# MODULE 21

# CHI-SQUARE TEST

## Contents

Situations where a categorical response variable is recorded would be summarized with a frequency or percentage table (see Modules 8 and 10). The appropriate test statistic in these situations is a chi-square rather than a t. The Chi-Square Test test statistic follows a chi-square distribution, which is introduced below. The rest of this module is dedicated to the general Chi-Square Test where the distribution of a categorical response variable is compared between two or more groups (or populations). The related goodness-of-fit test for a categorical response recorded for only one group (or population) is introduced in Module 22.

## 21.1 Chi-Square Distribution

A chi-square ($\chi^2$) distribution is generally right-skewed (Figure 21.1), with the exact shape dictated by the degrees-of-freedom (df; as df increase, the sharpness of the skew decreases; Figure 21.1). In its simplest form, the $\chi^2$ distribution arises as a sampling distribution for the $\chi^2$ test statistic,

$$\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$$

where "Observed" and "Expected" represent the observed and expected individuals in the cells of frequency tables (see Module 8 and Module 10) and "cells" generically represents the number of cells in one of these tables. Thus, the $\chi^2$ distribution arises from comparing frequencies in two tables.[1]

[1] Subsequent sections demonstrate how this test statistic is used to compare observed frequencies (i.e., from a sample) to a table of expected frequencies (i.e., from a null hypothesis).

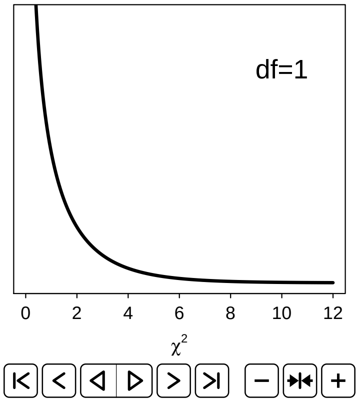

Figure 21.1. $\chi^2$ distributions with varying degrees-of-freedom.

Unlike the normal and t distributions, the $\chi^2$ distribution always represents the two-tailed situation, although the "two tails" will appear as one tail on the right side of the distribution. The simplest explanation for this characteristic is that the "squaring" in the calculation of the $\chi^2$ test statistic results in what would be a "negative tail" being "folded over" onto what is the "positive tail." Thus, all probability (i.e., area) calculations on a $\chi^2$ distribution represent the two-tailed alternative hypotheses.

Proportional areas on a $\chi^2$ distribution are computed with `disrib()`, similarly to what was described for normal and t distributions in Modules 9, 13, and 19. The major difference for using `distrib()` with a $\chi^2$ distribution is that `distrib="chisq"` must be used and the degrees-of-freedom must be given to `df=` (how to find the df will be discussed in subsequent sections). In addition, if calculating a p-value, then `lower.tail=FALSE` is always used because the upper-tail probability represents the two-tailed alternative hypothesis inherent to all Chi-Square Tests. For example, the area right of $\chi^2 = 6.456$ on a $\chi^2$ distribution with 2 df is 0.0396 (Figure 21.2).

```
> ( distrib(6.456,distrib="chisq",df=2,lower.tail=FALSE) )
[1] 0.03963669
```

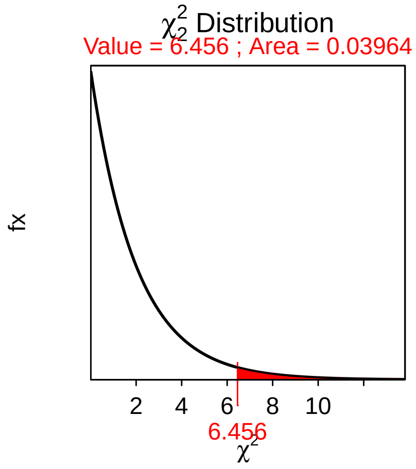

Figure 21.2. Depiction of the area to the right of $\chi^2 = 6.456$ on a $\chi^2$ distribution with 2 df.

## 21.2 Chi-Square Test Specifics

Researchers commonly want to compare the distribution of individuals into the levels of a categorical variable among two or more groups (or populations). For example, researchers may want to determine if the disribution of failing students differs between males and females, if the distribution of kids playing sports differs between kids from high- or low-income families, if the distribution of four major plant species differs between two locations, or if the distribution of responses to a five-choice question differs between respondents from neighboring counties. All of these questions have a categorical response variable (fail or not, play sport or not, plant species, answer to five-choice question) compared among two or more groups (gender, income category, two locations, neighboring counties). The Chi-Square Test, the subject of this module, can be used for each of these situations.[2]

### 21.2.1 Hypotheses

The statistical hypotheses for a Chi-Square Test are "wordy." To explore this, let's first assume that a two-way frequency table (see Module 10) will summarize the data where the rows correspond to separate groups and the columns correspond to levels of the response variable. In this organization, the Chi-Square Test null hypothesis is that the row percentages are equal – i.e., "the percentage distribution of individuals into the levels of the response variable is the same for all groups." The alternative hypothesis states that there is some difference among the row percentages – i.e., "the percentage distribution of individuals into the levels of the response variable is NOT the same for all groups."

