

# 17. Chi Square

- A. Chi Square Distribution
- B. One-Way Tables
- C. Contingency Tables
- D. Exercises

Chi Square is a distribution that has proven to be particularly useful in statistics. The first section describes the basics of this distribution. The following two sections cover the most common statistical tests that make use of the Chi Square distribution. The section “One-Way Tables” shows how to use the Chi Square distribution to test the difference between theoretically expected and observed frequencies. The section “Contingency Tables” shows how to use Chi Square to test the association between two nominal variables. This use of Chi Square is so common that it is often referred to as the “Chi Square Test.”

# Chi Square Distribution

by David M. Lane

## *Prerequisites*

- Chapter 1: Distributions
- Chapter 7: Standard Normal Distribution
- Chapter 10: Degrees of Freedom

## *Learning Objectives*

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

A *standard normal deviate* is a random sample from the standard normal distribution. The Chi Square distribution is the distribution of the sum of squared standard normal deviates. The *degrees of freedom* of the distribution is equal to the number of standard normal deviates being summed. Therefore, Chi Square with one degree of freedom, written as  $\chi^2(1)$ , is simply the distribution of a single normal deviate squared. The area of a Chi Square distribution below 4 is the same as the area of a standard normal distribution below 2, since 4 is  $2^2$ .

Consider the following problem: you sample two scores from a standard normal distribution, square each score, and sum the squares. What is the probability that the sum of these two squares will be six or higher? Since two scores are sampled, the answer can be found using the Chi Square distribution with two degrees of freedom. A Chi Square calculator can be used to find that the probability of a Chi Square (with 2 df) being six or higher is 0.050.

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. Figure 1 shows density functions for three Chi Square distributions. Notice how the skew decreases as the degrees of freedom increase.

