Chapter 24

Two-Way Tables and the Chi-Square Test

Lecture Slides

Case Study: Two-Way Tables and the Chi-Square Test* 1

Purdue is a Big Ten university that emphasizes engineering, scientific, and technical fields.

In the 2013–2014 academic year, Purdue had 1670 professors of all ranks (assistant, associate, and full), of whom 483 were women. That's just under 29%, or slightly less than 3 out of every 10 professors.

Getty Images/Fuse

Case Study: Two-Way Tables and the Chi-Square Test* 2

These numbers don’t tell us much about the fields of expertise of women on the faculty.

We need to look at relationships among several variables, not just at sex alone.

Female faculty members are more common in the humanities than in engineering.

We can also look at the relationship between sex and a variable particularly important to faculty members, academic rank.

Case Study: Two-Way Tables and the Chi-Square Test* 3

Universities tend to be run by full professors.

Here is a two-way table that categorizes Purdue's 1670 faculty members by both sex and academic rank (professors typically start at the rank of assistant, then associate, then full professor):

	Female	Male	Total
Assistant professors	160	177	337
Associate professors	191	374	565
Professors	132	636	768
Total	483	1187	1670

Case Study: Two-Way Tables and the Chi-Square Test* 4

	Female	Male	Total
Assistant professors	160	177	337
Associate professors	191	374	565
Professors	132	636	768
Total	483	1187	1670

What does this table tell us about the rank of women on the faculty?

In this chapter, we will learn how to interpret such tables. By the end of the chapter, you will be able to interpret this table.

Two-Way Tables 1

One measurement of student success is the percent completing a bachelor's degree within 6 years.

The following **two-way table** is representative of the percents of students completing a bachelor's degree within 6 years by race/ethnicity:

Race/ ethnicity	White	Black	Hispanic	Asian	American Indian/ Alaska Native	Two or more races	Total
Graduated	6963	1063	1388	792	69	175	10,450
Did not graduate	3933	1614	1163	295	110	119	7234
Total	10,896	2667	2551	1087	179	294	17,684

Two-Way Tables 2

Graduation status (completing a bachelor's degree within 6 years or not) and race of students are both *categorical variables*.

Use a two-way table to display the relationship between two categorical variables.

Race/ ethnicity	White	Black	Hispanic	Asian	American Indian/ Alaska Native	Two or more races	Total
Graduated	6963	1063	1388	792	69	175	10,450
Did not graduate	3933	1614	1163	295	110	119	7234
Total	10,896	2667	2551	1087	179	294	17,684

Two-Way Tables 3

Graduation status is the **row variable** because each row in the table describes one of the possible admission decisions for a student.

Race is the **column variable** because each column describes one of the racial/ethnic groups.

Race/ ethnicity	White	Black	Hispanic	Asian	American Indian/ Alaska Native	Two or more races	Total
Graduated	6963	1063	1388	792	69	175	10,450
Did not graduate	3933	1614	1163	295	110	119	7234
Total	10,896	2667	2551	1087	179	294	17,684

Two-Way Tables 4

Race/ ethnicity	White	Black	Hispanic	Asian	American Indian/ Alaska Native	Two or more races	Total
Graduated	6963	1063	1388	792	69	175	10,450
Did not graduate	3933	1614	1163	295	110	119	7234
Total	10,896	2667	2551	1087	179	294	17,684

The “Total” column at the right of the table contains the totals for graduation status for all students (for all racial/ethnic groups combined).

The “Total” row at the bottom of the table gives the distribution of race for students (with both categories of graduation status combined).

