

CHAPTER 9 – ANALYSIS OF TWO-WAY TABLES

In Chapter 12 (One-Way ANOVA) we learned how to study the effect of a categorical factor (X) on a quantitative response variable (Y). In chapter 9, we are going to learn how to study this relationship when both variables are categorical.

TWO-WAY TABLE

Remember from Stat I: - **n** represents the number of **observations** in a group
 - **X** represents the number of **successes** in the group

To describe categorical data, we use the **counts** (frequencies) or **percents** (relative frequencies) of individuals that fall into various **categories**. We use a two-way table to record all of our counts.

Since it is the practice to label our explanatory variable x, we put our explanatory variables in columns in our two-way table (like the x-axis in regression) and it is therefore called the **column variable**. Our response variable is usually called y, so we put our response variable in rows in our two-way table (like the y-axis in regression) and it is therefore called the **row variable**. Each combination of values for these two variables is called a **cell**.

Example #1a

Two-Way table for Frequent Binge Drinking and Gender			
Frequent Binge Drinker	Gender		Total
	Male	Female	
Yes	1630	1684	3314
No	5550	8232	13782
Total	7180	9916	17096

So in this example, we want to test weather gender can explain whether someone is a frequent binge drinker, so gender is considered the explanatory variable and is placed in columns and called our column variable. Being a frequent binge drinker is considered the response variable, so it is placed in rows and called our row variable. The cell corresponding to men who are frequent binge drinkers contains the number 1630.

JOINT DISTRIBUTION

We can use the counts found in a two-way table to calculate **proportions**. The collection of these proportions is called the **joint distribution** of the two categorical variables. Since we are dealing with a distribution (joint distribution) the sum of our proportions should always be 1.

Example #1b

Joint Distribution of Frequent Binge Drinking and Gender		
	<u>Gender</u>	
Frequent binge drinker	Men	Women
Yes	0.095	0.099
No	0.325	0.482

Men Yes: $1630/17096^{**} = 0.095$
 Women Yes: $1684/17096 = 0.099$
 Men no: $5550/17096 = 0.325$
 Women no: $8232/17096 = 0.482$

** It is important to remember that we are finding the proportion of people who are **both** a specific gender and yes/no binge drinkers, so we need to use $n = 17096$ as our sample size.

Sum to 1: $0.095 + 0.099 + 0.325 + 0.482 = 1.001 \approx 1$ (rounding error)

MARGINAL DISTRIBUTION

When we want to examine one of our categorical variables by itself, we can look at its **marginal distribution**. There are two marginal distributions for a two-way table: one for each categorical variable.

Example #1c

Marginal Distribution of Gender		
	Men	Women
Proportion	0.420	0.580

Men: $7180/17096^{**} = 0.420$

Women: $9916/17096 = 0.580$

Sum to 1: $0.420 + 0.580 = 1$

**Since we are only interested in the proportion of males in our group, we must use $X = 7180$, since in this case being a man is considered to be a success.

Marginal Distribution of Frequent Binge Drinking		
	Yes	No
Proportion	0.194	0.806

Yes: $3314/17096^{**} = 0.194$

No: $13782/17096 = 0.806$

Sum to 1: $0.194 + 0.806 = 1$

**Since we are only interested in the proportion of frequent binge drinkers in our group, we must use $X = 3314$, since in this case being a frequent binge drinker is considered a success.

Therefore, each marginal distribution from a two-way table is a distribution for a single categorical variable.

CONDITIONAL DISTRIBUTIONS

When we condition on the value of one variable (or know that we are dealing with a specific level of one of our variables) and we calculate the distribution of the other variable, we obtain a **conditional distribution**.

Example #1d

Conditional Distribution of Frequent Binge Drinking for Men		
	Yes	No
Percent	0.227	0.773

Yes: $1630/7180^{**} = 0.227$

No: $5550/7180 = 0.773$

Sum to 1: $0.227 + 0.773 = 1$

We are given a specific level of gender (men), so to calculate the percent of frequent male binge drinkers, we look at the number of frequent binge drinkers, but **only for men.

Conditional Distribution of Frequent Binge Drinking for Women		
	Yes	No
Percent	0.17	0.83

Yes: $1684/9916^{**} = 0.17$
Sum to 1: $0.17 + 0.83 = 1$

No: $8232/9916 = 0.83$

****We are given a specific level of gender (women), so to calculate the percent of frequent female binge drinkers, we look at the number of frequent binge drinkers, but only for women.**

Conditional Distribution of Gender for Frequent Binge Drinkers		
	Men	Women
Percent	0.492	0.508

Men: $1630/3314^{**} = 0.492$
Sum to 1: $0.492 + 0.508 = 1$

Women: $1684/3314 = 0.508$

****We are given a specific level of Binge Drinkers (Yes), so to calculate the percent of frequent male binge drinkers, we look at the number of men, but only for frequent binge drinkers.**

Conditional Distribution of Gender for Non-Frequent Binge Drinkers		
	Men	Women
Percent	0.403	0.597

Men: $5550/13782^{**} = 0.403$
Sum to 1: $0.403 + 0.597 = 1$

Women: $8232/13782 = 0.597$

****We are given a specific level of Binge Drinkers (No), so to calculate the percent of non-frequent male binge drinkers, we look at the number of men, but only for non-frequent binge drinkers.**

Once we have found our conditional distributions, we can study the true nature of the association between our two categorical variables.