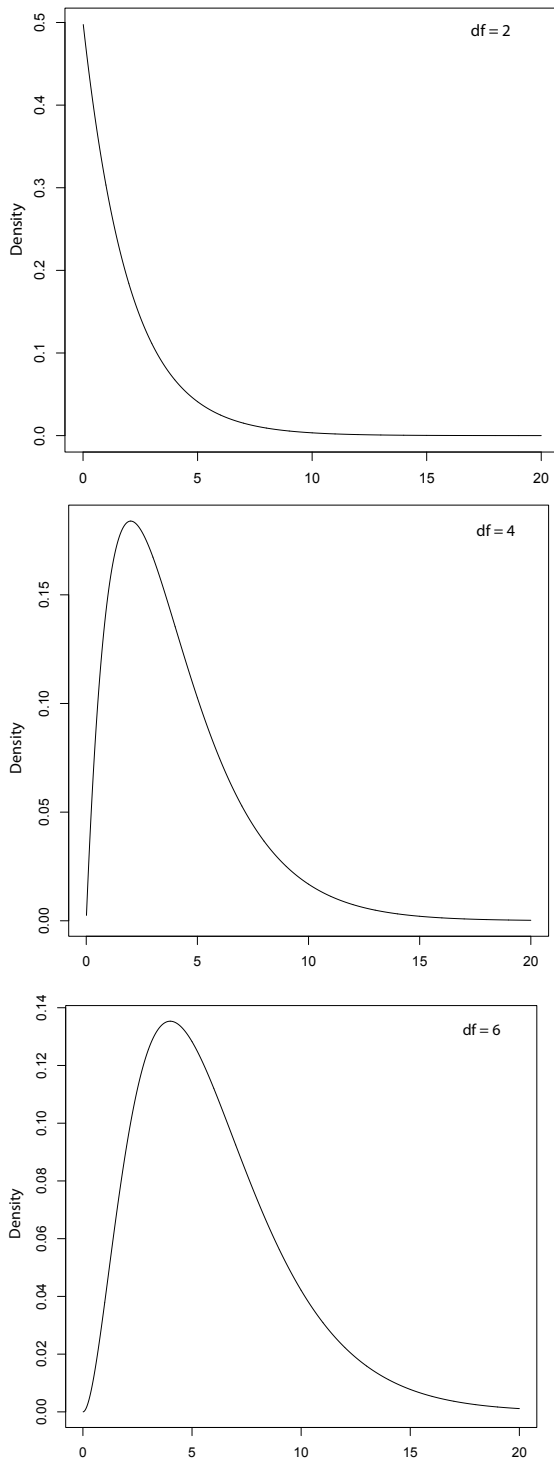


Figure 1. Chi Square distributions with 2, 4, and 6 degrees of freedom.

The Chi Square distribution is very important because many test statistics are approximately distributed as Chi Square. Two of the more common tests using the

Chi Square distribution are tests of deviations of differences between theoretically expected and observed frequencies (one-way tables) and the relationship between categorical variables (contingency tables). Numerous other tests beyond the scope of this work are based on the Chi Square distribution.

# One-Way Tables (Testing Goodness of Fit)

by David M. Lane

## *Prerequisites*

- Chapter 5: Basic Concepts of Probability
- Chapter 11: Significance Testing
- Chapter 17: Chi Square Distribution

## *Learning Objectives*

1. Describe what it means for there to be theoretically-expected frequencies
2. Compute expected frequencies
3. Compute Chi Square
4. Determine the degrees of freedom

The Chi Square distribution can be used to test whether observed data differ significantly from theoretical expectations. For example, for a fair six-sided die, the probability of any given outcome on a single roll would be  $1/6$ . The data in Table 1 were obtained by rolling a six-sided die 36 times. However, as can be seen in Table 1, some outcomes occurred more frequently than others. For example, a “3” came up nine times, whereas a “4” came up only two times. Are these data consistent with the hypothesis that the die is a fair die? Naturally, we do not expect the sample frequencies of the six possible outcomes to be the same since chance differences will occur. So, the finding that the frequencies differ does not mean that the die is not fair. One way to test whether the die is fair is to conduct a significance test. The null hypothesis is that the die is fair. This hypothesis is tested by computing the probability of obtaining frequencies as discrepant or more discrepant from a uniform distribution of frequencies as obtained in the sample. If this probability is sufficiently low, then the null hypothesis that the die is fair can be rejected.

Table 1. Outcome Frequencies from a Six-Sided Die

Outcome	Frequency
1	8
2	5
3	9

4	2
5	7
6	5

The first step in conducting the significance test is to compute the expected frequency for each outcome given that the null hypothesis is true. For example, the expected frequency of a “1” is 6, since the probability of a “1” coming up is 1/6 and there were a total of 36 rolls of the die.

$$\text{Expected frequency} = (1/6) (36) = 6$$

Note that the expected frequencies are expected only in a theoretical sense. We do not really “expect” the observed frequencies to match the “expected frequencies” exactly.

The calculation continues as follows. Letting  $E$  be the expected frequency of an outcome and  $O$  be the observed frequency of that outcome, compute

$$\frac{(E - O)^2}{E}$$

for each outcome. Table 2 shows these calculations.

Table 2. Outcome Frequencies from a Six-Sided Die

Outcome	E	O	$(E-O)^2/E$
1	6	8	0.667
2	6	5	0.167
3	6	9	1.500
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

Next we add up all the values in Column 4 of Table 2.

$$\sum \frac{(E - O)^2}{E} = 5.33$$

This sampling distribution of

$$\sum \frac{(E - O)^2}{E}$$

is approximately distributed as Chi Square with k-1 degrees of freedom, where k is the number of categories. Therefore, for this problem the test statistic is

$$\chi_5^2 = 5.333$$

which means the value of Chi Square with 5 degrees of freedom is 5.333.

From a Chi Square calculator it can be determined that the probability of a Chi Square of 5.333 or larger is 0.377. Therefore, the null hypothesis that the die is fair cannot be rejected.