As one example (more are shown below), consider the following:

> *An association of Christmas tree growers in Indiana sponsored a survey of Indiana households to help improve the marketing of Christmas trees. Of the 261 rural households, 64 had a natural tree (as compared to an artificial tree). Of the 160 urban households, 89 had a natural tree. Use these results to determine, at the 10% level, if the distribution of households with a natural tree differed between rural and urban households.*

The hypotheses for this situation are,

$H_0$ : "the distrubution of households into the tree types is the same for urban and rural households"
$H_A$ : "the distrubution of households into the tree types is NOT the same for urban and rural households"

### 21.2.2 Tables

As noted above, all two-way frequency tables used for a Chi-Square Test will be organized such that the response variable forms the columns and the groups to be compared form the rows. With this organization, the row-percentage table becomes the table of primary interest because it relates directly to the hypotheses described above. The question of a Chi-Square Test then becomes one of determining whether each row of the row-percentage table is equal, given sampling variability.

The observed raw data must be organized into a two-way frequency table as described in Module 10. For example, the Christmas tree data is summarized as in Table 21.1. The actual calculations for a Chi-Square Test are performed on this observed table. However, the hypothesis test, as described above, is best viewed as a method to determine if each row of the row-percentage table is statistically equivalent or not. Thus, the row-percentage table computed from the frequency table is useful when interpreting the results of a Chi-Square Test (Table 21.2).

[2] The Chi-Square Test is quite flexible and can be derived from different types of hypotheses than those described here.

Table 21.1. Frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

| | Tree Type | | |
|---|---|---|---|
| Household | Natural | Artificial | |
| Urban | 89 | 172 | **261** |
| Rural | 64 | 96 | **160** |
| | **153** | **268** | **421** |

Table 21.2. Percentage of individuals within urban and rural households that have a natural or an artificial Christmas tree.

| | Tree Type | | |
|---|---|---|---|
| Household | Natural | Artificial | |
| Urban | 34.1 | 65.9 | **100.0** |
| Rural | 40.0 | 60.0 | **100.0** |
| | **36.3** | **63.7** | **100.0** |

The Chi-Square Test requires constructing a table of expected values that are derived from the null hypothesis. Specifically, the "expected" table contains the expected frequency of individuals in each level of the response variable for each group assuming that the distribution of responses does not differ among groups. These expected table are computed from the margins of the observed table, but are best explained with an illustrative example.

In the Christmas tree example, the null hypothesis states that there is no difference in the distribution of households with a natural tree between the rural and urban areas. Thus, under this null hypothesis, one would expect the proportion (or percentage) of households with a natural tree to be the same in both groups. The proportion of households with a natural tree, regardless of location, is $\frac{153}{421}$=0.363. Thus, under the null hypothesis, the proportion of rural AND the proportion of urban households with a natural tree is 0.363. Because there is a different number of urban and rural households in the study, the actual NUMBER (rather than proportion) of households expected to have a natural tree will differ. The NUMBER of urban households expected to HAVE a natural tree is found by multiplying the number of urban households by the common proportion computed above – i.e., $261 * 0.363$=94.743. The remaining urban households would be expected to NOT have a natural tree – i.e., $261 - 94.743$=$261(1 - 0.363)$=166.257. Similar calculations are made for the rural households (i.e., $160*0.363 = 58.080$ expected to have a natural tree and $160*(1-0.363) = 101.920$ expected to NOT have a natural tree.

These expected frequencies are computed directly and easily from the marginal totals of the observed frequency table (Table 21.1). For example, substituting the fractional representation of the decimal proportions into the calculation of the expected number of urban households with a natural tree gives $261 * \frac{153}{421} = \frac{261*153}{421}$=94.853[3]. A close examination of this formula and the marginal totals in Table 21.1 shows that this value is equal to the product of the corresponding row and column marginal totals in the observed table divided by the total number of individuals. The other expected values follow a similar pattern as follows,

- $261 * \frac{268}{421} = \frac{261*268}{421} = 166.147$ urban households to NOT have a natural tree.
- $160 * \frac{153}{421} = \frac{160*153}{421} = 58.147$ rural households to have a natural tree.
- $160 * \frac{268}{421} = \frac{160*268}{421} = 101.853$ rural households to NOT have a natural tree.

