# The Chi-Squared Family of Distributions- $\chi^2$

# Learning Objectives

- Define the Chi Square distribution in terms of squared normal deviates
- Describe how the shape of the Chi Square distribution changes as its degrees of freedom increase
- Chi Square test

**Nonparametric Tests** – statistical procedures that can be used when the assumptions of parametrical procedures cannot be met.

Given a choice, use a parametric test; they are more powerful.

We will use chi-square ( $\chi^2$).

Often used for nominal and ordinal data in the form of frequency counts.

# The importance of Chi-Square ( $\chi^2$)

- It is one of the most commonly used non-parametric test, in which the sampling distribution of the test statistics is a chi-square distribution, when the null hypothesis is true.

- It is very important because many test statistics are approximately distributed as Chi Square.

- It can be applied when there are few or no assumptions about the population parameter.

- It can be applied on categorical data or qualitative data using a contingency table.

- Used to evaluate unpaired/unrelated samples and proportions.

Answers such questions as:

- Which of the three leading brands of bottled water do most Americans prefer?
- How does the number of male nurses compare to the number of female nurses in the profession?
- Is there a relationship between gender and the types of movies (e.g., drama, comedies, adventure) watched most?

All of these questions involve determining the frequencies of observations in various categories.

## The Chi-Square Statistic

Measures the amount of discrepancy between observed frequencies and the frequencies that would be expected due to chance, or random sampling error.  The formula for chi square is:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

where:  $f_o$ = observed frequencies
$f_e$ = expected frequencies

- If we find no differences between our observed and expected frequencies, then our obtained $\chi^2$ value will equal 0.

- But if our observed frequencies differ from those that would be expected, then $\chi^2$ will be greater than 0.

- How much greater than 0 our chi-square value has to be to reject $H_0$ will be determined by comparing our obtained chi-square value to a theoretical distribution of chi-square values.

- Chi-square values will range from 0 to positive infinity.

- Because $\chi^2$ is calculated from squared deviations, there will be no negative values.

- Two types of chi-square tests:

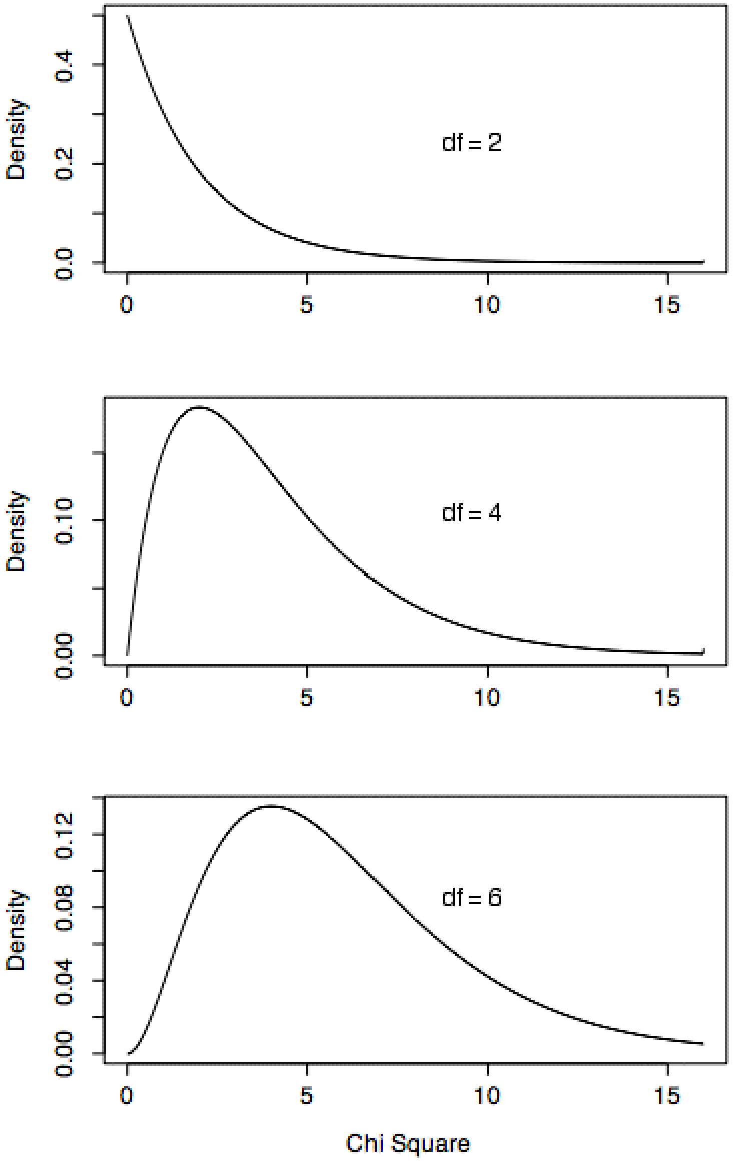
  - goodness of fit test
  - test of independence