The Chi Square test can also be used to test other deviations between expected and observed frequencies. The following example shows a test of whether the variable “University GPA” in the SAT and College GPA case study is normally distributed.

The first column in Table 3 shows the normal distribution divided into five ranges. The second column shows the proportions of a normal distribution falling in the ranges specified in the first column. The expected frequencies (E) are calculated by multiplying the number of scores (105) by the proportion. The final column shows the observed number of scores in each range. It is clear that the observed frequencies vary greatly from the expected frequencies. Note that if the distribution were normal, then there would have been only about 35 scores between 0 and 1, whereas 60 were observed.

Table 3. Expected and Observed Scores for 105 University GPA Scores.

Range	Proportion	E	O
Above 1	0.159	16.695	9
0 to 1	0.341	35.805	60
-1 to 0	0.341	35.805	17
Below -1	0.159	16.695	19

The test of whether the observed scores deviate significantly from the expected scores is computed using the familiar calculation.

$$\chi^2_3 = \sum \frac{(E - O)^2}{E} = 30.09$$

The subscript “3” means there are three degrees of freedom. As before, the degrees of freedom is the number of outcomes minus one, which is three in this example. The Chi Square distribution calculator shows that  $p < 0.001$  for this Chi Square. Therefore, the null hypothesis that the scores are normally distributed can be rejected.



# Contingency Tables

by David M. Lane

## Prerequisites

- Chapter 17: Chi Square Distribution
- Chapter 17: One-Way Tables

## Learning Objectives

1. State the null hypothesis tested concerning contingency tables
2. Compute expected cell frequencies
3. Compute Chi Square and df

This section shows how to use Chi Square to test the relationship between nominal variables for significance. For example, Table 1 shows the data from the Mediterranean Diet and Health case study.

Table 1. Frequencies for Diet and Health Study

Diet	Outcome				Total
	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

The question is whether there is a significant relationship between diet and outcome. The first step is to compute the expected frequency for each cell based on the assumption that there is no relationship. These expected frequencies are computed from the totals as follows. We begin by computing the expected frequency for the AHA Diet/Cancers combination. Note that 22/605 subjects developed cancer. The proportion who developed cancer is therefore 0.0364. If there were no relationship between diet and outcome, then we would expect 0.0364 of those on the AHA diet to develop cancer. Since 303 subjects were on the AHA diet, we would expect  $(0.0364)(303) = 11.02$  cancers on the AHA diet. Similarly, we would expect  $(0.0364)(302) = 10.98$  cancers on the Mediterranean diet. In

general, the expected frequency for a cell in the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column is equal to

$$E_{i,j} = \frac{T_i T_j}{T}$$

where  $E_{i,j}$  is the expected frequency for cell  $i,j$ ,  $T_i$  is the total for the  $i^{\text{th}}$  row,  $T_j$  is the total for the  $j^{\text{th}}$  column, and  $T$  is the total number of observations. For the AHA Diet/Cancers cell,  $i = 1$ ,  $j = 1$ ,  $T_i = 303$ ,  $T_j = 22$ , and  $T = 605$ . Table 2 shows the expected frequencies (in parenthesis) for each cell in the experiment. Table 2 shows the expected frequencies (in parenthesis) for each cell in the experiment.

Table 2. Observed and Expected Frequencies for Diet and Health Study

	Outcome				
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)	302
Total	22	38	33	512	605

The significance test is conducted by computing Chi Square as follows.

$$\chi^2_3 = \sum \frac{(E - O)^2}{E} = 16.55$$

The degrees of freedom is equal to  $(r-1)(c-1)$ , where  $r$  is the number of rows and  $c$  is the number of columns. For this example, the degrees of freedom is  $(2-1)(4-1) = 3$ . The Chi Square calculator can be used to determine that the probability value for a Chi Square of 16.55 with three degrees of freedom is equal to 0.0009. Therefore, the null hypothesis of no relationship between diet and outcome can be rejected.