Two-Way Tables 5

It is often clearer to present these distributions using percentages. We might report the distribution of race as:

$$\text{percentage white} = \frac{10{,}896}{17{,}684} \times 100\% = 0.616 \times 100\% = 61.6\%$$

$$\text{percentage black} = \frac{2677}{17{,}684} \times 100\% = 0.151 \times 100\% = 15.1\%$$

$$\text{percentage Hispanic} = \frac{2551}{17{,}684} \times 100\% = 0.144 \times 100\% = 14.4\%$$

$$\text{percentage Asian} = \frac{1087}{17{,}684} \times 100\% = 0.061 \times 100\% = 6.1\%$$

$$\text{percentage Amer. Ind./Alaska Native} = \frac{179}{17{,}684} \times 100\% = 0.017 \times 100\% = 1.0\%$$

$$\text{percentage Two or more races} = \frac{294}{17{,}684} \times 100\% = 0.010 \times 100\% = 1.7\%$$

Two-Way Tables 6

The nature of the relationship between graduation status and race cannot be deduced from the separate distributions but requires the full table.

To describe relationships among categorical variables, calculate appropriate percents from the counts given.

Two-Way Tables 7

Race/ ethnicity	White	Black	Hispanic	Asian	American Indian/ Alaska Native	Two or more races	Total
Graduated	6963	1063	1388	792	69	175	10,450
Did not graduate	3933	1614	1163	295	110	119	7234
Total	10,896	2667	2551	1087	179	294	17,684

Because there are only two categories of graduation status, we can see the relationship between race and graduation status by comparing the percents of those who completed a bachelor’s degree within 6 years for each race:

Two-Way Tables 8

There are conditional percentages

P(G|W)

$$\text{percentage of white students who graduated} = \frac{6963}{10{,}896} \times 100\% = 0.639 \times 100\% = 63.9\%$$

P(G|B)

$$\text{percentage of black students who graduated} = \frac{1063}{2677} \times 100\% = 0.397 \times 100\% = 39.7\%$$

$$\text{percentage of Hispanic students who graduated} = \frac{1388}{2551} \times 100\% = 0.544 \times 100\% = 54.4\%$$

P(G|A)

$$\text{percentage of Asian students who graduated} = \frac{792}{1087} \times 100\% = 0.729 \times 100\% = 72.9\%$$

$$\text{percentage of Amer. Ind./Alaska Native} = \frac{69}{179} \times 100\% = 0.385 \times 100\% = 38.5\%$$

$$\text{percentage of Two or more races} = \frac{175}{294} \times 100\% = 0.595 \times 100\% = 59.5\%$$

Over 60% of the white students and more than 70% of the Asian students completed a bachelor's degree within 6 years, but less than 40% of black students and American Indian/Alaska Native students completed a bachelor's degree within 6 years.

Inference for a Two-Way Table 1

Imagine displaying data in a two-way table to see if two categorical variables are related to each other.

Look at sample data, turn them into percentages, and look for an association between the row and column variables.

Is the association in the sample evidence of an association between these variables in the entire population?

This is a question for a significance test.

Example: Treating Cocaine Addiction

Cocaine addicts need the drug to feel pleasure. Perhaps, giving them a medication that fights depression will help them resist cocaine.

A 3-year study compared an antidepressant called desipramine, lithium (a standard treatment for cocaine addiction), and a placebo.

The subjects were 72 chronic users of cocaine who wanted to break their drug habit. An equal number of the subjects were randomly assigned to each treatment.

Example: Treating Cocaine Addiction (continued)

Group	Treatment	Subjects	Successes	Percent
1	Desipramine	24	14	58.3
2	Lithium	24	6	25.0
3	Placebo	24	4	16.7

The sample proportions of subjects who did not use cocaine are quite different.

The percentage of subjects in the desipramine group who did not use cocaine was much higher than for the lithium or placebo group.

Inference for a Two-Way Table 2

The test that answers this question starts with a two-way table.

Here's the table for the data:

	Success	Failure	Total
Desipramine	14	10	24
Lithium	6	18	24
Placebo	4	20	24
Total	24	48	72

Inference for a Two-Way Table 3

Our null hypothesis, as usual, says that the treatments have no effect.

That is, addicts do equally well on any of the three treatments. The differences in the sample are just the result of chance. Our null hypothesis is:

H_0: There *is no association* between the treatment an addict receives and whether or not there is success in not using cocaine in the population of all cocaine addicts.

Inference for a Two-Way Table 4

The alternative hypothesis says, “Yes, there is some association between the treatment an addict receives and whether or not he succeeds in staying off cocaine.”