# Sum: The feature of $\chi^2$

1. The mean of $\chi^2$ distribution is equal to the number of degrees of freedom (n)
2. The variance is equal to two times the number of degrees of freedom, i.e. the variance od $\chi^2$ distribution is equal to 2n.
3. The median of $\chi^2$ distribution divides, the area of the curve into two equal parts, each part being 0.5
4. The mode of $\chi^2$ distribution is equal to n-2
5. Since $\chi^2$ values always positive, the $\chi^2$ curve is always positively skewed.
6. Sine $\chi^2$ values increases with the increase in the degrees of freedom, there is a new $\chi^2$ distribution with every increase in the number of degrees of freedom.
7. The lowest value of $\chi^2$ is zero and the highest value is infinity ie $\chi^2 \geq 0$
8. As the degrees of freedom increase, $\chi^2$ curve approaches a normal distribution.

The mean of a Chi Square distribution is its degrees of freedom. Chi Square distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom.

As the degrees of freedom increases, the Chi Square distribution approaches a normal distribution.



## Assumptions

- *Independent Observations*. Each participant can contribute frequencies to only one category.

- Minimum Size. The expected frequency ($f_e$) for all categories should be a minimum of 5.

## The One-Way Goodness of Fit Test

- Examines *one* variable, such as religion, in several categories – Buddhism, Christianity, Hinduism, Islam, Judaism, Taoism, Other.

- Used to determine how well the frequencies observed in our sample match the frequencies that would be expected in the population if the null hypothesis were true.

## Hypothesis Testing for Goodness of Fit

- The null hypothesis specifies that a certain proportion of the population will fall into each category and therefore what frequencies should be expected in our sample of *n* subjects if $H_0$ is true.

- The alternative hypothesis, as always, will state the opposite.

**No difference from a known population**. In some cases, a researcher may want to compare the frequencies of a randomly drawn sample to the frequencies of a known population distribution.

For example, a professor collects data to determine if the political affiliations (i.e., Democratic, Republican, Libertarian, Green Party, or Independent) of students who major in political science correspond to the patterns illustrated by voters in the last U.S. presidential election.

In this case, the null and alternative hypotheses would be written as:

$H_0$: The political affiliations of political science students are the
same as U.S. voters.
$H_1$: The political affiliations of political science students are
different from U.S. voters.

If our obtained $\chi^2$ value is greater than would be expected by chance if the null hypothesis is true then, as always, we reject $H_0$ and assume instead that the political preferences of political science students are different from those of U.S. voters in general.

**No preference**. At other times, the null hypothesis might predict equal frequencies in each category, or no preference.

For example, a candy manufacturer may want to know if there is a preference for one of the five candy bars produced by his company and distributed by a particular store outlet. The null and alternative hypotheses would be written as follows:

$H_0$: Preferences for candy bars are equally divided.
$H_1$: Preferences for candy bars are not equally divided.

If our obtained $\chi^2$ value is greater than would be expected by chance if the null hypothesis were true, then we would reject $H_0$ and assume instead that the preferences for candy bars are not equally divided.

**No preference**. At other times, the null hypothesis might predict equal frequencies in each category, or no preference.

For example, a candy manufacturer may want to know if there is a preference for one of the five candy bars produced by his company and distributed by a particular store outlet. The null and alternative hypotheses would be written as follows:

$H_0$: Preferences for candy bars are equally divided.
$H_1$: Preferences for candy bars are not equally divided.