A key assumption of this Chi Square test is that each subject contributes data to only one cell. Therefore, the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment. Consider an experiment in

which each of 16 subjects attempted two anagram problems. The data are shown in Table 3.

Table 3. Anagram problem data

	Anagram 1	Anagram 2
Solved	10	4
Did not Solve	6	12

It would not be valid to use the Chi Square test on these data since each subject contributed data to two cells: one cell based on their performance on Anagram 1 and one cell based on their performance on Anagram 2. The total of the cell frequencies in the table is 32, but the total number of subjects is only 16.

The formula for Chi Square yields a statistic that is only approximately a Chi Square distribution. In order for the approximation to be adequate, the total number of subjects should be at least 20. Some authors claim that the correction for continuity should be used whenever an expected cell frequency is below 5. Research in statistics has shown that this practice is not advisable. For example, see: Bradley, Bradley, McGrath, & Cutcomb (1979). The correction for continuity when applied to 2 x 2 contingency tables is called the Yates correction.

# Statistical Literacy

by David M. Lane

## *Prerequisites*

- Chapter 17: Contingency Tables

An experiment was conducted to test whether the spice saffron can inhibit liver cancer. Two groups of rats were tested. Both groups were injected with chemicals known to increase the chance of liver cancer. The experimental group was fed saffron ( $n = 24$ ) whereas the control group was not ( $n = 8$ ). The experiment is [described here](#).

Only 4 of the 24 subjects in the saffron group developed cancer as compared to 6 of the 8 subjects in the control group.

## **What do you think?**

What method could be used to test whether this difference between the experimental and control groups is statistically significant?

The Chi Square test of contingency tables could be used. It yields a Chi Squared ( $df = 1$ ) of 9.50 which has an associated  $p$  of 0.002. □

## References

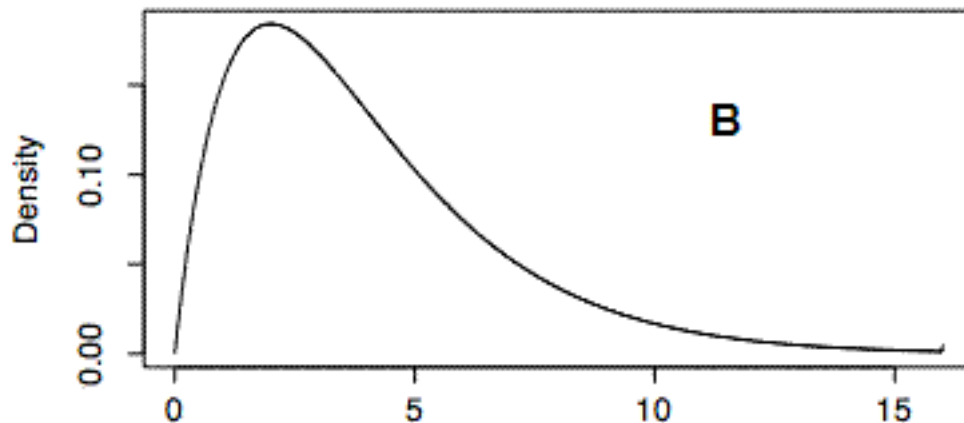
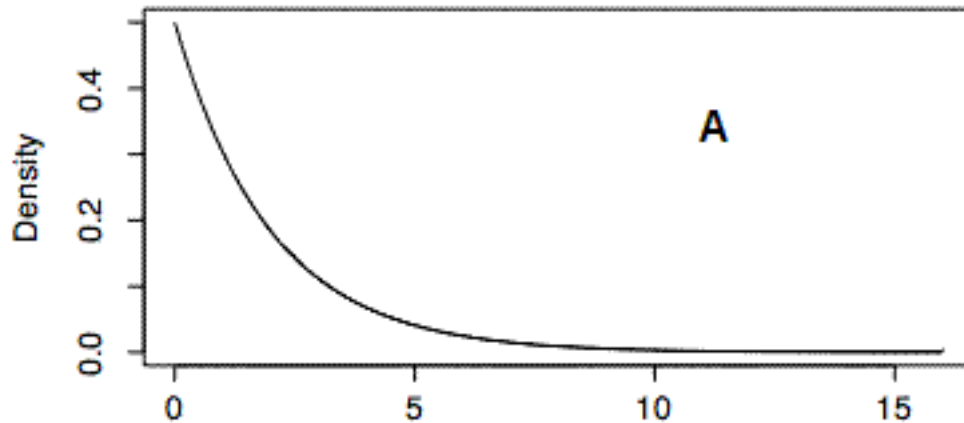
Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979) Type I error rate of the chi square test of independence in  $r \times c$  tables that have small expected frequencies. *Psychological Bulletin*, 86, 1200-1297.