H_a: There *is an association* between the treatment an addict receives and whether or not there is success in not using cocaine in the population of all cocaine addicts.

Inference for a Two-Way Table 5

To test H_0, we compare the observed counts in a two-way table with the expected counts.

The expected counts are what we would expect—except for random variation—if H_0 were true.

If the observed counts are far from the expected counts, that is evidence against H_0.

Inference for a Two-Way Table 6

In all, 24 of the 72 subjects succeeded, which is an overall success rate of one-third because ${}^{24}/_{72}$ is one-third.

If the null hypothesis is true, there is no difference among the treatments. So we expect one-third of the subjects in each group to succeed.

There were 24 subjects in each group, so we expect 8 successes and 16 failures in each group.

If the treatment groups differ in size, the expected counts will differ.

Inference for a Two-Way Table 7

Fortunately, there is a rule that makes it easy to find expected counts.

Expected counts

The expected count in any cell of a two-way table when H_0 is true is

$$expected\ count = \frac{row\ total \times column\ total}{table\ total}$$

Inference for a Two-Way Table 8

O = observed E = expected

	Success	Failure	Total
Desipramine	O = 14 E = 8	10	24
Lithium	6	18	24
Placebo	4	20	24
Total	24	48	72

The expected count for successes in the desipramine condition is

$$expected\ count = \frac{24 \times 24}{72} = 8$$

The Chi-Square Test 1

observed

To see if the data give evidence against the null hypothesis of “no relationship,” compare the counts in the two-way table with the counts we would expect if there really were no relationship.

If the observed counts are far from the expected counts, that’s the evidence we were seeking.

The significance test uses a statistic that measures how far apart the observed and expected counts are.

The Chi-Square Test 2

Chi-square statistic

The chi-square statistic, denoted χ^2, is a measure of how far the observed counts in a two-way table are from the expected counts. The formula for the statistic is

$$\chi^2 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$$

The symbol $\sum$ means "sum over all cells in the table."

The Chi-Square Test 3

	Observed		Expected	
	Success	**Failure**	**Success**	**Failure**
Desipramine	14	10	8	16
Lithium	6	18	8	16
Placebo	4	20	8	16

And the chi-square statistic becomes:

$$\chi^2 = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$$

$$= \frac{(14-8)^2}{8} + \frac{(6-8)^2}{8} + \frac{(4-8)^2}{8} + \frac{(10-16)^2}{16}$$

$$+ \frac{(18-16)^2}{16} + \frac{(20-16)^2}{16} = 10.5$$

The Chi-Square Test 4

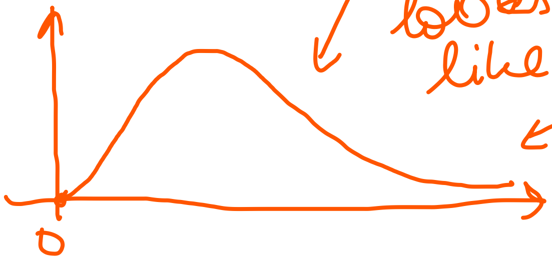
Because χ^2 measures how far the observed counts are from what would be expected if H_0 were true, large values are evidence against the null hypothesis.

Is $\chi^2 = 10.5$ a large value?

You know the drill: compare the observed value 10.5 against the sampling distribution that shows how χ^2 would vary if the null hypothesis were true.

The Chi-Square Test 5

χ^2 density looks like this



This is *not* a normal distribution.

It is a right-skewed distribution, whose values cannot be negative.

This sampling distribution is different for two-way tables of different sizes.