Example #1d continued

From our conditional distribution of binge drinking given gender, we can see that both males and females in our sample are more likely to be non-frequent binge drinkers, and that males in our sample are slightly more likely to be frequent binge drinkers than females in our sample.

INFERENCE FOR TWO-WAY TABLES

To examine the relationship between two categorical variables, we need to look at conditional distributions. From the conditional distributions, we can see if there is a difference between the different levels of our explanatory variable. However, we need to perform a significance test in order to determine if the differences are due to chance, or because the different levels really are different. So for our first example, we saw from the conditional distribution of gender given that the person was a frequent binge drinker that there was a difference between the levels (men and women) of our explanatory variable (gender). In order to be convinced that there is a true difference between male and female frequent binge drinkers, and that our conclusions weren't the result of random chance, we then perform a significance test. Specifically, we want to know if the two populations have the same distribution, how likely is it that a sample would show differences as large or larger than those found from the conditional distributions.

When we perform a significance test to determine if there is an association (dependence) between our row variable and our column variable, our **null hypothesis** is that there is **no association**, or that our two variables are **independent**.

Our **alternative hypothesis** is that there is an **association** between our two variables, or that the response variable is **dependent** on the explanatory variable. Our alternative does not specify any particular direction for the association. We cannot describe H_a as either one-sided or two-sided because it includes all of the many kinds of association that are possible.

When our two variables have different levels, we can represent their levels with symbols. The number of levels of our row variable is represented by r , whereas the number of levels of our column variables is represented by c . So for **$r \times c$ tables**, where the columns correspond to independent samples from distinct populations, there are c distributions for the row variable, one for each population. The null hypothesis then says that the c distributions of the row variable are identical. The alternative hypothesis is that the distributions are not all the same.

It is generally sufficient to state the null and alternative hypotheses as

Ho: No association

Vs

Ha: There is an association between our row and column variables.

Theoretically, however, there are two specific variations of these general statements, and the method of sampling dictates which variation is appropriate in a particular situation. If samples have been taken from separate populations, the null hypothesis is a statement about homogeneity (sameness) among the populations, that is to say all of the levels of our row variable (response) are the same. If a sample has been taken from a single population, and two categorical variables measured for each individual, the statement of no relationship is a statement of independence between the two variables.

Example #2

Students in a statistics class were asked “with whom do you find it easiest to make friends?” possible responses were ‘opposite sex’, ‘same sex’, and ‘no difference’. Students were also categorized as male or female. We would like to determine whether there is a statistically significant relationship between the gender of the respondent and the response.

a) State the two categorical variables in our study, and specify which should be the row variable and which the column variable.

Solution:

Row Variable (response variable): the response of “whom is it easier to make friends with?” (three levels: opposite sex, same sex, no difference)

Column Variable (explanatory): gender of the person answering the question (two levels: male and female)

b) Give r and c.

Solution:

$r = 3, c = 2$

c) State our null and alternative hypotheses.

Solution:

H_0 : There is no association between gender and who they think it is easier to make friends with.

H_a : An association does exist between the two variables

To test our null hypothesis, we compare the observed cell values in our $r \times c$ table with the expected cell counts, the values that we would expect to find if our null hypothesis of no relationship were true. To compute the expected cell count, we need our row totals and the column totals. That is, we need the sum of each row as well as the sum of each column. Then our expected cell count is the product of the row total and column total divided by the table total:

Expected cell count:
$$\frac{\text{row total} \times \text{column total}}{n}$$

$$e_{ij} = \frac{r_i \text{ total} \times c_j \text{ total}}{n} \quad \text{where } i=1,2,\dots,r \text{ and } j=1,2,\dots,c$$

Example #3

Calculate the row and column totals for the data found in the ‘with whom do you find it easiest to make friends?’ example. Then calculate the expected cell counts.

