed under the “E” column below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | O | E | O- E | (O-E)2 | (O-E)2 ÷ E |
| 1 | 17 | 20 | -3 | 9 | 9/20 = 0.45 |
| 2 | 21 | 20 | 1 | 1 | 1/20 = 0.05 |
| 3 | 22 | 20 | 2 | 4 | 4/20 = 0.20 |
| 4 | 19 | 20 | -1 | 1 | 1/20 = 0.05 |
| 5 | 16 | 20 | -4 | 16 | 16/20 = 0.80 |
| 6 | 25 | 20 | 5 | 25 | 25/20 = 1.25 |
|  | 120 | 120 | 0 |  | χ2 = 2.80 |

The table above shows the step-by-step procedure in computing the chi-square formula. Notice that both **O** and **E** columns add to the same value (N=120).

**Testing the Chi Square Value:**

The computed value of χ***2*** is compared to the appropriate critical value. The critical value is found in the Chi-square Table. **Using** α **and df**, locate the critical value from the table. For the Goodness of Fit Test, the degrees of freedom (df) equal the number of categories (k) minus one (**df = k-1**). In our example above, the critical value (α=0.05, df =5) is **11.07**. Since the computed value (**2.80**) is less than the critical value (11.07), we declare the χ***2 not significant***.

What does this non-significant χ***2*** mean in English? The observed frequencies of the six categories of die rolls do not significantly differ from the expected frequencies. The observed frequencies have a “good fit” with what was expected. Or, simply stated, **“The die is fair.”** Had the computed value been greater than 11.07, the χ***2*** would have been declared significant. This would mean that the difference between observed and expected values is greater than we expect by chance. The observed frequencies would have a “bad fit” with what was expected. Or simply stated, “The die is loaded.”

Equal E is usually an unrealistic assumption of the break-down of categories. A better approach is to compute proportional expected frequencies (PE).

***Proportional Expected (PE) Frequencies:***

With proportional expected frequencies, the expected values are derived from a known population. Suppose you are in an Advanced Greek class of 100 students. You notice a large number of women in the class, and wonder if there are more women in the class than one might expect, given the student population. Using equal E’s, you would use the value (E=N/k) of 50. But you know that women make up only 15% of the student population. This gives you expected frequencies of 15 women (.15 x 100) and 85 men (.85 x 100). This latter design is far more accurate than the EE value of 50.

**The Example of Political Party Preference**

Suppose you want to study whether political party preference has changed since the last Presidential election. A poll of 1200 voters taken four years before showed the following breakdown: 500 Republicans, 400 Democrats, and 300 Independents. The ratio equals 5:4:3. In your present study, you poll 600 registered voters and find 322 Republicans, 184 Democrats, and 94 Independents. The null hypothesis for this study is that **party preference has not changed in four years.** That is, your hypothesis is that the present observed preferences are in a ratio of 5:4:3.

**Computing the Chi Square Value**

Compute the expected frequencies as follows. The ratio of 5:4:3 means there are 5+4+3=12 parts. Twelve parts divided into 600 voters yield *50 voters per part* (600/12=50).

The first category, Republicans, has 5 parts (**5**:4:3), or 5x50=***250* E**xpected voters. The second, Democrats, has 4 (5:**4**:3) parts, or 4x50=**200 E**xpected voters. The third, Independents, has 3 parts (5:4:**3**), or 3x50=**150 E**xpected voters. Putting this in a table as before, we have the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Party | O | E | (O-E) | (O-E)2 | (O-E)2 ÷ E |
| Republic | 322 | 250 | 72 | 5184 | 5184/250 = 20.74 |
| Democratic | 184 | 200 | -16 | 256 | 256/200 = 1.28 |
| Independent | 94 | 150 | -56 | 3136 | 3136/150 = 20.91 |
|  | 600 | 600 | 0 |  | **χ2 = (O-E)2/E= 42.93** |

Notice that both **O** and **E** columns add to 600 (N). Notice that the **O-E** column adds to zero. Notice that the E values are unequal, reflecting the 5:4:3 ratio derived from the earlier poll. The resulting χ***2*** value equals **42.93**.

**Testing the Chi Square**

The critical value (α=0.05, df = 2) is **5.991**. Since the computed value of **42.93** is greater than the critical value of 5.991, we declare the **chi-square value significant**. The observed values **do not fit** the expected values.

Since the recent poll does not fit the ratio of 5:4:3 found in the earlier poll, we can say that ***party preference has changed over the last four years****.*

**The Chi-Square Test for Independence**

The second chi-square test, the **chi-square test for independence**, can be used and interpreted in two different ways:

1. Testing hypotheses about the relationship between two variables in a population, or
2. Testing hypotheses about differences between proportions for two or more populations.

Although the two versions of the test for independence appear to be different, they are equivalent and they are interchangeable.

The first version of the test emphasizes the relationship between chi-square and a correlation, because both procedures examine the relationship between two variables.