The Chi-Square Test 6

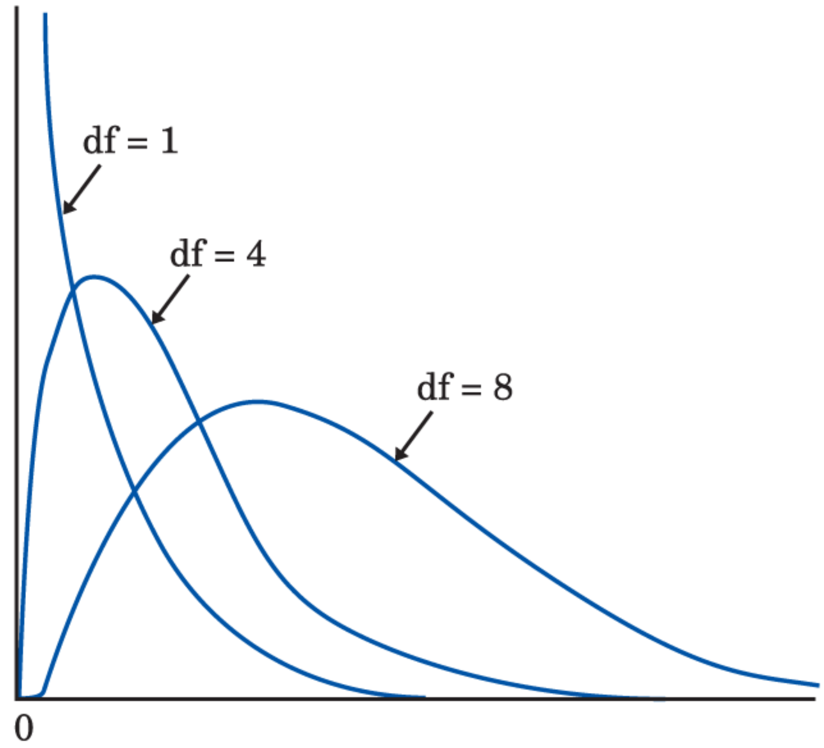
The sampling distribution of the chi-square statistic χ^2 when the null hypothesis of no association is true is called a **chi-square distribution**.

The chi-square distributions are a family of distributions that take only non-negative values and are skewed to the right.

A specific chi-square distribution is specified by giving its **degrees of freedom**.

The **chi-square test** for a two-way table with r rows and c columns uses critical values from the chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

The Chi-Square Test 7



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

The Chi-Square Test 8

We let software get the areas under the chi-square curve to calculate *P*-values for us. Table 24.1 is a shortcut.

It shows how large the chi-square statistic must be in order to be significant at various levels. Each number of degrees of freedom has a separate row in the table. We see, for example, that a chi-square statistic with 3 degrees of freedom is significant at the 5% level if it is greater than 7.81 and is significant at the 1% level if it is greater than 11.34.

The Chi-Square Test 9

We let software get the areas under the chi-square curve to calculate *P*-values for us. Table 24.1 is a shortcut.

It shows how large the chi-square statistic must be in order to be significant at various levels.

This isn't as good as an actual *P*-value, but it is often good enough.

The Chi-Square Test 10

$\chi^2 = 10.5$ from data

TABLE 24.1 To be significant at level α, a chi-square statistic must be larger than the table entry for α

$r=3$, $c=2$

$(r-1)(c-1)=2$

	Significance Level α						
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.01**	**0.001**
1	1.32	1.64	2.07	2.71	3.84	6.63	10.83
2	2.77	3.22	3.79	4.61	5.99	9.21	13.82
3	4.11	4.64	5.32	6.25	7.81	11.34	16.27
4	5.39	5.99	6.74	7.78	9.49	13.28	18.47
5	6.63	7.29	8.12	9.24	11.07	15.09	20.51
6	7.84	8.56	9.45	10.64	12.59	16.81	22.46
7	9.04	9.80	10.75	12.02	14.07	18.48	24.32
8	10.22	11.03	12.03	13.36	15.51	20.09	26.12
9	11.39	12.24	13.29	14.68	16.92	21.67	27.88

The Chi-Square Test 11

The two-way table of three treatments by two outcomes for the cocaine study has three rows and two columns.

That is, $r = 3$ and $c = 2$.

The chi-square statistic, therefore, has degrees of freedom $(r-1)(c-1) = (3-1)(2-1) = (2)(1) = 2$.

The Chi-Square Test 12

Look in the *df* = 2 row of Table 24.1.