If our obtained $\chi^2$ value is greater than would be expected by chance if the null hypothesis were true, then we would reject $H_0$ and assume instead that the preferences for candy bars are not equally divided.

## For Example,

## Goodness of Fit for Known Proportions

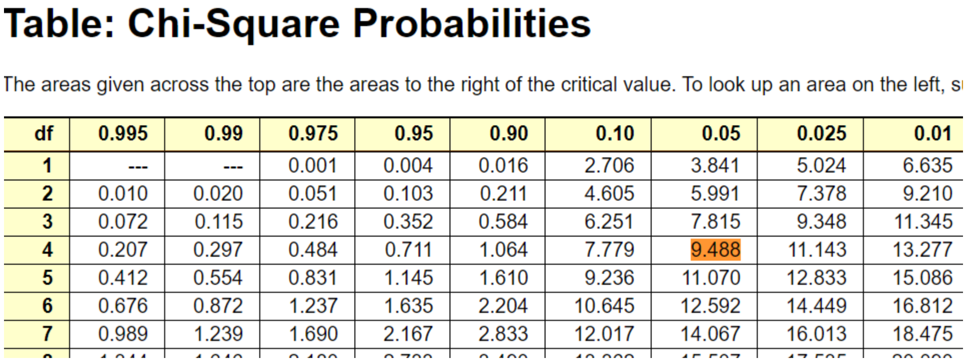
Research Problem. A new professor at a mid-sized college wanted to see if her grade distribution, after her first year of teaching, was comparable to the overall college grade distribution which has the following percentages: A – 10%; B – 22%; C – 40%; D – 21%; and F – 7%. The distribution of the new professor's grades for 323 students at the end of her first year was as follows: 38 students received A's, 78 received B's, 139 received C's, 55 received D's, and 13 received F's. Does the new professor's grade distribution fit the overall college's distribution? Test, using $\alpha = .05$.

Step 1: Formulate Hypotheses:

We are using the proportions/percentages from a **known population** distribution for our comparison. Thus, our hypotheses would be as follows:

$H_0$: The distribution of grades for the new professor fits the overall grade distribution of the college.

$H_1$: The distribution of grades for the new professor does not fit the overall grade distribution of the college.



## Table: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, s

| df | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | --- | --- | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 |

## Step 2:  Indicate the Alpha Level and Determine Critical Values:

After we calculate our $\chi^2$ value in the next step, we will need to compare it to a critical $\chi^2$ value.

For **goodness of fit**, $df = k - 1$, where $k$ refers to the number of categories.

Our problem has 5 categories (A, B, C, D, and F).  Thus,

$\alpha = .05$
$df = k - 1 = 5 - 1 = 4$
$\chi^2_{\text{crit}} = 9.488$

https://people.richland.edu/james/lecture/m170/tbl-chi.html

Step 3: Calculate Relevant Statistics:

- To use the chi-square formula, observed and expected frequencies are required.

  - Observed frequencies are the actual frequencies obtained from our sample which are given in the problem.
  - For research questions dealing with a known population distribution, expected frequency is determined by multiplying the known proportion in the population by *n*. In other words,

$$f_e = (\text{known proportion})(n)$$

The null hypothesis specifies that the new professor’s grades won’t differ significantly from the proportion found in the population.

We know that 10% of the college population earned A’s, 22% earned B’s, and so forth.

If we multiply those known proportions by our sample size, then we can determine what frequencies would be expected if $H_0$ were true.  Thus,

$$\begin{aligned}
A &= .10 \times 323 = 32.30 \\
B &= .22 \times 323 = 71.06 \\
C &= .40 \times 323 = 129.20 \\
D &= .21 \times 323 = 67.83 \\
F &= .07 \times 323 = 22.61
\end{aligned}$$

Use a table to keep track of observed and expected frequencies.  In the table below, expected frequencies are in parentheses.

| A | B | C | D | F |
|---|---|---|---|---|
| 38 (32.30) | 78 (71.06) | 139 (129.20) | 55 (67.83) | 13 (22.61) |

Both observed frequency counts and expected frequency values, when added, should equal *n*.  If they do not, then you have made a mistake in your calculations.

