Bookmark this page

Case C→C: Overview

Learning Objective: Choose the appropriate inferential method for examining the relationship between two variables and justify the choice.

Inference for the Relationships between Two Categorical Variables (Chi-Square Test for Independence)

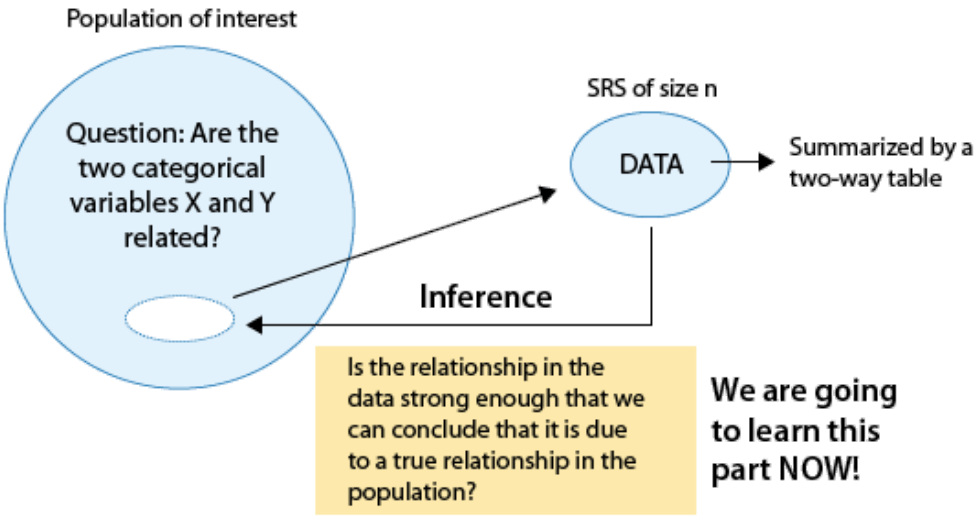
Overview

The last three procedures that we studied (two-sample t, paired t, and ANOVA) all involve the relationship between a categorical explanatory variable and a quantitative response variable, corresponding to Case C→Q in the role/type classification table below. Next, we will consider inferences about the relationships between two categorical variables, corresponding to case C→C.

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C→C	√C→Q
	Quantitative	Q→C	Q→Q

In the Exploratory Data Analysis section of the course, we summarized the relationship between two categorical variables for a given data set (using a two-way table and conditional percents), without trying to generalize beyond the sample data.

Now we perform statistical inference for two categorical variables, using the sample data to draw conclusions about whether or not we have evidence that the variables are related in the larger population from which the sample was drawn. In other words, we would like to assess whether the relationship between X and Y that we observed in the data is due to a real relationship between X and Y in the population or if it is something that could have happened just by chance due to sampling variability.



The statistical test that will answer this question is called the **chi-square test for independence**. Chi is a Greek letter that looks like this: χ, so the test is sometimes referred to as: The χ^2 test for independence.

The structure of this section will be very similar to that of the previous ones in this module. We will first present our leading example, and then introduce the chi-square test by going through its 4 steps, illustrating each one using the example. We will conclude by presenting another complete example. As usual, you'll have activities along the way to check your understanding, and to learn how to use software to carry out the test.

Let's start with our leading example.

Example

In the early 1970s, a young man challenged an Oklahoma state law that prohibited the sale of 3.2% beer to males under age 21 but allowed its sale to females in the same age group. The case (Craig v. Boren, 429 U.S. 190, 1976) was ultimately heard by the U.S. Supreme Court.

The main justification provided by Oklahoma for the law was traffic safety. One of the 3 main pieces of data presented to the court was the result of a "random roadside survey" that recorded information on gender, and whether or not the driver had been drinking alcohol in the previous two hours. There were a total of 619 drivers under 20 years of age included in the survey.

Here is what the collected data looked like:

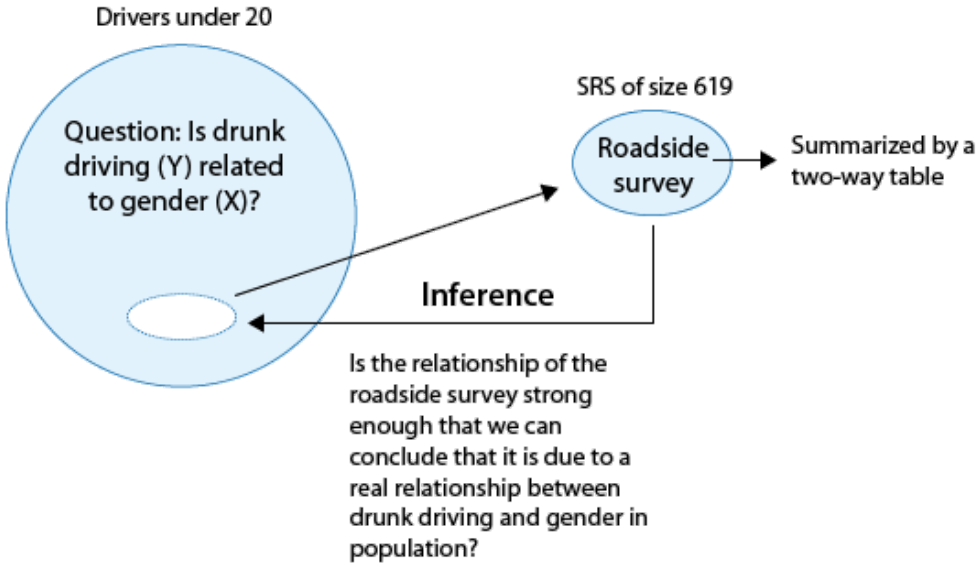
	Gender	Drove drunk?
Driver 1	M	Y
Driver 2	F	N
Driver 3	F	Y
•	•	•
•	•	•
•	•	•
Driver 619	M	N

The following two-way table summarizes the observed counts in the roadside survey:

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Our task is to assess whether these results provide evidence of a significant ("real") relationship between gender and drunk driving.