Thus, all expected values in a Chi-Square Test are calculated by multiplying the row and column totals of the frequency table and dividing by the total number of individuals. These expected values are summarized in a two-way table, called the expected frequencies table (Table 21.3).

[3] Note a slight difference here because 0.363 was rounded to three decimals, whereas the fraction is not rounded.

Table 21.3. The expected frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

| | Tree Type | | |
|---|---|---|---|
| Household | Natural | Artificial | |
| Urban | 94.853 | 166.147 | **261** |
| Rural | 58.147 | 101.853 | **160** |
| | **153** | **268** | **421** |

### 21.2.3 Specifics

The Chi-Square Test is characterized by a categorical response variable recorded for two or more groups (or populations). The specifics of the Chi-Square Test are in Table 21.4.

Table 21.4. Characteristics of a Chi-Square Test.

- **Null Hypothesis:** "The distribution of individuals into the levels of the response variable is the same for all groups"
- **Alternative Hypothesis:** "The distribution of individuals into the levels of the response variable is NOT the same for all groups."
- **Statistic:** Observed frequency table.
- **Test Statistic:** $\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$
- **df:** $(r-1)(c-1)$ where $r$ = number of rows and $c$ = number of columns
- **Assumptions:** Expected value for each category is $\geq 5$.
- **Use with:** Categorical response, two or more groups (or populations).

In general, a confidence region is not constructed for a Chi-Square Test because of the complexity of the statistics and parameter. Thus, in this course, Step 11 for a hypothesis test will not be computed for a Chi-Square Test.

### 21.2.4 Example – Christmas Trees

Below are the 11-steps (Section 18.1) for a full hypothesis test for the Christmas tree example.

1. $\alpha$=0.10.
2. $H_0$: "distribution of households by type of tree is the same for urban and rural households" vs. $H_A$: "distribution of households by type of tree is NOT the same for urban and rural households."
3. A Chi-Square Test is required because (i) a categorical response variable was recorded (type of tree) and (ii) two groups are being compared (urban and rural households).
4. The data appear to be part of an observational study with no clear indication of randomization.
5. The expected frequency in each of the four cells is greater than five (Table 21.3).
6. The observed frequency table is in Table 21.1.
7. $\chi^2 = \frac{(89-94.853)^2}{94.853} + \frac{(172-166.147)^2}{166.147} + \frac{(64-58.147)^2}{58.147} + \frac{(96-101.853)^2}{101.853} = 0.3611 + 0.2062 + 0.5891 + 0.3363 = 1.4927$ with 1 df.
8. p-value=0.2218.
9. $H_0$ is not rejected because the p-value is $> \alpha$.
10. There does not appear to be a significant difference in the distribution of Christmas tree types among rural and urban households.
11. Not performed for Chi-Square Test.

**R Appendix:**

```
( distrib(1.4927,distrib="chisq",df=1,lower.tail=FALSE) )
```

## 21.3 Chi-Square test in R (Raw Data)

The data for a Chi-Square Test may be computed from raw data on individuals (this section) or entered from summarized data (see Section 21.4). Raw data must be in stacked format where one column in the data.frame represents the response variable and another column represents the groups (see Sections 4.3.2 and 20.3). Raw data must be summarized into a two-way frequency table with `xtabs()` as described in Module 10. The two-way table must contain frequencies, not proportions or percentages (don't use `percTable()`), without marginal totals (don't use `addMargins()`).

The Chi-Square Test is performed with `chisq.test()`, which takes an observed frequency table either entered through `matrix()` or summarized with `xtabs()` as the first argument. The only other argument needed is `correct=FALSE` so that the continuity correction is not used.[4] The results of `chisq.test()` should be assigned to an object. The Chi-Square test statistic and p-value are extracted by simply printing the saved object. The expected frequency table is returned by appending `$expected` to the saved object.

Rejecting the null hypothesis in a Chi-Square Test indicates that there is some difference in the distribution of individuals into the levels of the response variable among some of the groups. However, rejecting the null hypothesis does not indicate which groups are different. In addition, as mentioned previously, confidence intervals are generally not performed with a Chi-Square Test. A post-hoc method for helping determine which groups differ is obtained by observing the Pearson residuals.