Now we can plug our values into the formula for $\chi^2$. The summation in the formula ($\Sigma$) tells us that the information following that symbol needs to be added for each category.

$$
\begin{aligned}
\chi^2 &= \Sigma \frac{(f_o - f_e)^2}{f_e} \\
&= \frac{(38-32.3)^2}{32.3} + \frac{(78-71.06)^2}{71.06} + \frac{(139-129.2)^2}{129.2} + \frac{(55-67.83)^2}{67.83} + \frac{(13-22.61)^2}{22.61} \\
&= 1.01 + .68 + .74 + 2.43 + 4.08 \\
&= 8.94
\end{aligned}
$$

Step 4:  Make a Decision and Report the Results:

The new professor’s grades did not differ significantly from those of the college at large.  Fail to reject $H_0$, $\chi^2$ (4 *df*, *n* = 323) = 8.94, *p* > .05.

The only new element is the inclusion of the sample size.

# Example Two

## Goodness of Fit for No Preference

Research Question. A manufacturer of women's clothing wants to know which of the colors red, blue, green, brown, and black, would be preferred for the fall collection. A random sample of 92 women shoppers was asked their preferences. The results were as follows:

| Red | Blue | Green | Brown | Black |
|---|---|---|---|---|
| 5 | 19 | 19 | 27 | 22 |

Test a no preference null hypothesis using $\alpha = .01$.

Step 1:

$H_0$: All colors were equally preferred
$H_1$: The colors were not equally preferred

Step 2:

$\alpha = .01$
$df = k - 1 = 5 - 1 = 4$
$\chi^2_{\text{crit}} = 13.277$

Step 3 :

For a $H_0$ that specifies no preference, calculate the expected frequency ($f_e$) by dividing $n$ by the number of categories ($k$). In other words,

$f_e = n \div k$

For our problem,
$f_e = 92 \div 5 = 18.4$ for each category

| Red | Blue | Green | Brown | Black |
|---|---|---|---|---|
| 5 (18.4) | 19 (18.4) | 19 (18.4) | 27 (18.4) | 22 (18.4) |

$$
\begin{aligned}
\chi^2 &= \Sigma \frac{(f_o - f_e)^2}{f_e} \\
&= \frac{(5-18.4)^2}{18.4} + \frac{(19-18.4)^2}{18.4} + \frac{(19-18.4)^2}{18.4} + \frac{(27-18.4)^2}{18.4} + \frac{(22-18.4)^2}{18.4} \\
&= 9.76 + .02 + .02 + 4.02 + .70 \\
&= 14.52
\end{aligned}
$$

Step 4:

Preferences for colors were not equally divided. Reject $H_0$, $\chi^2$ (4 *df*, *n* = 92) = 14.52, $p < .01$.

## Test of Independence

- Testing for independence of variables involves examining the frequencies for *two* variables to determine whether one variable is independent of the other or if the variables are related.

- In F-test, we used the Pearson correlation to test similar hypotheses; however, the Pearson requires interval or ratio level data.

- Now we will test for such relationships using nominal or ordinal level data.

## Test of Independence

- Testing for independence of variables involves examining the frequencies for *two* variables to determine whether one variable is independent of the other or if the variables are related.

- In F-test, we used the Pearson correlation to test similar hypotheses; however, the Pearson requires interval or ratio level data.

- Now we will test for such relationships using nominal or ordinal level data.

## Test of Independence

- Testing for independence of variables involves examining the frequencies for *two* variables to determine whether one variable is independent of the other or if the variables are related.

- In F-test, we used the Pearson correlation to test similar hypotheses; however, the Pearson requires interval or ratio level data.

- Now we will test for such relationships using nominal or ordinal level data.

Hypothesis Testing for the Test of Independence.

Suppose a researcher wants to know if there is a relationship between gender and introversion.

$H_0$: There is no relationship between gender and introversion.

$H_1$: A relationship exists between gender and introversion.