The following figure summarizes this example:



Note that as the figure stresses, since we are looking to see whether drunk driving is related to gender, our explanatory variable (X) is gender, and the response variable (Y) is drunk driving. Both variables are two-valued categorical variables, and therefore our two-way table of observed counts is 2-by-2. It should be mentioned that the chi-square procedure that we are going to introduce here is not limited to 2-by-2 situations, but can be applied to any r-by-c situation where r is the number of rows (corresponding to the number of values of one of the variables) and c is the number of columns (corresponding to the number of values of the other variable).

Before we introduce the chi-square test, let's conduct an exploratory data analysis (that is, look at the data to get an initial feel for it). By doing that, we will also get a better conceptual understanding of the role of the test.

Exploratory Analysis

Recall that the key to reporting appropriate summaries for a two-way table is deciding which of the two categorical variables plays the role of explanatory variable, and then calculating the conditional percentages — the percentages of the response variable for each value of the explanatory variable — separately. In this case, since the explanatory variable is gender, we would calculate the percentages of drivers who did (and did not) drink alcohol for males and females separately.

Here is the table of conditional percentages:

Gender (X)	Drank Alcohol in Last 2 Hours (Y)? Yes	No	Total
Male	77/481=16.0%	404/481=84.0%	100%
Female	16/138=11.6%	122/138=88.4%	100%

For the 619 sampled drivers, a larger percentage of males were found to be drunk than females (16.0% vs. 11.6%). Our data, in other words, provide some evidence that drunk driving is related to gender; however, this in itself is not enough to conclude that such a relationship exists in the larger population of drivers under 20. We need to further investigate the data and decide between the following two points of view:

- The evidence provided by the roadside survey (16% vs 11.6%) is strong enough to conclude (beyond a reasonable doubt) that it must be due to a relationship between drunk driving and gender in the population of drivers under 20.

- The evidence provided by the roadside survey (16% vs. 11.6%) is not strong enough to make that conclusion, and could have happened just by chance, due to sampling variability, and not necessarily because a relationship exists in the population.

Actually, these two opposing points of view constitute the null and alternative hypotheses of the chi-square test for independence, so now that we understand our example and what we still need to find out, let's introduce the four-step process of this test.

Scenario: Alcoholism Risk in 9/11 Responders

The purpose of this activity is to introduce you to the example that you are going to work through in this section, and for you to get a feeling for the data by conducting exploratory analysis.

Background: Alcoholism Risk in 9/11 Responders

Among firefighters and other "first responders" to the World Trade Center on September 11, 2001, there have been reports of increased alcohol-related difficulties (e.g., DUI). A survey of 9/11 first responders (On the Front Line: The Work of First Responders in a Post-9/11 World) conducted by Cornell researcher Samuel Bacharach was released in 2004. To see the report, click here. Based on the research, we can construct the following two-way table of observed counts:

Firefighters* vs. Alcohol Risk
Based on 2004 Study of NY Firefighters

*does not include officers

	No risk for alcohol problems**	Moderate to Severe risk for alcohol problems**	
Participated in 911 rescue	793	309	1102
Did Not Participate in 911 rescue	441	110	551
	1234	419	1653

** as defined by the DSM criteria
(also used by the National Institute on Alcohol Abuse) and determined by survey results

Using the data from this research, we would like to investigate whether alcohol risk among New York firefighters is significantly related to participation in the 9/11 rescue.

Learn By Doing (1/1 point)

There are two categorical variables in this problem:* Alcohol risk (none, moderate to severe)* Participation in the 9/11 rescue (yes, no)Which is the explanatory variable and which is the response?

Your Answer:

Explanatory = participation in 9/11 rescue
Response = alcohol risk

Our Answer:
Since we want to investigate whether participation in the 9/11 rescue had an effect on alcohol risk, the explanatory variable is "participation in the 9/11 rescue" and the response variable is "alcohol risk."

Resubmit Reset

Learn By Doing (1/1 point)

Conduct exploratory analysis of the data by calculating the conditional percentages. Summarize your findings. Recall that we calculate the percentages of the response variable for each category of the explanatory variable separately.

Your Answer:

Only 309 responders /1102 responders had alcoholism problems

Our Answer:

According to our data, a larger percentage of the firefighters who participated in the 9/11 rescue are at risk for alcohol problems compared to the percentage among firefighters who did not participate in the 9/11 rescue (28% vs. 20%). The question now is whether this difference in the percentages is significant or not. In other words, the next step would be to carry out a significance test that will assess whether observing data like ours (where there is a difference of 8% between the firefighters who participated in the 9/11 rescue and those who didn't) is likely to happen just by chance or the fact the we observed these data provides enough evidence that the risk of alcohol-related problems among New York firefighters and first responders is indeed related to participation in the 9/11 rescue.

Resubmit Reset