A Pearson residual is computed for each cell in the table as,

$$\frac{Observed - Expected}{\sqrt{Expected}}$$

which is the appropriately signed square root of the parts in the $\chi^2$ test statistic calculation. Therefore, cells that have Pearson residuals far from zero contributed substantially to the large $\chi^2$ test statistic that resulted in a small p-value and the ultimate rejection of $H_0$. Patterns in where the large Pearson residuals are found may allow one to qualitatively determine which groups differ and, thus, which levels of the response differ the most. This process will be illustrated more fully in the examples and review exercises. The Pearson residuals are obtained from the saved `chisq.test()` object by appending `$residuals`.

### 21.3.1 Example - Father Present at Birth

Below are the 11-steps (Section 18.1) for completing a full hypothesis test for the following situation:

> *Daniel Weiss (in "100% American") reported the results of a survey of 300 first-time fathers from four different hospitals (labeled as A, B, C, and D). Each father was asked if he was present (or not) in the delivery room when his child was born. The results of the survey are in FatherPresent.csv. Use these data to determine if there is a difference, at the 5% level, in the proportion of fathers present in the delivery room among the four hospitals.*

[4] The continuity correction is not used here simply so that the results using R will match hand-calculations. The continuity correction should usually be used.

1. $\alpha$=0.05.
2. $H_0$:"distribution of fathers presence (or not) during the birth of their child is the same for all four hospitals" vs. $H_A$ : "the distribution of fathers presence during the birth of their child is NOT the same for all four hospitals."
3. A Chi-Square Test is required because (i) a categorical variable (present or absent) was recorded and (ii) four groups are being compared (the hospitals).
4. The data appear to be part of an observational study with no clear indication of randomization (likely a voluntary response survey).
5. There are at least five individuals in each cell of the expected table (Table 21.5).
6. The statistic is the observed frequency table (Table 21.6).
7. $\chi^2$=5.000 with 3 df (Table 21.7).
8. p-value=0.1718 (Table 21.7).
9. $H_0$ is not rejected because the p-value is $> \alpha$.
10. The distribution of father's presence (or not) at their child's birth does not seem to differ significantly among hospitals where that birth occurred. For comparative purposes, the row-percentage table is in Table 21.8.

**R Appendix:**

```
setwd("c:/data/")
fp <- read.csv("FatherPresent.csv")
( fp.obs <- xtabs(~hospital+father,data=fp) )
( fp.chi <- chisq.test(fp.obs,correct=FALSE) )
fp.chi$expected
percTable(fp.obs,margin=1,digits=1)
```

Table 21.5. Expected frequency table for father's presence (or absence) during child birth among four hospitals.

```
        father
hospital Absent Present
       A  15.25   59.75
       B  15.25   59.75
       C  15.25   59.75
       D  15.25   59.75
```

Table 21.6. Observed frequency table for father's presence (or absence) during child birth among four hospitals.

```
        father
hospital Absent Present
       A      9      66
       B     15      60
       C     18      57
       D     19      56
```

Table 21.7. Results from the Chi-Square Test for differences in father's presence during child birth among four hospitals.

```
X-squared = 5.0003, df = 3, p-value = 0.1718
```

Table 21.8. Percentage of father's presence (or absence) during child birth among four hospitals.

```
        father
hospital Absent Present   Sum
       A   12.0    88.0 100.0
       B   20.0    80.0 100.0
       C   24.0    76.0 100.0
       D   25.3    74.7 100.0
```

## 21.4 Chi-Square test in R (Summarized Data)

Two-way frequency datat that has already been summarized (outside of R) must be entered into a two-dimensional matrix. The frequencies must first be entered into a vector with the first row of values followed by the second row and so on. This vector is then the first argument to `matrix()`, which will also include the number of rows in the frequency table in `nrow=` and `byrow=TRUE` (which causes the values in the vector to be entered into the matrix in a row-wise manner). The process of entering summarized data into a matrix is better explained by example.

Suppose that you are given this observed frequency table.

| | Species | | | | | | |
|---|---|---|---|---|---|---|---|
| Location | A | B | C | D | E | F | |
| DI | 34 | 22 | 14 | 13 | 12 | 5 | **100** |
| BP | 62 | 12 | 8 | 7 | 6 | 5 | **100** |
| | **96** | **34** | **22** | **20** | **18** | **10** | **200** |

The observed frequencies, ignoring the marginal sums, are first entered into a vector called `freq` below, which is then transformed into a two-row matrix called `obstbl`.

```
> ( freq <- c(34,22,14,13,12,5,62,12,8,7,6,5) )
 [1] 34 22 14 13 12  5 62 12  8  7  6  5
> ( obstbl <- matrix(freq,nrow=2,byrow=TRUE) )
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]   34   22   14   13   12    5
[2,]   62   12    8    7    6    5
```

The matrix is more informative if the rows and columns are named with `rownames()` and `colnames()` as shown below.

```
> rownames(obstbl) <- c("DI","BP")
> colnames(obstbl) <- c("A","B","C","D","E","F")
> obstbl
    A  B  C  D  E F
DI 34 22 14 13 12 5
BP 62 12  8  7  6 5
```

Once this observed table is constructed, the chi-square tests is performed exactly as described in the previous sections (i.e., starting with `chisq.test()`).