## Contingency Table

- The table for the chi-square test of independence is slightly different from the table for one-way goodness of fit because we are now working with *two* variables.

- Called a contingency table, it will include rows for the values of one variable, and columns which will contain the values for the second variable.

- Similar to the goodness of fit table, the expected frequencies, after being calculated, will be placed in parenthesis in the same cells as the corresponding observed frequencies.

# Example Three

## Test of Independence

Research Problem. A professor at the veterinary school of medicine is curious about whether or not a relationship exists between gender and type of pets owned in childhood. She randomly asks a sample of $n$ = 260 students (132 female and 128 male) about their pet ownership, the frequency of which is recorded in the table below. Is there a relationship between gender and type of pet ownership? Test at $\alpha$ = .05.

**Type of Pet**

| | | **Dogs** | **Cats** | **Birds** | **Reptiles** | **Rodents** | Row Total |
|---|---|---|---|---|---|---|---|
| **Gender** | **Female** | 58 | 36 | 22 | 4 | 12 | 132 |
| | **Male** | 62 | 22 | 14 | 10 | 20 | 128 |
| | Column Total | 120 | 58 | 36 | 14 | 32 | *n* = 260 |

Notice that row totals and column totals have also been computed. These are important for determining expected frequencies.

For the **test of independence**, expected frequency is determined by the following formula:

$$f_e = \frac{(f_c)(f_r)}{n}$$

Where $f_c$ = column total
$f_r$ = row total
$n$ = sample size

| | **Dogs** | **Cats** | **Birds** | **Reptiles** | **Rodents** | Row Total |
|---|---|---|---|---|---|---|
| **Female** | 58 | 36 | 22 | 4 | 12 | 132 |
| **Male** | 62 | 22 | 14 | 10 | 20 | 128 |
| Column Total | 120 | 58 | 36 | 14 | 32 | *n* = 260 |

Using this formula, calculate the expected frequencies for each cell. For example, the expected frequency for **females** who owned **dogs** in childhood would be:

$$f_e = \frac{(f_c)(f_r)}{n} = \frac{(120)(132)}{260} = 60.92$$

Using the same procedure, calculate the expected frequencies for all observed frequencies and place them in parentheses in the appropriate cells of the contingency table.

**Type of Pet**

| | | **Dogs** | **Cats** | **Birds** | **Reptiles** | **Rodents** | Row Total |
|---|---|---|---|---|---|---|---|
| **Gender** | **Female** | 58 (60.92) | 36 (29.45) | 22 (18.28) | 4 (7.11) | 12 (16.25) | 132 |
| | **Male** | 62 (59.08) | 22 (28.55) | 14 (17.72) | 10 (6.89) | 20 (15.75) | 128 |
| | Column Total | 120 | 58 | 36 | 14 | 32 | *n* = 260 |

We can now test our hypotheses in the usual manner.

Step 1:

$H_0$: Gender and pet ownership are independent and unrelated.
$H_1$: There is a relationship between gender and pet ownership.

Step 2:

For **tests of independence**, *df* = (R – 1)(C – 1)

where R = number of rows in contingency table
C = number of columns in contingency table

For our example,
$\alpha = .05$
$df = (R - 1)(C - 1) = (2 - 1)(5 - 1) = 4$
$\chi^2_{crit} = 9.488$

Step 3:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(58-60.92)^2}{60.92} + \frac{(36-29.45)^2}{29.45} + \frac{(22-18.28)^2}{18.28} + \frac{(4-7.11)^2}{7.11} + \frac{(12-16.25)^2}{16.25}$$

$$+ \frac{(62-59.08)^2}{59.08} + \frac{(22-28.55)^2}{28.55} + \frac{(14-17.72)^2}{17.72} + \frac{(10-6.89)^2}{6.89} + \frac{(20-15.75)^2}{15.75}$$

$$= .14 + .1.46 + .76 + 1.36 + 1.11 + .14 + 1.50 + .78 + 1.40 + 1.15$$

$$= 9.80$$

Step 4:

There is a relationship between gender and type of pet owned. Reject $H_0$, $\chi^2$ (4 *df*, *n* = 260) = 9.80, *p* < .05.