## Exercises

### Prerequisites

- All material presented in the Chi Square Chapter

1. Which of the two Chi Square distributions shown below (A or B) has the larger degrees of freedom? How do you know?



2. Twelve subjects were each given two flavors of ice cream to taste and then were asked whether they liked them. Two of the subjects liked the first flavor and nine of them liked the second flavor. Is it valid to use the Chi Square test to determine whether this difference in proportions is significant? Why or why not?
3. A die is suspected of being biased. It is rolled 25 times with the following result:

Outcome	Frequency
1	9
2	4
3	1
4	8
5	3
6	0

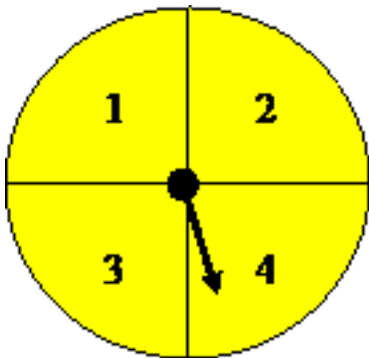
Conduct a significance test to see if the die is biased. (a) What Chi Square value do you get and how many degrees of freedom does it have? (b) What is the p value?

4. A recent experiment investigated the relationship between smoking and urinary incontinence. Of the 322 subjects in the study who were incontinent, 113 were smokers, 51 were former smokers, and 158 had never smoked. Of the 284 control subjects who were not incontinent, 68 were smokers, 23 were former smokers, and 193 had never smoked.
  - a. Create a table displaying this data.
  - b. What is the expected frequency in each cell?
  - c. Conduct a significance test to see if there is a relationship between smoking and incontinence. What Chi Square value do you get? What p value do you get?
  - d. What do you conclude?
  
5. At a school pep rally, a group of sophomore students organized a free raffle for prizes. They claim that they put the names of all of the students in the school in the basket and that they randomly drew 36 names out of this basket. Of the prize winners, 6 were freshmen, 14 were sophomores, 9 were juniors, and 7 were seniors. The results do not seem that random to you. You think it is a little fishy that sophomores organized the raffle and also won the most prizes. Your school is composed of 30% freshmen, 25% sophomores, 25% juniors, and 20% seniors.
  - a. What are the expected frequencies of winners from each class?
  - b. Conduct a significance test to determine whether the winners of the prizes were distributed throughout the classes as would be expected based on the percentage of students in each group. Report your Chi Square and p values.
  - c. What do you conclude?

6. Some parents of the West Bay little leaguers think that they are noticing a pattern. There seems to be a relationship between the number on the kids' jerseys and their position. These parents decide to record what they see. The hypothetical data appear below. Conduct a Chi Square test to determine if the parents' suspicion that there is a relationship between jersey number and position is right. Report your Chi Square and p values.