The second version of the test emphasizes the relationship between chi-square and an independent-measures t- test (or ANOVA) because both tests use data from two (or more) samples to test hypotheses about the difference between two (or more) populations.

The first version of the chi-square test for independence views the data as one sample in which each individual is classified on two different variables.

The data are usually presented in a matrix with the categories for one variable defining the rows and the categories of the second variable defining the columns.

The data, called **observed frequencies**, simply show how many individuals from the sample are in each cell of the matrix.

The null hypothesis for this test states that there is no relationship between the two variables; that is, the two variables are independent.

The second version of the test for independence views the data as two (or more) separate samples representing the different populations being compared.

The same variable is measured for each sample by classifying individual subjects into categories of the variable.

The data are presented in a matrix with the different samples defining the rows and the categories of the variable defining the columns.

The data, again called **observed frequencies**, show how many individuals are in each cell of the matrix.

The null hypothesis for this test states that the proportions (the distribution across categories) are the same for all of the populations.

Both chi-square tests use the same statistic. The calculation of the chi-square statistic requires two steps:

1. The null hypothesis is used to construct an idealized sample distribution of **expected frequencies** that describes how the sample would look if the data were in perfect agreement with the null hypothesis.

For the goodness of fit test, the expected frequency for each category is obtained by expected frequency = fe = pn (p is the proportion from the null hypothesis and n is the size of the sample)

For the test for independence, the expected frequency for each cell in the matrix is obtained by

Expected frequency = fe =

Where;

Row total = sum of all frequencies in the row

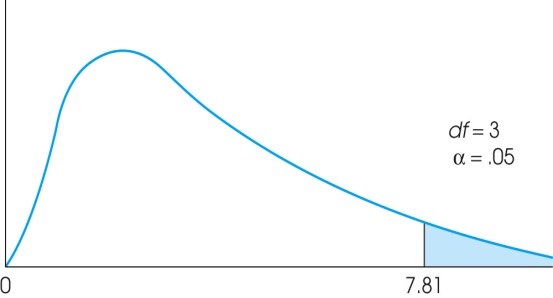
Column total = sum of all frequencies in the column

n = overall sample size

**Decision Rule:**

If χ2 > χ2U, reject H0, otherwise, do not reject H0

Where χ2U is from the chi-squared distribution with (r – 1)(c – 1) degrees of freedom



1. A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies from H0) and the actual sample data (the observed frequencies = fo).

A large discrepancy results in a large value for chi-square and indicates that the data do not fit the null hypothesis and the hypothesis should be rejected.

The calculation of chi-square is the same for all chi-square tests:

Chi-square = χ2 =

The fact that chi‑square tests do not require scores from an interval or ratio scale makes these tests a valuable alternative to the t- tests, ANOVA, or correlation, because they can be used with data measured on a nominal or an ordinal scale.

***Example:***

**The meal plan selected by 200 students is shown below:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class  Standing | **Number of meals per week** | | | **Total** |
| 20/week | 10/week | none |
| Fresh. | 24 | 32 | 14 | **70** |
| Soph. | 22 | 26 | 12 | **60** |
| Junior | 10 | 14 | 6 | **30** |
| Senior | 14 | 16 | 10 | **40** |
| **Total** | **70** | **88** | **42** | **200** |

**The hypothesis to be tested is:**

H0: Meal plan and class standing are independent (i.e., there is no relationship between them)

H1: Meal plan and class standing are dependent (i.e., there is a relationship between them)

**Example: Expected Cell Frequencies**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class  Standing | **Number of meals**  **per week** | | | **Total** |
| 20/wk | 10/wk | none |
| Fresh. | 24.5 | 30.8 | 14.7 | **70** |
| Soph. | 21.0 | 26.4 | 12.6 | **60** |
| Junior | 10.5 | 13.2 | 6.3 | **30** |
| Senior | 14.0 | 17.6 | 8.4 | **40** |
| **Total** | **70** | **88** | **42** | **200** |



**The test statistic value is:**



χ2U = 12.592 for α = 0.05 from the chi-squared distribution with (4 – 1)(3 – 1) = 6 degrees of freedom

**Decision and Interpretation:**



Decision Rule:

If χ2 > 12.592, reject H0, otherwise, do not reject H0

Here,

χ2 = 0.709 < χ2U = 12.592, so do not reject H0

Conclusion: there is not sufficient evidence that meal plan and class standing are related at α = 0.05

**Analysis of Variance (ANOVA)**

To test the significance of mean of one sample or significance of difference of means of two samples, t- test or χ2- test are very useful. But if there are more than two samples, the method of analysis of variance is used.

**Components of Total Variance:**

The total variation is split into two components- (a) Variance between samples and (b) Variance within samples, i.e.,

*Total Variance = Variance between samples + Variance within samples.*

**Assumptions:**

1. Normality;
2. Independence;
3. Additive Property.

**Uses or Applications or Importance of ANOVA:**

1. To test the significance of differences between means of more than two samples;
2. To test the significance of differences between variances;
3. Use in two- way classification;
4. To test the significance of correlation and regression.