	<u>With Whom Is It Easiest to Make Friends?</u>			
	Opposite Sex	Same Sex	No Difference	Total
Females	58	16	63	
Males	15	13	40	
Total				

Solution:

The row and column totals are filled in here:

	<u>With Whom Is It Easiest to Make Friends?</u>			
	Opposite Sex	Same Sex	No Difference	Total
Females	58	16	63	137
Males	15	13	40	68
Total	73	29	103	205

Expected counts:

- For females and opposite sex: $\text{expected cell count} = (137 \times 73) / 205 = 48.79$
- For females and same sex: $(29 \times 137) / 205 = 19.38$
- For females and no difference: $(103 \times 137) / 205 = 68.83$
- For males and opposite sex: $(73 \times 68) / 205 = 24.21$
- For males and same sex: $(29 \times 68) / 205 = 9.62$
- For males and no difference: $(103 \times 68) / 205 = 34.17$

CHI-SQUARED TEST

To test H_0 that there is no association between the row and column variables, we use a statistic that **compares our observed cell counts (x_{ij}) with their corresponding expected cell counts (e_{ij})**. First, we take the difference between each observed count and its respective expected cell count, and we square the differences so that they are all zero or positive values. A large difference means less if it comes from a cell we think will have a large count, so we divide each squared difference by the expected cell count, which has a **standardizing effect**. Finally, we sum up all of these values and the sum is called the chi-squared distribution, X^2 .

$$X^2 = \sum \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}} = \sum \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$

If the expected cell counts and the observed cell counts are very different, a large value of X^2 will result. Large values of X^2 provide evidence against the null hypothesis of no association.

The p-value of a chi-squared test is the probability that our calculated chi-squared test statistic X^2 could be as large as or larger than it is if the null hypothesis is true. Therefore, we need to compare our calculated chi-squared test statistic X^2 with the distribution of a chi-squared statistic when the null hypothesis is true.

If H_0 is true, then the chi-squared test statistic X^2 has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Just like the t and F families of distributions, the shape of a chi-squared distribution depends on the number of degrees of freedom, which are $(r-1)(c-1)$.

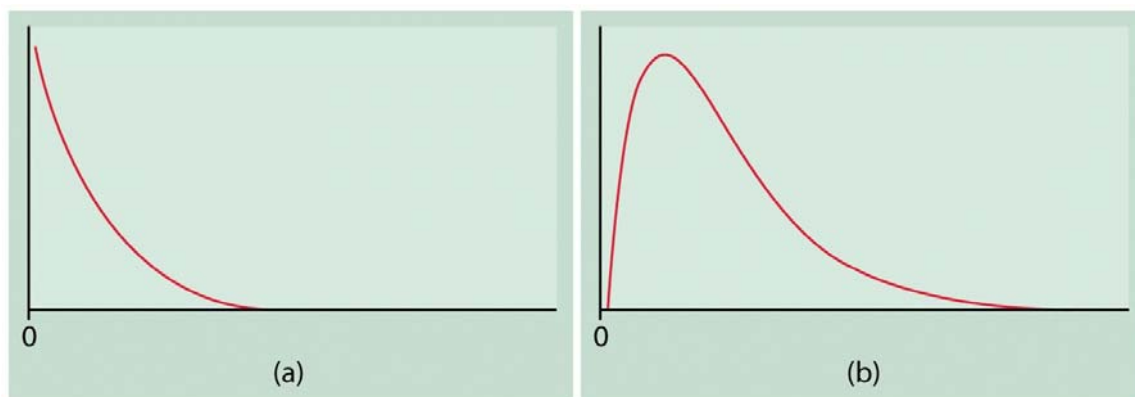
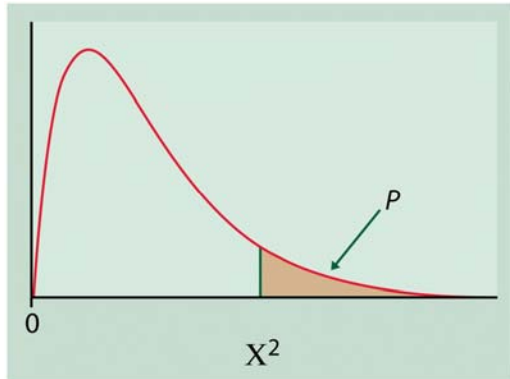


Figure 9.4 (a) the χ^2 density with 2 df or $\chi^2(2)$ (b) the χ^2 density with 4 df or $\chi^2(4)$

So the p-value for the chi-squared test is the probability of getting a statistic greater than or equal to our calculated chi-squared test statistic.



The p-value for the chi-squared test is $p = P(\chi^2 \geq X^2)$, where χ^2 is a random variable having the $\chi^2(df)$ distribution with $df = (r-1)(c-1)$.

Just like the **F and t distributions are related ($t^2 = F$)**, so are the chi-squared distribution and the normal distribution. The **χ^2 statistic is exactly equal to the square of the z statistic**, and $\chi^2(1)$ critical values are equal to the square of the corresponding $N(0,1)$ critical values, $\chi^2 = [Z^2]$.