	<b>Infield</b>	<b>Outfield</b>	<b>Pitcher</b>	<b>Total</b>
<b>0-9</b>	12	5	5	22
<b>10-19</b>	5	10	2	17
<b>20+</b>	4	4	7	15
<b>Total</b>	21	19	14	54

7. True/false: A Chi Square distribution with 2 df has a larger mean than a Chi Square distribution with 12 df.
8. True/false: A Chi Square test is often used to determine if there is a significant relationship between two continuous variables.
9. True/false: Imagine that you want to determine if the spinner shown below is biased. You spin it 50 times and write down how many times the arrow lands in each section. You will reject the null hypothesis at the .05 level and determine that this spinner is biased if you calculate a Chi Square value of 7.82 or higher.



*Questions from Case Studies*

SAT and GPA (SG) case study



10. (SG) Answer these items to determine if the math SAT scores are normally distributed. You may want to first standardize the scores.
- If these data were normally distributed, how many scores would you expect there to be in each of these brackets: (i) smaller than 1 SD below the mean, (ii) in between the mean and 1 SD below the mean, (iii) in between the mean and 1 SD above the mean, (iv) greater than 1 SD above the mean?
  - How many scores are actually in each of these brackets?
  - Conduct a Chi Square test to determine if the math SAT scores are normally distributed based on these expected and observed frequencies.

#### Diet and Health (DH) case study

11. (DH) Conduct a Pearson Chi Square test to determine if there is any relationship between diet and outcome. Report the Chi Square and p values and state your conclusions.

The following questions are from ARTIST (reproduced with permission)



12. A study compared members of a medical clinic who filed complaints with a random sample of members who did not complain. The study divided the complainers into two subgroups: those who filed complaints about medical treatment and those who filed nonmedical complaints. Here are the data on the total number in each group and the number who voluntarily left the medical clinic. Set up a two-way table. Analyze these data to see if there is a relationship between complaint (no, yes - medical, yes - nonmedical) and leaving the clinic (yes or no).

	<b>No Complaint</b>	<b>Medical Complaint</b>	<b>Nonmedical Complaint</b>
<b>Total</b>	743	199	440
<b>Left</b>	22	26	28

13. Imagine that you believe there is a relationship between a person's eye color and where he or she prefers to sit in a large lecture hall. You decide to collect data from a random sample of individuals and conduct a chi-square test of independence. What would your two-way table look like? Use the information to construct such a table, and be sure to label the different levels of each category.

14. A geologist collects hand-specimen sized pieces of limestone from a particular area. A qualitative assessment of both texture and color is made with the following results. Is there evidence of association between color and texture for these limestones? Explain your answer.

	<b>Colour</b>		
<b>Texture</b>	<b>Light</b>	<b>Medium</b>	<b>Dark</b>
<b>Fine</b>	4	20	8
<b>Medium</b>	5	23	12
<b>Coarse</b>	21	23	4

15. Suppose that college students are asked to identify their preferences in political affiliation (Democrat, Republican, or Independent) and in ice cream (chocolate, vanilla, or straw- berry). Suppose that their responses are represented in the following two-way table (with some of the totals left for you to calculate).

	<b>Chocolate</b>	<b>Vanilla</b>	<b>Strawberry</b>	<b>Total</b>
<b>Democrat</b>	26	43	13	82
<b>Republican</b>	45	12	8	65
<b>Independent</b>	9	13	4	26
<b>Total</b>	80	68	25	173

- What proportion of the respondents prefer chocolate ice cream?
- What proportion of the respondents are Independents?
- What proportion of Independents prefer chocolate ice cream?
- What proportion of those who prefer chocolate ice cream are Independents?

- e. Analyze the data to determine if there is a relationship between political party preference and ice cream preference.
16. NCAA collected data on graduation rates of athletes in Division I in the mid-1980s. Among 2,332 men, 1,343 had not graduated from college, and among 959 women, 441 had not graduated.
- a. Set up a two-way table to examine the relationship between gender and graduation.
- b. Identify a test procedure that would be appropriate for analyzing the relationship between gender and graduation. Carry out the procedure and state your conclusion