χ^2 = 10.5 is larger than the critical value of 9.21 for the α = 0.01 level of significance.

The cocaine study shows a significant relationship (P < 0.01) between treatment and success.

The significance test says only that we have strong evidence of some association between treatment and success. We must look at the two-way table to see the nature of the relationship: desipramine works better than the other treatments.

Using the Chi-Square Test 1

As does our test for a population proportion, the chi-square test uses some approximations that become more accurate as we take more observations.

You can safely use the chi-square test when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater.

The cocaine study easily passes this test: All the expected cell counts are either 8 or 16.

At least 80% of the expected counts must be ≥ 5.

All of them are ≥ 1.

Simpson's Paradox

As is the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables.

In the following example, we will find that, sometimes, a lurking variable might reverse the relationship of what we would expect to find in the data.

Example: Do $\boxed{\text{Medical Helicopters}}$ Save Lives? 1

Back to our very first lecture!

Accident victims are sometimes taken by helicopter from the accident scene to a hospital.

Helicopters save time. Do they also save lives?

Let's compare the percent of accident victims who die with helicopter evacuation and the percent with the usual transport to a hospital by road.

The numbers here are hypothetical, but they illustrate a phenomenon that often appears in real data.

Example: Do Medical Helicopters Save Lives? 2

	Helicopter	Road	Total
Victim died	64	260	324
Victim survived	136	840	976
Total	200	1100	1300

0.32 > 0.24

We see that 32% (64 out of 200) of helicopter patients died, but only 24% (260 out of 1100) of the others died.

That seems discouraging.

Example: Do Medical Helicopters Save Lives? 3

The explanation is that the helicopter is sent mostly to serious accidents, so the victims transported by helicopter are more often seriously injured.

They are more likely to die with or without helicopter evacuation.

We will break the data down into a three-way table that classifies the data by the seriousness of the accident.

We will present a three-way table as two or more two-way tables side-by-side, one for each value of the third variable.

Example: Do Medical Helicopters Save Lives? 4

How can it happen that the helicopter does better for both groups of victims but worse when all victims are combined? Look at the data: half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients. So the helicopter carries patients who are more likely to die. The original two-way table did not take into account the seriousness of the accident and was therefore misleading. This is an example of Simpson's paradox.

Example: Do Medical Helicopters Save Lives? 5

Serious Accidents		
	Helicopter	**Road**
Died	48	60
Survived	52	40
Total	100	100

0.48 < 0.6

Less Serious Accidents		
	Helicopter	**Road**
Died	16	200
Survived	84	800
Total	100	1000

0.16 < 0.2

How can it happen that the helicopter does better for both groups of victims but worse when all victims are combined?

Look at the data: Half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients.

Example: Do Medical Helicopters Save Lives? 6

Serious Accidents		
	Helicopter	Road
Died	48	60
Survived	52	40
Total	100	100

Less Serious Accidents		
	Helicopter	Road
Died	16	200
Survived	84	800
Total	100	1000

So the helicopter carries patients who are more likely to die.

The original two-way table did not take into account the seriousness of the accident and was therefore misleading.

This is an example of Simpson's paradox.

Example: Do Medical Helicopters Save Lives? 7

Simpson's Paradox

An association or comparison that holds for all of several groups can disappear or even reverse direction when the data are combined to form a single group. This situation is called **Simpson's paradox**.

Simpson's paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables. Remember the caution from Chapter 15: beware the lurking variable.

Statistics in Summary 1

- Categorical variables group individuals into classes. To display the relationship between two categorical variables, make a **two-way table** of counts for the classes. We describe the nature of an association between categorical variables by comparing selected percentages.

- As always, lurking variables can make an observed association misleading. In some cases, an association that holds for every level of a lurking variable disappears or changes direction when we lump all levels together. This is **Simpson's paradox**.

Statistics in Summary 2

- The **chi-square test** tells us whether an observed association in a two-way table is statistically significant. The **chi-square statistic** compares the counts in the table with the counts we would expect if there were no association between the row and column variables. The sampling distribution is not Normal. It is a new distribution, the **chi-square distribution**.