### 21.4.1 Example - Apostle Islands Plants

Below are the 11-steps (Section 18.1) for completing a full hypothesis test for the following situation:

> *In her Senior Capstone project a Northland College student recorded the dominant (i.e., most abundant) plant species in 100 randomly selected plots on both Devil's Island and the Bayfield Peninsula (i.e., the mainland). There were a total of six "species" (one group was called "other") recorded (labeled as A, B, C, D, E, and F). The results are shown in the table below. Determine, at the 5% level, if the frequency of dominant species differs between the two locations.*

| | Species | | | | | | |
|---|---|---|---|---|---|---|---|
| Location | A | B | C | D | E | F | |
| DI | 34 | 22 | 14 | 13 | 12 | 5 | **100** |
| BP | 62 | 12 | 8 | 7 | 6 | 5 | **100** |
| | **96** | **34** | **22** | **20** | **18** | **10** | **200** |

1. $\alpha$=0.05.
2. $H_0$: "the distribution of dominant plants species is the same between Devil's Island and the Bayfield Peninsula" vs. $H_A$: "the distribution of dominant plants species is NOT the same between Devil's Island and the Bayfield Peninsula."
3. A Chi-Square Test is required because (i) a categorical variable with six levels (plant species) was recorded and (ii) two grouops are being compared (Devil's Island and Bayfield Peninsula).
4. The data appear to be part of an observational study where the plots were randomly selected.
5. There are more than five individuals in each cell of the expected table (Table 21.9).
6. The statistic is the observed frequency table given in the background.
7. $\chi^2$=16.54 with 5 df (Table 21.10).
8. p-value=0.0055 (Table 21.10).
9. $H_0$ is rejected because the p-value is $< \alpha$.
10. There does appear to be a significant difference in the distribution of the dominant plants between the two sites. A look at the Pearson residuals (Table 21.11) and the row-percentage table (Table 21.12) both suggest that the biggest difference between the two locations is due to "plant A."[5]

**R Appendix:**

```
freq <- c(34,22,14,13,12,5,62,12,8,7,6,5)
ai.obs <- matrix(freq,nrow=2,byrow=TRUE)
rownames(ai.obs) <- c("DI","BP")
colnames(ai.obs) <- c("A","B","C","D","E","F")
( ai.chi <- chisq.test(ai.obs) )
ai.chi$expected
ai.chi$residuals
percTable(ai.obs,margin=1,digits=1)
ai.obs1 <- ai.obs[,-1]
( ai.chi1 <- chisq.test(ai.obs1) )
```

[5] When "Plant A" is removed from the observed table, the Chi-Square Test performed on the remaining plant species showed no difference in the distribution of the remaining plants between the two locations ($p = 0.9239$). Thus, most of the difference in plant distributions between Devil's Island and the Bayfield Peninsula appears to be due primarily to "plant A" with more of "plant A" found on the Bayfield Peninsula than on Devil's Island.

Table 21.9. Expected frequency table for dominant plant species on Devil's Island and the Bayfield Peninsula.

```
    A  B  C  D E F
DI 48 17 11 10 9 5
BP 48 17 11 10 9 5
```

Table 21.10. Results from the Chi-Square Test for differences in the distribution of dominant plant species between Devil's Island and the Bayfield Peninsula.

```
X-squared = 16.5442, df = 5, p-value = 0.00545
```

Table 21.11. Pearson residuals from the Chi-Square Test for differences in the distribution of dominant plant species between Devil's Island and Bayfield Peninsula.

```
           A         B         C          D  E F
DI -2.020726  1.212678  0.904534  0.9486833  1 0
BP  2.020726 -1.212678 -0.904534 -0.9486833 -1 0
```

Table 21.12. Percentage of dominant plant species within each location (Devil's Island and Bayfield Peninsula).

```
    A  B  C  D  E F Sum
DI 34 22 14 13 12 5 100
BP 62 12  8  7  6 5 100
```