Therefore, we can compare two population proportions using either the chi-squared test or the two-sample z test from Stat I, and the results will be exactly the same. The advantage of the z test is that we can specifically test either one-sided or two-sided alternatives, whereas the chi-squared test only tests for an association, but does not specify the type of association. However, the chi-squared test can compare more than two populations, whereas the z test compares only two.

If our chi-squared test results in rejecting H_0 and concluding that there is a relationship between our two variables, use the information from our two-way table to describe the type of association.

Example #4

Calculate the chi-squared test statistic for the friends' data and then find its corresponding p-value. Use the two-way table to expand on your conclusion.

Solution:

$$\chi^2 = ((58 - 48.79)^2 / 48.79) + ((16 - 19.38)^2 / 19.38) + ((63 - 68.83)^2 / 68.83) + ((15 - 24.21)^2 / 24.21) + ((13 - 9.62)^2 / 9.62) + ((40 - 34.17)^2 / 34.17) = 8.5 \text{ with } df = (3-1) \times (2-1) = 2.$$

This test statistic corresponds to a p-value of $P(\chi^2 > 8.5) = 0.014$. We have evidence to reject the null hypothesis of no association.

For females, our sample's probability that they thought the opposite sex was easier to make friends with was 0.42. The male's probability was 0.22. So females seem to think that making friend's with the opposite sex is easier than males think it is. Males tend to think that making friend's with the same sex is easier than females think it is. Also, males seem to think that there is no difference more than females do.

Example # 5

In the 1993 General Social Survey conducted by the National Opinion Research Center at the University of Chicago, participants were asked:

*Do you favor or oppose the death penalty for persons convicted of murder?
Do you think the use of marijuana should be made legal or not?*

A two-way table of counts for the responses to theses two questions is:

		Marijuana?		Total
		Legal	Not Legal	
Death Penalty?	Favor	152	561	
	Oppose	61	159	
Total				

a) Calculate the row and column totals for the table. Then calculate the joint distribution of Marijuana support and Death Penalty support.

Solution:

Here are the column and row totals filled in:

		Marijuana?		
		Legal	Not Legal	Total
Death Penalty?	Favor	152	561	713
	Oppose	61	159	220
	Total	213	720	933

		Marijuana?	
		Legal	Not Legal
Death Penalty?	Favor	0.163	0.601
	Oppose	0.065	0.170

To find these values: for favor/legal cell, this is equal to $152/933$, for favor/not legal cell, this is equal to $561/933$, etc.

b) Calculate the conditional distribution of death penalty support for each of the two marijuana support categories. (Distribution of death penalty support given a specific level of marijuana support)

Solution:

Conditional Distribution of Death Penalty Support for Legalizing Marijuana		
	Favor	Oppose
Percent	$152/213 = 0.714$	$61/213 = 0.286$

Conditional Distribution of Death Penalty Support for Keeping Marijuana Illegal		
	Favor	Oppose
Percent	$561/720 = 0.779$	$159/720 = 0.221$

c) Calculate the conditional distribution of marijuana support for each of the two death penalty support categories. (Distribution of marijuana support given a specific level of death penalty support)

Solution:

Conditional Distribution of Marijuana Support for Favoring Death Penalty		
	Legal	Illegal
Percent	$152/713 = 0.213$	$561/713 = 0.787$

Conditional Distribution of Marijuana Support for Opposing Death Penalty		
	Legal	Illegal
Percent	$61/220 = 0.277$	$159/220 = 0.723$

d) State the null and alternative hypotheses about the two variables that have been used to create the two-way table.

Solution:

H_0 : There is no association between one's stance on marijuana and the death penalty

H_a : An association exists

e) Calculate the expected cell counts.

For favor/legal, expected cell count = $(213 \times 713) / 933 = 162.77$

For favor/not legal, expected cell count = $(720 \times 713) / 933 = 550.23$

For oppose/legal, expected cell count = $(213 \times 220) / 933 = 50.23$

For oppose/not legal, expected cell count = $(720 \times 220) / 933 = 169.77$

		Marijuana?		
		Legal	Not Legal	Total
Death Penalty?	Favor	152	561	713
	Oppose	61	159	220
	Total	213	720	933

f) Calculate the Chi-squared test statistic.

Solution:

$$\chi^2 = ((152 - 162.77)^2/162.77) + ((561 - 550.23)^2/550.23) + ((61 - 50.23)^2/50.23) + ((159 - 169.77)^2/169.77) = 3.9 \text{ with } df = (2-1) \times (2-1) = 1.$$

g) Find the p-value and use it, the joint and conditional distributions to make your conclusion.

Solution:

This test statistic corresponds to a p-value of $P(\chi^2 > 3.9) = 0.048$. At the level $\alpha = 0.05$, we have evidence to reject the null hypothesis of no association. From the table, it seems that someone who is in favor of the death penalty is less likely to support the legalization of marijuana than someone opposing the death penalty.